CHAPTER

13

Clustering analysis

Rafael Garcia-Dias[1], *Sandra Vieira*[1],
Walter Hugo Lopez Pinaya[1, 2], *Andrea Mechelli*[1]

[1] Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, United Kingdom; [2] Centre of Mathematics, Computation, and Cognition, Universidade Federal do ABC, Santo André, São Paulo, Brazil

13.1 Introduction

Clustering analysis aims to find the most natural way of grouping a dataset. Normally, this is achieved by the application of unsupervised algorithms which organize a collection of n observations $(X_1, \quad X_2 \quad , \ldots, \quad X_n)$ into K groups $(g_1, \quad g_2 \quad , \ldots, \quad g_K)$ based on a similarity criterion, such that observations in the same group are more alike than observations in different groups.

Fig. 13.1 illustrates three possible distributions that we can find in a dataset. In Fig. 13.1A, we see a distribution of red dots with a clear-cut separation between two groups. If we look at the density distribution across variable x_2 (represented with a red line in Fig. 13.1E), we can see a single peak. However, if we look at its density distribution across x_1 (represented with a red line in Fig. 13.1D), we can see two clear peaks. When a dataset presents with more than one peak in the density distribution of at least one of its variables, we call this a multipeak distribution (also known as multimodal distribution, referring to the presence of multiple statistical modes in the distribution). In Fig. 13.1B, we can also see a multipeak distribution; however, in this case there is no clear-cut separation between groups. Finally, in Fig. 13.1C, we have a uniform distribution in which we cannot see any clear peaks.

Clustering analysis is a common approach when there is a multipeak distribution of observations in the dataset. However, clustering analysis can also be used to classify observations in distributions without clear-cut separation among groups and even to classify observations in uniform

https://doi.org/10.1016/B978-0-12-815739-8.00013-4

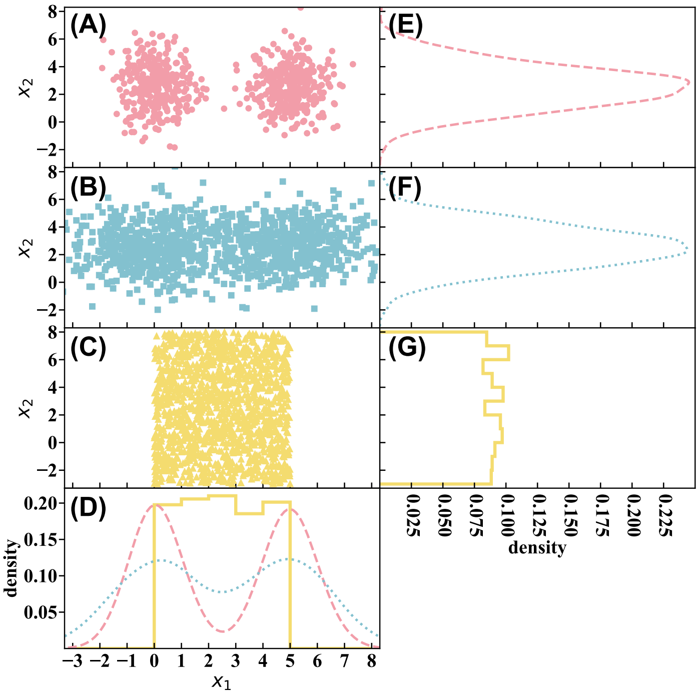

FIGURE 13.1 Examples of multipeak and uniform distributions. Panel (A) show a multipeak distribution with a clear-cut separation between groups; panel (B) shows a multipeak distribution without a clear-cut separation between groups; panel (C) shows a uniform distribution; panel (D) shows the densities of the three distributions across; and panels (E), (F), and (G) show the density of each distribution across variable. Distribution in panel (A) is shown as a dashed line, distribution in panel (B) is shown as a doted line and distribution in panel (C) is shown as a solid line.

distributions. An example would be the classification of the frequency distribution of the light spectrum into distinct groups (i.e., colors), which is routinely used in many scientific contexts and in everyday life. Likewise, most disorders are characterized by a continuous distribution of clinical symptoms, with patients lying at different points across the disorder spectrum. In addition, some disorders can present multipeak continuous distributions, with unclear boundaries between the different diagnoses. A potential issue with some clustering methods is that distinct groups will be found even when there is no multipeak distribution in the dataset. For this reason, before interpreting the results of these methods, it is important to check if the data do have a multipeak distribution. In the absence of such distributions, one can still use clustering to determine underlying patterns of interest in the data; however, the results should not be interpreted in terms of well-separated categories.

A key benefit of clustering analysis is the ability to shed light on the underlying structure of a dataset even when its properties are not obvious. Two-dimensional distributions are often self-explanatory with distinct peaks being easily identified; here clustering analysis has limited applicability. In contrast, in multidimensional distributions, it can be very difficult to spot groups or overdensities; here clustering analysis can be very useful as an alternative to classification when labeled data are not available. Furthermore, in some cases, one can use labeled data in clustering analysis—an approach known as semisupervised learning. While we do not cover semisupervised learning in this chapter, we refer the reader to Chapelle, Schölkopf, Zien, Schlkopf, and Zien (2006) for a detailed description of this approach.

There are several cluster algorithms available in the literature (e.g., Fahad et al., 2014; Shirkhorshidi, Aghabozorgi, Wah, & Herawan, 2014). In general, all involve the following main tasks: (1) feature selection, i.e., the selection of the features to be included in the clustering analysis, taking theoretical and practical considerations into account, with each observation X being defined by N features, $X(x_1, x_2, \dots, x_N)$; (2) choice of a similarity metric, the mathematical function that defines the similarity between observations in the dataset (e.g., Euclidean, Manhattan distances); (3) application of the grouping criterion via a clustering algorithm that organizes the observations according to their similarities; and (4) cluster validation, an evaluation of the reliability of the derived groups. The grouping criterion is the core of clustering analysis, as it defines how the observations are assigned to each cluster.

The clustering algorithms can be classified based on three main characteristics. First, they are either partitional or hierarchical, meaning that they divide the observations into simple groups (i.e., partitional) or into groups and subgroups (i.e., hierarchical). Second, they are either hard or soft clustering algorithms. In hard clustering, each observation is assigned to a single class, whereas in soft clustering each observation receives a probability of belonging to each class. Finally, they are either centroid-based or density-based. Centroid-based clustering assigns the observation with respect to their distance to the center of the cluster, while density-based algorithms assign objects based on the local density around the observation.

