Machine Learning-Based Clustering Analysis: Foundational Concepts, Methods, and Applications

12

Miquel Serra-Burriel and Christopher Ames

12.1 Introduction

On a day-to-day basis, one after another, we make unconscious classifications around the things we perceive. From colors to personalities, we classify observations into groups. A recent hypothesis in neuroscience suggests that brains spontaneously learn statistical structure of images by extracting their properties such as geometry or illumination [1]. Clustering analysis is the branch of statistics that formally deals with this task, learning from patterns, and its formal development is relatively new in statistics compared to other branches.

Statistical learning can be broadly defined as supervised, unsupervised, or a combination of the previous two. While supervised learning aims at mapping inputs to pre-specified outputs, unsupervised learning aims at grouping objects so that elements in each group are more similar to each other than those in other groups. The advantage of this approach is that it does not require any assumptions regarding the underlying joint distribution of patterns, also unsupervised learning also does not require labelling, which is usually time- and cost-sensitive or entirely impossible for large, unstructured datasets.

There are a lot of types of clustering. However, the main thing that they share in common is the fact that they try to explain variance in the data with discrete partitions. Cluster analysis made its first public appearance in human anthropology by Driver and Kroeber in 1932 in their quantitative expression of cultural relationships [2]. They used a simple trait-count model of the populations of Polynesia, Plains Sun Dance, America Northwest Coast, and Peru to cluster them. Much has happened since, and the number of applications of such a simple principle is almost infinite. Marketing [3], genetics [4], politics [5], physics [6], ecology [7], and many more fields benefit from it. Most digital companies use it to segment their market and customer base according to their online preferences and behaviors.

How can we cluster? There are a lot of approaches to cluster observations, namely: connectivity-based clustering or hierarchical clustering, centroid-based clustering, and density-based clustering. We will go through each approach, with applications, review dimensionality reduction and two examples of papers that we find meaningful. The Supplementary Content 12.1 presents the R code to replicate our results and create your own, while following these examples.

12.2 Connectivity-Based Clustering

Connectivity-based clustering is based on the idea of building a hierarchy of similar elements within a sample. It can be performed in two ways, bottom-up or agglomerative and top-down or divisive. The former begins with each observation being its own cluster and later pairing them recursively, the later starts with one cluster containing all observations and recursively splitting them into smaller clusters until each observation forms its own group. The results of clustering are usually presented in dendrograms, tree-shaped objects that represent the hierarchy of the clustering product.

To illustrate the basic functionality, let us begin with a toy example of hierarchical clustering with two dimensions or features of a population. We have a sample of 1000 individuals who were subject to two visual perception tasks, one of

Supplementary Information The online version contains supplementary material available at [https://doi.org/10.1007/978-3-030-85292-4_12].

M. Serra-Burriel (✉)
Epidemiology, Biostatistics and Prevention Institute, University of Zurich (UZH), Zurich, Switzerland
e-mail: miquel.serraburriel@uzh.ch

C. Ames
Department of Neurological Surgery, University of California San Francisco (UCSF), San Francisco, CA, USA
e-mail: Christopher.Ames@ucsf.edu

V. E. Staartjes et al. (eds.), *Machine Learning in Clinical Neuroscience*, Acta Neurochirurgica Supplement 134,
https://doi.org/10.1007/978-3-030-85292-4_12