Clustering analysis is not without limitations. The main drawback is that it relies on expert knowledge of the field to interpret the results, as in many cases there are no labeled data and no other means to derive meaning from the resulting groups. Another drawback relates to the determination of the number of groups. While some clustering algorithms require the researcher to specify this number as an input of the model, others have hyperparameters that influence the number of groups derived. There are some cluster validation techniques that can help determine the optimal number of groups; however, these are not always reliable (Milligan & Cooper, 1985). A final drawback is that, sometimes, the most natural way of grouping the available data does not necessarily reflect the groups of interest within the context of a research project—an issue that could be addressed by selecting features that have greater relevance to the scientific aim of the study.

In this chapter, we present *K*-means, one of the simplest clustering algorithms, and in particular, we use this method to illustrate the main virtues and limitations of clustering analysis (Section 13.1.1.1). In Section 13.1.1.2, we discuss the important topic of cluster validation, whereas in Section 13.1.4.1 we cover the limitations of *K*-means. Alternative algorithms are suggested in Section 13.1.4.2. In the final part of the chapter, we show some applications of clustering analysis to brain disorders (Section 13.1.5), before presenting some conclusive remarks (Section 13.1.6).

13.2 Method description

13.2.1 The algorithm

K-means is a type of clustering analysis that was first developed in the 1950s (Ball & Hall, 1965; Macqueen, 1967; Lloyd, 1982; Steinhaus, 1956) and has a long history of being applied to the investigation of brain disorders. For instance, in the 1980s, it was used to identify subgroups of patients with schizophrenia who showed different clinical presentations (Farmer, McGuffin, & Spitznagel, 1983). Because of its efficiency and simplicity, it is one of the most used algorithms in the literature. Fig. 13.2 shows the steps taken by the algorithm to split a dataset into nonoverlapping clusters. The first step is to choose the number of clusters, the *K* in *K*-means. The second step is to define *K* initial cluster centers, as illustrated in Fig. 13.2B; this second step is called the initialization of the algorithm and can be done randomly or using some more sophisticated algorithms, such as *K*-means++ (Arthur & Vassilvitskii, 2007). The third step is to assign each observation in the sample to the most similar cluster, according to the chosen similarity metric; in this example, we are using Euclidean distances, as illustrated in Fig. 13.2C. The convergence of the algorithm is achieved via the repetition of steps 3 and 4, by redefining the cluster centers based on the centroid of the observations assigned to each cluster, as shown in Fig. 13.2D. Fig. 13.2E illustrates the repetition of steps 3 and 4 until a convergence criterion is met. Usually, the convergence criterion is either a threshold in the within-cluster variance or a minimal number of reassignations of the observations between two consecutive iterations.

13.2.2 Cluster validation

In this section, we discuss the important topic of cluster validation by focusing on three key issues: (1) measure of reliability, (2) choice of *k*, and (3) testing for a multipeak distribution.

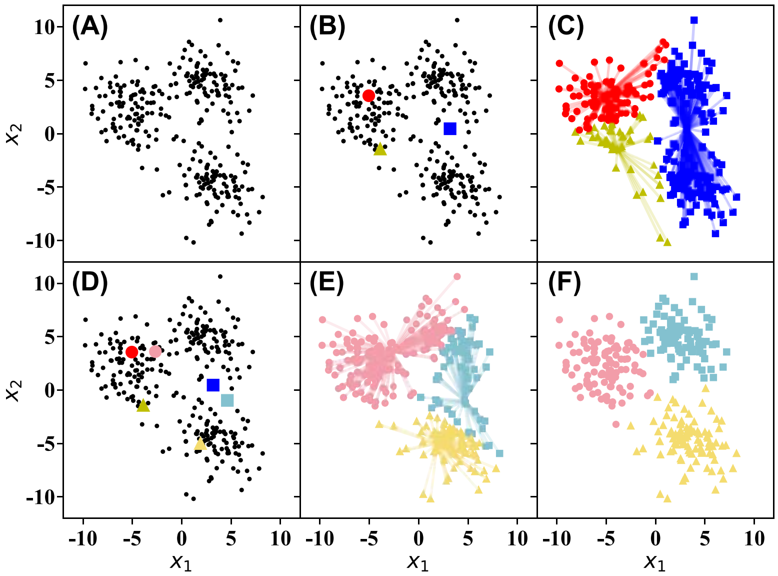

FIGURE 13.2 This figure illustrates the main steps in *K*-means clustering. Panel (A) shows the unlabeled data; panel (B) shows the initial cluster centers (step 2); panel (C) illustrates the Euclidean distance measurements from the cluster centers to each point and the assignation of the objects to the closest center (step 3); panel (D) illustrates the reassignation of the centers to the centroid of the objects in each cluster, with the previous centers shown in darker colors, and the new centers shown in lighter colors (step 4); panel (E) represents the repetition of steps 3 and 4; and panel (F) shows the result of the algorithm converged.

Measure of reliability

In supervised classification, one splits the labeled data into two and then uses one part to train the machine learning model and the other part to validate it. In unsupervised learning, the labeled data are often unavailable, and therefore it is important to establish some metrics to determine the reliability of the models. This is known as cluster validation. Here, the simplest measurement of reliability is the sum of squared error (SSE), defined as

$$\mathrm{SSE} = \sum_{j}^{K} \sum_{i \epsilon g_j} \left(\frac{X_i - \mu_j}{n_j} \right)$$

where μ_j is the mean of the n_j objects in the group g_j. The SSE is the cumulative difference among observations in each group; in other words, it measures how similar are the observations within groups. As mentioned at the beginning of this chapter, clustering analysis aims to group observations such that the ones in the same group are similar to each other. Therefore, the lower the SSE, the closer we are to achieving this aim.

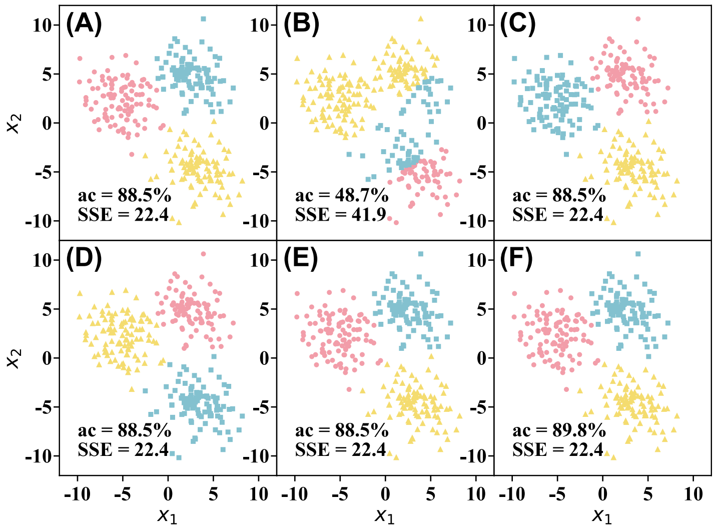

FIGURE 13.3 Examples of how initialization impacts in the final groups derived by *K*-means. Each panel corresponds to a different *K*-means initialization. In all panels, the three groups are indicated using the colors red (gray in print version) circles, yellow triangles (light gray in print version), and blue squares (dark gray squares in print version), respectively. The percentage of correctly classified observations (ac) is shown in the bottom left insets followed by the sum of squared error (SSE).

Given the randomized nature of the initialization of the algorithm, each time *K*-means is run on a dataset will lead to a different result, and each run will have its corresponding SSE. This is shown in Fig. 13.3, which illustrates how the initialization can impact on the final solution. In this case, we have the ground truth labels for the observations, and therefore we can use them to calculate the accuracy score to assess how *K*-means behaves (see Chapter 2 for information on how to perform this calculation). The figure shows that 4 out of 5 times the algorithm can give accuracies of around 89%; however, in one of the runs, it results in a solution with about 49% accuracy. In the context of optimization, this is known as convergence to local minima. To overcome this problem, *K*-means algorithms are often implemented to run multiple times and select the solution with the lowest SSE. In Fig. 13.3 we show the accuracy of scores and the SSE for each classification. It ca be seen that the highest accuracies coincide with the lowest SSEs. As the SSE is very similar for solutions with the highest accuracies, it does not help us choose the best solution; however, it clearly helps us discard poor solutions.

Choosing K *with silhouette scores*

In *K*-means, as in many other clustering algorithms, the number of clusters is a fundamental input parameter. Determining the optimal number of clusters, therefore, is also one aspect of cluster validation. In Fig. 13.2, it is trivial to conclude there are three main groups, but in multidimensional datasets, determining the number of clusters can be very challenging. There is no universal solution to this problem; however, the literature offers a range of possible methods which are reviewed in detail elsewhere (Halkidi, Batistakis, & Vazirgiannis, 2001; Jain, 2010; Steinley, 2006). One of these methods is the silhouette score (Rousseeuw, 1987). This score corresponds to the mean silhouette values s_i, defined as

$$s_i = \begin{cases} 1 - d_w(i)/d_b(i), & \textit{if } d_w(i) < d_b(i) \\ 0, & \textit{if } d_w(i) = d_b(i) \\ d_w(i)/d_b(i) - 1, & \textit{if } d_w(i) > d_b(i) \end{cases},$$

where $d_w(i)$ is the mean within-cluster distance, i.e., the mean of the distances from observation i to all the observation within its group, and $d_b(i)$ is the mean between-cluster distance, the mean of the distances from observation i to all the observation in other groups, for each X_i in the sample. As s_i will always be between -1 and 1, so will the silhouette score. The silhouette score is 1 when all the clusters are well separated and is 0 when the clusters are completely overlapping. To determine the optimal number of clusters, we need to measure the silhouette score for groupings with a different number of clusters. The classification with the highest silhouette score corresponds to the group configuration with the maximum separation among clusters.

The silhouette score has its limitations. First, the algorithm requires the calculation of the complete distance matrix of the dataset, which can be impractical for large volumes of data. The silhouette score can also return the wrong number of clusters depending on the geometry of the groups in the dataset, and furthermore, it can return random values when confronted with a homogeneous random distribution. Therefore, one should only trust this method when a clear peak is present in the silhouette score measurements. In addition, when applying clustering analysis to distributions without clear-cut separation among groups, the choice of K should be driven by the meaning of the derived groups, based on previous qualitative analysis. Adolfsson, Ackerman, and Brownstein (2018) have introduced a novel methodology to determine whether or not there is an underlying multipeak distribution in a dataset. This methodology could be used to confirm the nature of the distribution and interpret the results of the silhouette score (see Adolfsson et al., 2018 for detail).

Testing for a multipeak distribution

As we mentioned earlier, it is important to know whether or not the underlying distribution is multipeak. Only when applying clustering analysis to a multipeak distribution, one can trust the results of analytical tools as the silhouette score. Therefore, the best practice is to start with some exploratory statistical analysis to determine the distribution of the observations before applying any clustering models. There are a number of methods that can be used to test for multipeak distribution. In these methods, the null hypothesis is that the data are generated by a single-peak distribution. When we apply this sort of tests to a dataset, the p-value reflects the probability of drawing that dataset from a single-peak distribution. Therefore, if the p-value is small enough, we can assume the distribution is multipeak. One of the most popular methods to test for multipeak distribution is the dip test (Hartigan & Hartigan, 1985), which measures the difference between the empirical distribution of the data and a hypothetical single-peak distribution. Although the dip test is unidimensional, it can be applied to multidimensional data using the pairwise distances of the observations in the dataset, as prescribed by Adolfsson et al. (2018). Once multimodality is confirmed, we can proceed with the silhouette score analysis to determine the optimal number of clusters in the dataset.

13.2.3 Main drawbacks of *K*-means

In this section, we present the most discussed drawbacks of *K*-means. We will see that some of these are common to most clustering algorithms, some are specific to *K*-means, and some are overstated concerns which do not represent actual drawbacks. Here is a comprehensive list:

(a) *K-means always return groups even when there are no actual groups in the distribution.* As mentioned before, *K*-means always returns groups of observations, much in the same way as linear regression always returns a line, even if applied to an exponential distribution. This is also true for most of the clustering algorithms and should not be seen as a drawback, but as a characteristic of the method to be taken into account. If we attempt to divide a uniform distribution of observations into two groups, *K*-means will successfully return two groups of observations. This becomes a problem only if one interprets this division of the data as proof of the existence of two well-defined groups in the distribution, which would be incorrect.

(b) *The number of clusters must be inputted into K-means.* *K*-means requires the number of clusters to be defined a priori. However, most of the other cluster algorithms rely on the same requirement. There are some algorithms that do not require this; however, these depend on other hyperparameters that impact on the number of derived clusters. There are some methods that can help in the search for the optimal number of clusters, such as the silhouette score discussed earlier, but all of them require careful application and interpretation. The same can be said about other clustering algorithms that use hyperparameters to determine the number of clusters; in this case, the issue is how to best tune these hyperparameters.