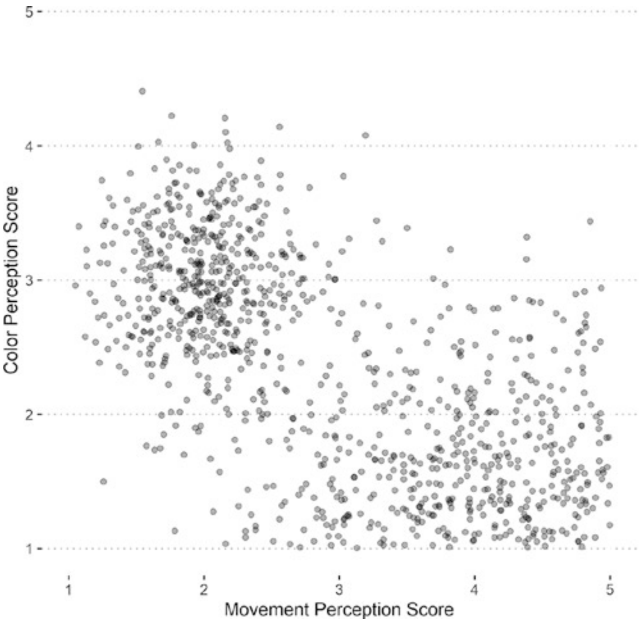

Fig. 12.1 Scatterplot of cognition performance

movement perception and another one of color perception. The scatterplot of the performance in both tasks is presented in Fig. 12.1.

Each dot presents an observation of our study, the *x* axis presents the score of the movement perception task and the *y*-axis presents the score of the color perception one. We want to create groups that are homogeneous within themselves and heterogenous across. The first step to clustering is to create a distance or dissimilarity matrix. This matrix contains the relative distance of each observation with respect to all other observations in the set. There are a lot of ways to create a distance matrix. The most widely used is the Euclidean distance (Eq. 12.1):

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{12.1}$$

where the distance between observations *p* and *q* is equal to the square root of the squared sum of differences in position of *p* and *q* for *n* dimensions. In our case, since we have 1000 observations, each Euclidean distance for participant *i*, with respect to participant *j*, takes form in the following way (Eq. 12.2):

$$d(i,j) = \sqrt{(\text{mps}_i - \text{mps}_j)^2 + (\text{cps}_i - \text{cps}_j)^2} \tag{12.2}$$

where mps is the movement perception score and cps is the color perception score. Figure 12.2 presents the matrix of distances as measured by different metrics.

The upper left panel of the figure presents the Euclidean distance, the upper right panel presents the maximum distance, the lower left the Manhattan distance, and the lower right panel presents the Canberra distance. It can be noted that irrespective of the distance measure, the overall structure of the matrix is fairly similar. Once the matrix has been constructed, two approaches are possible, the above-mentioned agglomerative (also called Agnes) and divisive (also called Diana) functions. In general terms, agglomerative methods are mainly used to find small clusters and divisive methods larger clusters. Let us use the Euclidean distance matrix from Fig. 12.2 to build a dendrogram for the bottom-up approach and split it into four clusters.

Figure 12.3 represents the resulting dendrogram. The *x*-axis presents each observation, while the *y*-axis connections present the pairs of observations and groups of observations. In the figure, the number of clusters is predefined to be 4. However, how can one determine the “natural” or optimal number of clusters in the sample? In our sample there are between 2 and 999 potential clusters. The hierarchy of the model aids us in distinguishing which subgroups stem from other bigger clusters recursively.

There are three main methods in determining the number of clusters: the elbow method [8], average silhouette method [9], and the gap statistic method [10].

The elbow method basically computes the resulting intra-cluster variation (also known as wss) for each of the potential cluster groupings. The location of the bend or “knee,” meaning the inflexion point is usually chosen as the indicator of the appropriate number of clusters. The silhouette method computes a silhouette value that considers how close each observation is to its own cluster compared to the others and the value ranges from −1 to 1, with higher values indicating better clustering for each iteration on the number of clusters. The gap statistic method is similar to the silhouette method; however, it compares the resulting difference in intra-cluster variation from each clustering distribution with a random Monte Carlo simulated sample. Figure 12.4 presents the results on the optimal number of clustering by each of the described methods.

Independently, each method points toward two underlying clusters. We rebuild the previous dendrogram and plot clustered scatterplot of cognition performance groups (Fig. 12.5).

What are the advantages and disadvantages of hierarchical clustering?

Advantages:

- The clustering model has an imposed structural hierarchy, which tends to be more interpretable than other outputs.
- Its construction process is independent of the number of clusters, thus conserving some information that can be of value for the researcher.
- Their simplicity and transparency foster interpretation and reproducibility in external settings.