(c) *K-means does not guarantee convergence to global minima.* As we have discussed in Section 13.1.1.2, sometimes *K*-means converges to a local minimum. We have seen that this problem is solvable by repeating the clustering process many times and comparing its outputs. Nanetti, Cerliani, Gazzola, Renken, and Keysers (2009) showed that, to find the global solution when performing connectivity-based cortical segmentation using diffusion-weighted images, it is necessary to run the algorithm more than 250 times. As *K*-means is much less computationally intense than other clustering algorithms, often it will be faster to run it hundreds of times than to run an algorithm that guarantees convergence to global minima just once. This can vary with the dimensionality of the problem and the shape of the clusters.

(d) *K-means does not work well when the data are not linearly separable.* *K*-means generates nonoverlapping clusters based on the similarity criterion measured with respect to the cluster center. Euclidean distance is the similarity criteria most often applied with *K*-means. In this case, the decision boundaries can only be linear, and therefore the algorithm does not work well when data are not linearly separable. However, if another similarity criterion is applied, or the data are previously transformed to become linearly separable, then the algorithm can work well.

(e) *K-means does not perform well when clusters have different scales, different shapes, or an unbalanced number of observations.* These problems are illustrated in Fig. 13.4. In this figure, we have four clusters violating the main assumptions of *K*-means: group I is nonspherical (diversity in shapes), group II has 4.5 times the size of groups III and IV (diversity in scales), and group IV has the same scale of group III but fewer observations (unbalanced groups). In Fig. 13.4A we apply *K*-means with four clusters. As *K*-means seeks to generate groups with approximately the same size, it splits group I into two and considers groups II and IV to be a single group. Therefore, the *K*-means solution does not match the four original groups as described. In Fig. 13.4B, we prescribe the use of five clusters. This results in a new solution which is in greater agreement with the underlying structure of the data than the earlier solution with the correct number of clusters. This example

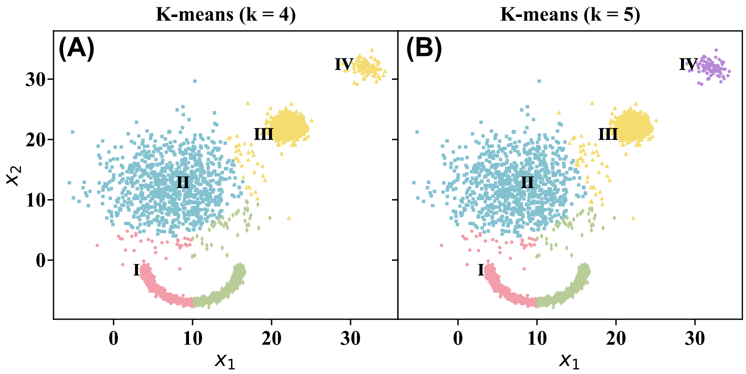

FIGURE 13.4 Example of *K*-means classification. Different shapes correspond to different classes. The shapes are colored according to the *K*-means grouping. Region I corresponds to a semicircular distribution of 1000 objects. Region II encloses a Gaussian distribution of 1000 objects with a standard deviation of 4.5 unities. Region III is a Gaussian distribution of 1000 objects with a standard deviation of one unity. Region IV encloses a Gaussian distribution of 100 objects with a standard deviation of 1 unity. Panel (A) shows the result of *K*-means with four clusters whereas panel (B) shows the solution for five clusters.

illustrates how *K*-means can return suboptimal results even when there is a multipeak underlying distribution of well-separated clusters, and the right number of clusters is given. However, for some applications, this could be an acceptable solution, in the same way, that a linear approximation of an exponential distribution can be useful in certain cases.

The above limitations could be considered characteristics of the algorithm rather than drawbacks per se. However, is not uncommon to find alarmed concerns about how *K*-means fails in some applications in the literature. These reactions come from the assumption that machine learning methods should be able to provide universal automatic solutions for every problem—an unrealistic expectation resulting from the association of these methods with the increasingly promising field of artificial intelligence. The limitations of *K*-means should motivate researchers to use this method in a thoughtful and critical manner, rather than abandoning it altogether.

13.2.4 Alternatives to *K*-means

In the last section, we have discussed some limitations of the *K*-means algorithm. Here, we briefly present some alternatives to perform clustering analysis when *K*-means is not the best fit for the problem. For this, we return to the example given in Fig. 13.4 and see how the Gaussian mixture model (GMM) and density-based spatial clustering of applications with noise (DBSCAN) can perform better in that situation.

13.2.4.1 *Gaussian mixture model*

The GMM (Pearson, 1894; Shanmugam, 2009) assumes that the underlying distribution in the dataset can be described as the mixture of a finite number of Gaussian (normal) distributions. The GMM fits Gaussian distribution to generate groups of observations from the data. The main limitation of GMM is that it can be successfully applied only when the underlying distribution is a combination of Gaussian distributions. There are other mixture models that assume other types of statistical distributions, but they will all have limitations when dealing with datasets including groups with diverse shapes. Fortunately, in many scientific applications, the distribution is actually a mixture of Gaussians, thus the GMM algorithm can be successfully applied.

13.2.4.2 *Density-based spatial clustering of applications with noise*

DBSCAN (Daszykowski & Walczak, 2010) uses the local density of the space to group objects together and to define the cluster boundaries. The main assumption of this algorithm is that the distance among edges of the different groups is smaller than the distance among observations inside the groups. This assumption is manifested through two fundamental parameters of the algorithm, the neighborhood radius (r) and the minimal number of observations necessary to define a core object (n_{min}). The core observations are defined as the observations in the distribution that have at least n_{min} observations, including the observation itself, inside of a radius r from it. A group is defined by at least one core observation and all *density-reachable* points from the core observations. An observation is density-reachable from a core observation if there is a path between them that passes only through core observations. If an observation is not density-reachable from any core observation, it is considered an outlier. Fig. 13.5 illustrates the process when $n_{\text{min}} = 4$, with core

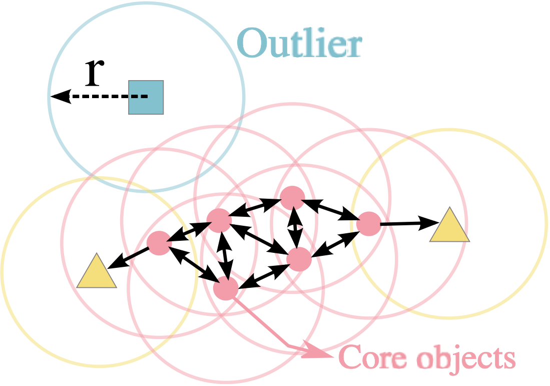

FIGURE 13.5 Clustering analysis with density-based spatial clustering of applications with noise with $n_{\text{min}} = 4$. The image shows core points (red [dark gray in print version] circles), density-reachable objects (yellow [light gray in print version] triangles), and an outlier point (blue [gray in print version] squares).

observations represented as red dots, density-reachable observations represented as yellow dots, and the outlier in the distribution represented as the blue dot. In this example, all yellow and red dots would be part of the same group. Therefore, in DBSCAN, there is no assumption that all observations must belong to a cluster, as is the case in *K*-means; instead, some observations can be classified as outliers.

The algorithm starts by selecting a random observation in the sample and establishing whether or not it is a core observation. If it is a core observation, then it propagates the classification through direct reachable observations until there are no density-reachable points from the starting core observation. The algorithm continues to randomly select other observations from the unevaluated ones, each time creating a new group until all observations in the dataset are assigned to a group or as an outlier.

13.2.4.3 Comparison of K*-means with GMM and DBSCAN*

In this section, we use the dataset in Fig. 13.4 to compare the performance of the three models. Although the comparison can be done by visual inspection, as the groups are well separated, we will use the homogeneity score as a quantitative metric to compare the results. The homogeneity score is similar to the accuracy score; however, the two scores will differ when cluster labels are not matched. This is because the random initialization of the clustering algorithms causes the labels of the clusters to be shuffled during the classification; such shuffling will not affect the homogeneity score—where the labels of the clusters are interchangeable—but will affect the accuracy score—where each cluster must be associated with a specific label. For instance, Table 13.1 shows two possible outcomes of the application of *K*-means to a dataset with four objects. We see that both classifications 1 and 2 detect the presence of two distinct groups within the dataset, resulting in a homogeneity score of 1 for both classifications; however, the measured accuracy score is 1 in the first classification and 0 in the second classification. A formal definition of the homogeneity score is given in Rosenberg and Hirschberg (2007).

TABLE 13.1 Homogeneity and the accuracy scores for two possible outcomes of a clustering analysis. In the second column, each number represents the classification of one observation. 0 means the observation is considered a member of group g_0 and 1 means the observation is assigned to group g_1. It can be seen that the homogeneity and accuracy scores differ when cluster labels are not matched.

	Group	Homogeneity score	Accuracy score
Classification 1	0, 0, 1, 1	1	1
Classification 2	1, 1, 0, 0	1	0
Ground truth	0, 0, 1, 1		

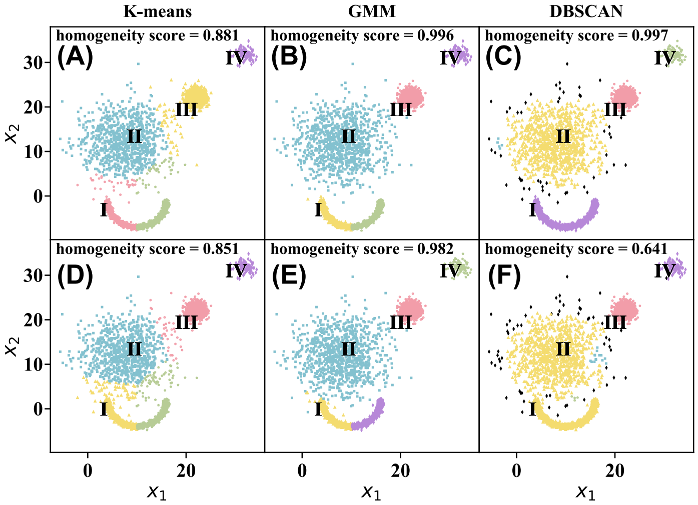

FIGURE 13.6 Comparison of the performances of *K*-means (left panels, (A) and (D)), Gaussian mixture model (GMM) (middle panels, (B) and (E)), and density-based spatial clustering of applications with noise (DBSCAN) (right panels, (C) and (F)). Two slightly different datasets are presented in the top and bottom panels, respectively; in particular, groups I and II have greater proximity in the dataset shown in the bottom panels than the one shown in the top panels. This slight difference affects the performance of the algorithms, especially in the case of DBSCAN.

Fig. 13.6 shows how *K*-means, GMM, and DBSCAN compare in terms of performance using two slightly different datasets. In Fig. 13.6A,B and C, we present the dataset described in Fig. 13.4 grouped by the three algorithms. We see that GMM performs slightly better than *K*-means. Although we have three Gaussian distributions, the presence of one non-Gaussian distribution is sufficient to negatively impact the performance of the GMM algorithm. Here, GMM fails to identify the semicircular distribution; instead it divides this into two groups. In contrast, we see the DBSCAN perform much better than *K*-means and GMM. Moreover, DBSCAN manages to successfully identify outliers in group II. It should be noted that groups are colored according to the label assigned by the algorithm; for instance, group III is colored in blue when *K*-means is used (Fig. 13.6A) but is colored in red when GMM and DBSCAN are used (Fig. 13.6B,C). This difference in coloring illustrates how the initialization shuffles the final labels of the clusters. In addition to shuffling the final labels, the randomness of the initialization impacts the results of the algorithm, as shown in Fig. 13.3. Therefore, when performing clustering analysis, it is important to feed the algorithms with known random seeds and report these numbers in any scientific publication. The use of known random seeds will guarantee that a random distribution can be exactly reproduced.

From the above example, considering only the top panels of Fig. 13.6, one could naively conclude that DBSCAN is better than *K*-means and GMM and decide to keep the former and discard the latter from their machine learning tool kit. However, with a small modification of the dataset, as shown in the bottom panels of Fig. 13.6, we can see that DBSCAN is not universally better than the other models. In particular, by shortening the distance between group I and group II, we break the main assumption in DBSCAN, i.e., the distance intergroup is no longer much larger than the distance among the observations within each group. In this case, DBSCAN performs worse than *K*-means and GMM. We also note that, in both cases, GMM performs better than *K*-means; however, there will be other instances where *K*-means performs better than GMM depending on the characteristics of the dataset. This example highlights the importance of understanding the models' main assumptions and the underlying structure of the dataset. When applying clustering analysis, it is good practice to try alternative clustering algorithms and use cluster validation metrics as well as existing knowledge of the field to interpret the results.