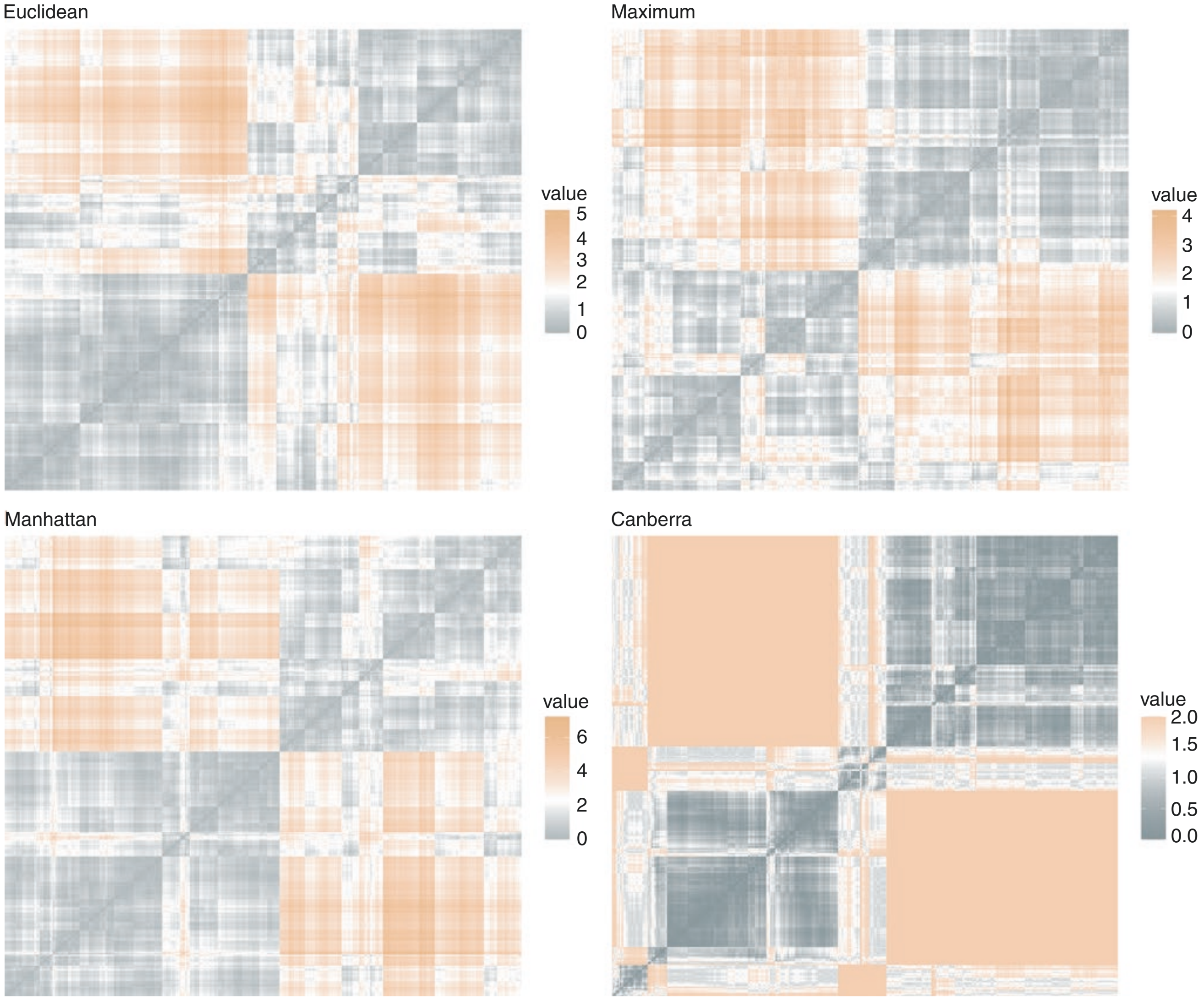

Fig. 12.2 Distance matrix according to different distance measures

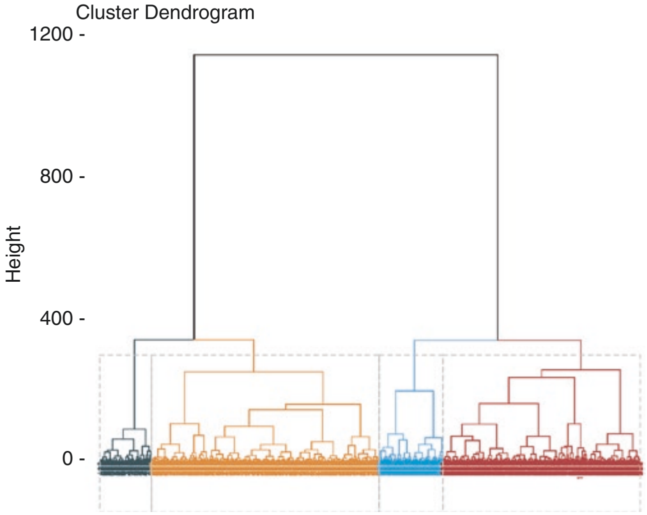

Fig. 12.3 Agnes Dendrogram built with Euclidean distance and four groups

Disadvantages:

- Given its static recursive approach, once a data point has been placed within a cluster, the model does not test for other potential combinations.
- It is more computationally demanding than other clustering algorithms.
- Its sensitivity to outliers requires caution in the pre-processing stage.
- Its results also depend on the metric used to compute the distance or dissimilarity matrix.

12.3 Centroid-Based Clustering

Instead of computing distance across observations and then recursively imposing a hierarchy over them, centroid-based clustering aims to partition observations into k groups in such a way that the sum of distances from points to the cen-