13.3 Applications to brain disorders

Clustering analysis has long been applied to the investigation of a wide range of brain disorders, including, among others, schizophrenia (Carpenter, Bartko, Carpenter, & Strauss, 1976), eating disorder (Stice et al., 2001), personality disorder (Petrocelli, Glaser, Calhoun, & Campbell, 2001), panic disorder (Zilcha-Mano et al., 2015), and Parkinson's disease (Verbaan et al., 2010). The most frequent application involves the use of *K*-means to identify subtypes of patients within a disorder, with the vast majority of studies using clinical measurements as selected features and Euclidean distance as the similarity metric. Another possible application of *K*-means involves the investigation of patterns in the data—for example, distinct neurocognitive profiles—which are expressed above and beyond diagnostic categories. A further possible application of *K*-means involves the estimation of functional connectivity patterns from functional magnetic resonance imaging (fMRI) data (Allen et al., 2014; Calhoun, Miller, Pearlson, & Adali, 2014; Rashid, Damaraju, Pearlson, & Calhoun, 2014). Below we review these applications to illustrate the potential benefits and challenges of applying clustering analysis to brain disorders.

13.3.1 Identifying disorder subtypes

When applying *K*-means to the identification of subtypes of patients within a brain disorder, most studies determined the number of clusters either by using the knowledge of the field or combining *K*-means with the Ward's linkage hierarchical clustering (Abramowitz, Franklin, Schwartz, & Furr, 2003; Aderka et al., 2012; Zilcha-Mano et al., 2015). The Ward's

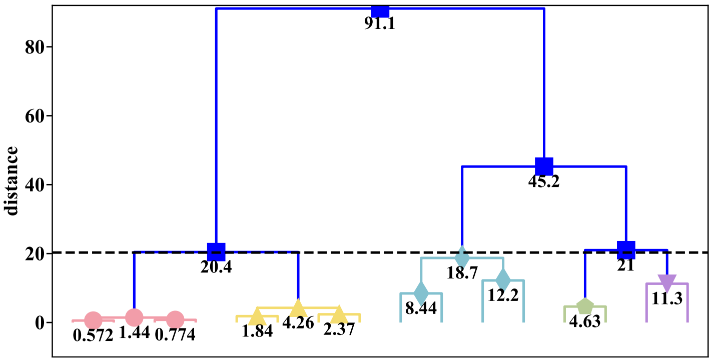

FIGURE 13.7 Dendrogram of the dataset presented in Fig. 13.4. Distances are measured using Ward's linkage clustering, and the *horizontal line* shows the distance used as a threshold for deciding whether or not to merge groups.

linkage is a decision rule used to combine clusters, with the aim of minimizing the ratio of within-group to between-group variation; this type of rule is used in agglomerative clustering (AG). In AG, we start by defining each observation in the dataset as a cluster; we then measure the distance between each pair of clusters and decide whether or not to merge them into a single cluster based in such distance. Using Ward's linkage, one can draw a tree of the distances between possible clusters in the sample, which is called a dendrogram. The visual inspection of the dendrogram can then help define the number of clusters in the sample. In Fig. 13.7, we show the dendrogram of the dataset presented in Fig. 13.4. The horizontal line represents the distance used as a threshold for deciding whether or not to merge two clusters into one. We can see that choosing a threshold between 18.7 and 20.4 would result in 5 clusters; a threshold between 20.4 and 21 would result in 4 clusters; and a threshold higher than 21 would result in 3 clusters. This shows that it is not trivial to use a dendrogram to determine the optimal number of clusters because there is no obvious way of deciding among the possible thresholds. If a dendrogram does not solve the problem of determining the number of clusters, the combination of *K*-means and Ward clustering can still help conversion to the global minimum.

Abramowitz et al. (2003) have used *K*-means together with Ward clustering to identify subtypes in a sample of 132 adult patients diagnosed with obsessive-compulsive disorder. The algorithms were applied on clinical measures obtained using a revised version of the Yale—Brown Obsessive—Compulsive Scale (Y-BOCS; Goodman et al., 1989a, 1989b). The number of clusters was determined using Ward clustering, and the centroid of the Ward-based clusters was used to initialize *K*-means. This led to a solution with five groups, named as harming, contamination, hoarding, unacceptable thoughts, and symmetry. Interestingly, when the researchers compared the outcome of cognitive behavioral therapy among the five groups, there were particularly poor responses in the hoarding group in comparison with the other groups. This example shows that the use of clustering analysis to identify disorder subtypes has the potential of supporting clinical decision-making, for example, by helping identify patients who are likely to show poor response to standard treatment and require additional support.

13.3.2 Identification of cross-diagnostic neurocognitive profiles

Our next exemplar application of clustering analysis refers to the identification of cross-diagnostic neurocognitive profiles. In Lewandowski, Sperry, Cohen, and Öngür (2014), the authors used clinical and cognitive measures to investigate neurocognitive variability in a cross-diagnostic sample of patients with psychotic disorders. The sample included 41 patients with schizophrenia, 53 patients with schizoaffective disorder, and 73 patients with bipolar disorder with psychosis. The researchers used Ward clustering and *K*-means independently and then compared the results; the aim was to look for a consistent pattern of results that was supported by both algorithms. In this case, therefore, the centers of the Ward classification were not used as input for *K*-means as in Abramowitz et al. (2003), as this would have introduced artificial consistency between the two results. The results suggested that the total sample could be divided into four groups with distinct neurocognitive profiles: a *neuropsychologically normal* cluster, a globally and significantly impaired cluster, and two clusters of mixed cognitive profiles. These four groups were distributed across the three diagnostic categories; in other words, there was no correspondence between neurocognitive profile and diagnosis. This may either indicate that the clustering model was not appropriate to the geometry of the problem or that diagnostic categories do not really map to distinct neurocognitive profiles. The authors also applied the analysis of variance (ANOVA) to the four clusters to confirm the existence of the four neurocognitive profiles. It is worth mentioning, however, that the application of ANOVA to the same data used to generate the clusters may not provide confirmation of the existence of a multipeak distribution. Instead, such application could lead to the conclusion that the clusters are distinct even when the underlying distribution is a single-peak distribution. A more informative approach might be to use different sets of data to perform the clustering and validate it. For example, one could use clinical data to group patients using *K*-means and then use ANOVA on brain imaging data from the same patients to confirm the significance of the groups.

13.3.3 Investigation of functional connectivity states in fMRI

Our final exemplar application of clustering analysis refers to the investigation of functional connectivity from fMRI data. Allen et al. (2014)

argue that the assumption that functional connectivity is stationary throughout the duration of a scanning session can limit the value of the findings. The authors addressed this issue by proposing an approach that combines independent component analysis (ICA) and *K*-means to estimate temporal variability in functional connectivity states across 405 subjects. ICA is a dimensionality reduction method, in some aspects similar to principal component analysis (PCA), discussed in the previous chapter. An important difference between ICA and PCA is that ICA derives components that are statistically independent but not necessarily orthogonal between each other; whereas PCA generates orthogonal components. In addition, while ICA creates components that are focused on more local features of the data, PCA creates components that express more global features. For an in-depth comparison between the two dimensionality reduction methods, we refer to Draper, Baek, Bartlett, and Beveridge (2003).

In the context of resting-state fMRI, ICA is used to decompose the whole-brain data into homogeneous regions that can be used to analyze the data consistently across subjects. In Allen et al. (2014) the components derived from ICA were used to build a covariance matrix for each subject. The values in the covariance matrices were then grouped using *K*-means. In particular, the authors used the ratio between within-group and between-group mean distances to determine the optimal number of clusters. This was implemented using Manhattan distances instead of Euclidian distances as in the previous examples; this decision was taken to mitigate the so-called *curse of dimensionality*, i.e., the fact that even simple problems, such as the grouping of two well-separated balanced Gaussian distributions, can be very challenging if the number of dimensions is too high. *K*-means is susceptible to the curse of dimensionality, mainly because the information in the distance measurement loses sensitivity as the number of dimensions increases (due to the growth of the volume of the space). However, Manhattan distances are less affected by the curse of dimensionality than Euclidian distances and therefore should be preferred when dealing with high-dimensional datasets. Using Manhattan distances, Allen et al. (2014) were able to identify, for the first time, time-dependent functional connectivity states in resting-state imaging data between regions in lateral parietal and cingulate cortex. These findings challenge the traditional notion of stationary connectivity patterns in the human brain and provide evidence for temporal flexibility in the functional coordination within and neural networks. The future investigation of this temporal flexibility in patients with psychiatric and neurological conditions could help refine our understanding of the neurofunctional basis of these disorders.

13.4 Conclusion

More than 60 years after its original development, *K*-means remains a relevant algorithm in many fields of science. Owing to its agility and simplicity, *K*-means offers a simple way of extracting data-driven groups from datasets. Data-driven grouping can be particularly useful when dimensionality is high and there is no labeled data to train any supervised models. In psychiatry and neurology, it has been extensively used to identify subtypes within a certain disorder of interest (Abramowitz et al., 2003; Calamari, Wiegartz, & Janeck, 1999). In addition, even when labeled data are available, clustering can be useful to review and refine the definition of these labels. For instance, Lewandowski et al. (2014) reported cross-diagnostic neurocognitive profiles that do not map to diagnostic categories.

Like all machine learning algorithms, *K*-means has its limitations. For example, the emphasis on minimizing the global SSE can result in misleading grouping, as shown in Fig. 13.4. Unbalanced samples and clusters with diverse shapes or sizes are particularly challenging to *K*-means. Defining the number of clusters is a fundamental step and can be challenging when we do not have a well-grounded knowledge of the data. To overcome these problems, we suggest the application of the dip test to determine if there is a multipeak distribution in the data, and the use the silhouette score in the definition of the number of clusters and in the interpretation of the results. When the data violate the main assumptions of *K*-means, we recommend the application of other clustering methods. A further limitation of *K*-means and other types of clustering analysis is that the results can be difficult to replicate (Verbaan et al., 2010). This is because of the lack of information on the implementation of the model. The performance of most algorithms depends on a series of hyperparameters including random seeds; therefore, to ensure replicability, we recommend the transparent reporting of how the model is implemented including its hyperparameters. We have argued that ANOVA does not help validate the clustering unless different sets of features are used to perform and validate the clustering. While there are some validation algorithms, the ultimate analysis and interpretation of groups must be informed by expert knowledge of the field.

Given the huge volume of unlabeled data constantly generated in many fields, unsupervised learning is a fundamental tool. In the context of brain disorders, it can potentially generate clinical useful insights, for example, by revealing subgroups of patients who are likely to show different clinical outcomes (Abramowitz et al., 2003).

13.5 Key points

- Clustering analysis is a type of unsupervised learning which aims to find the most natural way of grouping a dataset.
- *K*-means is the most popular clustering algorithm and is fast and simple to implement.
- Alternative clustering algorithms include DMM and DBSCAN.
- Before applying any clustering algorithm, it is important to understand the nature of the dataset and the aim of the analysis.
- In clustering analysis, the choice of the number of clusters is a critical step. Some algorithms explicitly ask for this information, others use hyperparameters to derive it.
- Cluster validation techniques, such as the SSE and the silhouette score analysis, can be used to help access the quality of the clustering and the number of groups in a dataset.
- Transparent reporting of how the clustering algorithm is implemented, including the random seeds used for initialization, is a key to guarantee reproducibility.

References

Abramowitz, J. S., Franklin, M. E., Schwartz, S. A., & Furr, J. M. (2003). Symptom presentation and outcome of cognitive-behavioral therapy for obsessive-compulsive disorder. *Journal of Consulting and Clinical Psychology, 71*(6), 1049—1057. https://doi.org/10.1037/0022-006X.71.6.1049.

Aderka, I. M., Hofmann, S. G., Nickerson, A., Hermesh, H., Gilboa-Schechtman, E., & Marom, S. (2012). Functional impairment in social anxiety disorder. *Journal of Anxiety Disorders, 26*(3), 393—400. https://doi.org/10.1016/J.JANXDIS.2012.01.003.

Adolfsson, A., Ackerman, M., & Brownstein, N. C. (2018). *To cluster, or not to cluster: An analysis of clusterability methods*. Retrieved from http://arxiv.org/abs/1808.08317.

Allen, E. A., Damaraju, E., Plis, S. M., Erhardt, E. B., Eichele, T., & Calhoun, V. D. (2014). Tracking whole-brain connectivity dynamics in the resting state. *Cerebral Cortex, 24*(3), 663—676. https://doi.org/10.1093/cercor/bhs352.

Arthur, D., & Vassilvitskii, S. (2007). K-Means++: The advantages of careful seeding. In *Proc ACM-SIAM symposium on discrete algorithms* (Vol. 8, pp. 1027—1035). https://doi.org/10.1145/1283383.1283494.

Ball, G. H., & Hall, D. J. (1965). *Isodata: A novel method of data analysis and pattern classification*. Retrieved from https://apps.dtic.mil/docs/citations/AD0699616.

Calamari, J. E., Wiegartz, P. S., & Janeck, A. S. (1999). Obsessive-compulsive disorder subgroups: A symptom-based clustering approach. *Behaviour Research and Therapy, 37*(2), 113—125. https://doi.org/10.1016/S0005-7967(98)00135-1.

Calhoun, V. D., Miller, R., Pearlson, G., & Adali, T. (October 22, 2014). *The chronnectome: Time-varying connectivity networks as the next frontier in fMRI data discovery. Neuron*. Cell Press. https://doi.org/10.1016/j.neuron.2014.10.015.

Carpenter, W. T., Bartko, J. J., Carpenter, C. L., & Strauss, J. S. (1976). Another view of schizophrenia subtypes: A report from the international pilot study of schizophrenia. *Archives of General Psychiatry, 33*(4), 508–516. https://doi.org/10.1001/archpsyc.1976.01770040068012.

Chapelle, O., Schölkopf, B., Zien, A., Schlkopf, B., & Zien, A. (2006). *Semi-supervised learning* (1st ed.). MIT Press.

Daszykowski, M., & Walczak, B. (2010). Density-based clustering methods. *Comprehensive Chemometrics, 2*, 635–654. https://doi.org/10.1016/B978-044452701-1.00067-3.

Draper, B. A., Baek, K., Bartlett, M. S., & Beveridge, J. R. (2003). Recognizing faces with PCA and ICA. *Computer Vision and Image Understanding, 91*(1–2), 115–137. https://doi.org/10.1016/S1077-3142(03)00077-8.

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., et al. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing, 2*(3), 267–279. https://doi.org/10.1109/TETC.2014.2330519.

Farmer, A. E., McGuffin, P., & Spitznagel, E. L. (1983). Heterogeneity in schizophrenia: A cluster-analytic approach. *Psychiatry Research, 8*(1), 1–12. https://doi.org/10.1016/0165-1781(83)90132-4.

Goodman, W. K., Price, L. H., Rasmussen, S. A., Mazure, C., Delgado, P., Heninger, G. R., et al. (1989). The yale-brown obsessive compulsive scale: II. Validity. *Archives of General Psychiatry, 46*(11), 1012–1016. Retrieved from https://jamanetwork.com/journals/jamapsychiatry/article-abstract/494744.

Goodman, W. K., Price, L. H., Rasmussen, S. A., Mazure, C., Fleischmann, R. L., Hill, C. L., et al. (1989). The yale-Brown obsessive compulsive scale. I. Development, use, and reliability. *Archives of General Psychiatry, 46*(11), 1006–1011. https://doi.org/10.1001/archpsyc.1989.01810110054008.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems, 17*(2/3), 107–145. https://doi.org/10.1023/A:1012801612483.

Hartigan, J. A., & Hartigan, P. M. (1985). The dip test of unimodality. *Annals of Statistics, 13*(1), 70–84. https://doi.org/10.1214/aos/1176346577.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011.

Lewandowski, K. E., Sperry, S. H., Cohen, B. M., & Öngür, D. (2014). Cognitive variability in psychotic disorders: A cross-diagnostic cluster analysis. *Psychological Medicine, 44*(15), 3239–3248. https://doi.org/10.1017/S0033291714000774.

Lloyd, S. P. (1982). Least squares quantization in PCM. Special issue on quantization. *IEEE Transactions on Information Theory, 28*(2), 129–137. Retrieved from https://evasion.imag.fr/Membres/Franck.Hetroy/Teaching/ProjetsImage/2007/Bib/lloyd-1982.pdf.

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1*(233), 281–297. https://doi.org/citeulike-article-id:6083430.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*(2), 159–179. https://doi.org/10.1007/BF02294245.

Nanetti, L., Cerliani, L., Gazzola, V., Renken, R., & Keysers, C. (2009). Group analyses of connectivity-based cortical parcellation using repeated k-means clustering. *NeuroImage, 47*(4), 1666–1677. https://doi.org/10.1016/J.NEUROIMAGE.2009.06.014.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 185*, 71–110. https://doi.org/10.1098/rsta.1894.0003.

Petrocelli, J. V., Glaser, B. A., Calhoun, G. B., & Campbell, L. F. (2001). Early maladaptive schemas of personality disorder subtypes. *Journal of Personality Disorders, 15*(6), 546–559. https://doi.org/10.1521/pedi.15.6.546.19189.

Rashid, B., Damaraju, E., Pearlson, G. D., & Calhoun, V. D. (2014). Dynamic connectivity states estimated from resting fMRI Identify differences among Schizophrenia, bipolar disorder, and healthy control subjects. *Frontiers in Human Neuroscience, 8*, 897. https://doi.org/10.3389/fnhum.2014.00897.

Rosenberg, A., & Hirschberg, J. (2007). *V-measure: A conditional entropy-based external cluster evaluation measure* (pp. 410–420). Retrieved from http://aclweb.org/anthology/D/D07/D07-1043.pdf.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*(C), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Shanmugam, R. (2009). Finite mixture models. *Technometrics, 44*(1), 82–82 https://doi.org/10.1198/tech.2002.s651.

Shirkhorshidi, A. S., Aghabozorgi, S., Wah, T. Y., & Herawan, T. (2014). Big data clustering: A review. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (Vol. 8583, pp. 707–720). LNCS. https://doi.org/10.1007/978-3-319-09156-3_49.

Steinhaus, H. (1956). Sur la division des corp materiels en parties. *Bulletin of the Polish Academy of Sciences, 1*(804), 801.

Steinley, D. (2006). K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology, 59*(1), 1–34. https://doi.org/10.1348/000711005X48266.

Stice, E., Agras, W. S., Telch, C. F., Halmi, K. A., Mitchell, J. E., & Wilson, T. (2001). Subtyping binge eating-disordered women along dieting and negative affect dimensions. *International Journal of Eating Disorders, 30*(1), 11–27. https://doi.org/10.1002/eat.1050.

Verbaan, D., van Hilten, J. J., van Rooden, S. M., Kok, J. N., Heiser, W. J., & Marinus, J. (2010). The identification of Parkinson's disease subtypes using cluster analysis: A systematic review. *Movement Disorders, 25*(8), 969–978. https://doi.org/10.1002/mds.23116.

Zilcha-Mano, S., McCarthy, K. S., Dinger, U., Chambless, D. L., Milrod, B. L., Kunik, L., et al. (2015). Are there subtypes of panic disorder? An interpersonal perspective. *Journal of Consulting and Clinical Psychology, 83*(5), 938–950. https://doi.org/10.1037/a0039373.