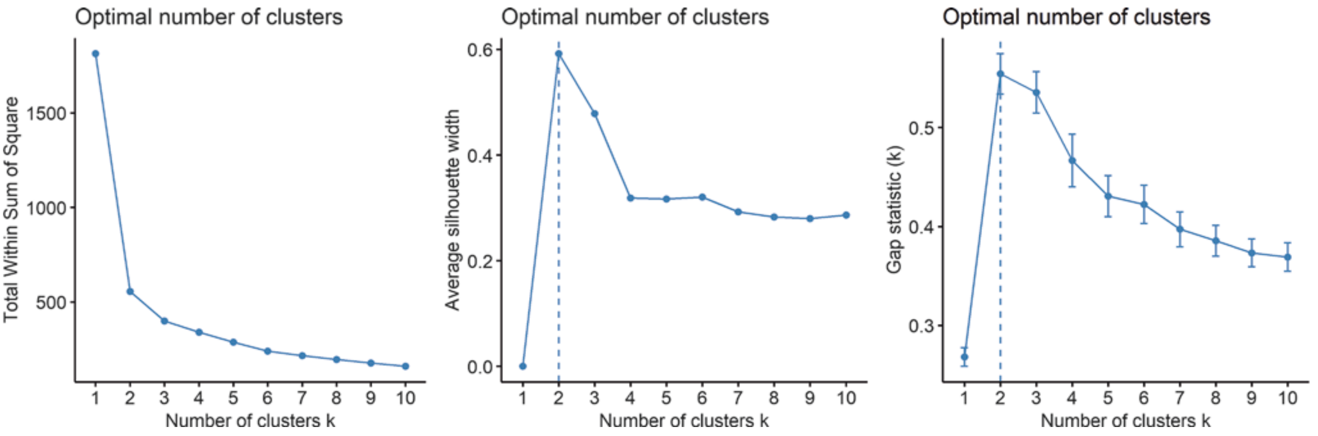

Fig. 12.4 Optimal number of clusters by method

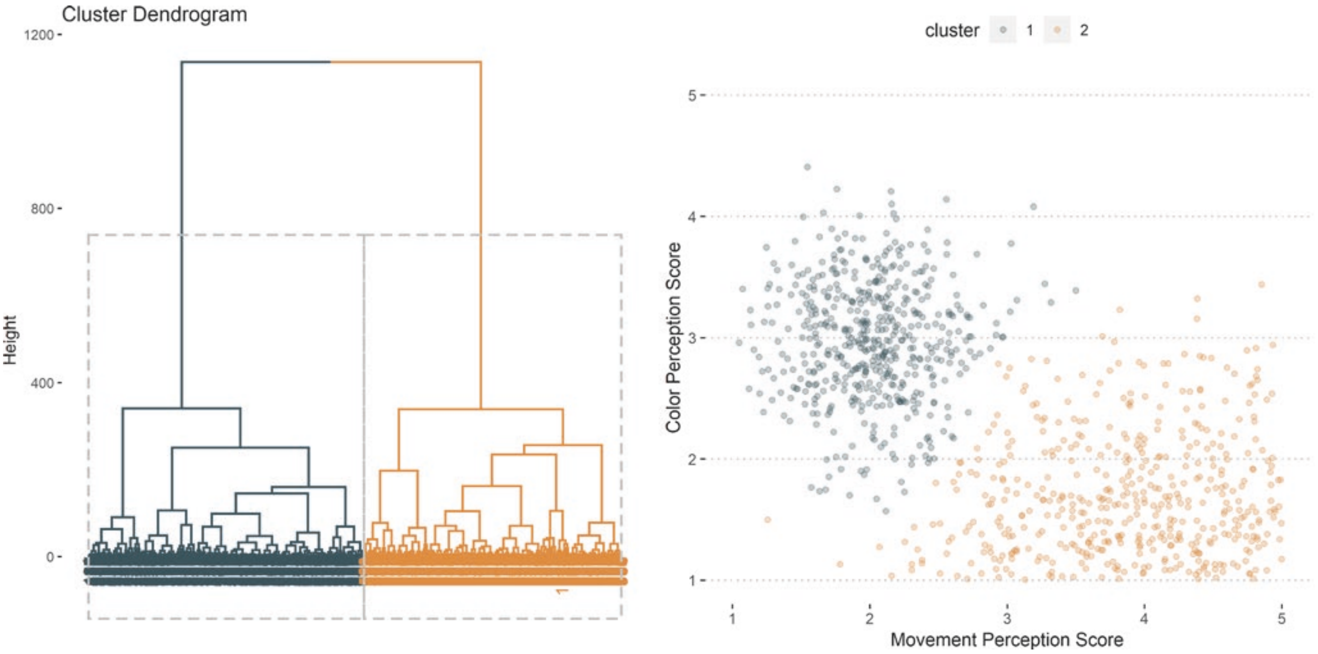

Fig. 12.5 Optimized Agnes Dendrogram Clustering built with Euclidean distances and scatterplot by cluster

troid of their respective clusters is minimized. A valid analogy would be to split a lot of identical pies into *k* pieces, not in even parts necessarily, and select the splitting pattern that is more satisfying. The history of this type of clustering started in the late 50s, with Hugo Steinhaus first in 1956 [11] and Stuart Lloyd in 1957 [12] as a technique for representing analog signals in a digital way. However, the algorithm was further refined by James MacQueen in 1967 [13] and the currently most used one was published in 1979 by Hartigan and Wong [14].

The algorithm has two steps, assignment, and update, preceded by an initialization method. The initialization can be done in two ways. Randomly choosing *k* (the same amount of desired clusters) observations and using them as the initial means or randomly assigning a cluster to each observation and using that cluster mean as the centroid. Then, with either method the assignment step follows. Each observation is assigned to the cluster that is nearer, measured with the Euclidean distance to the centroid as described in Eq. (12.1). Then, the update step follows by simply computing the centroid or mean again for the observations assigned to it. The process is repeated until the observations classified to each cluster do not change. Note that this process does not need to converge necessarily, and the general recommendation is to initialize the algorithm with several random starts, which sometimes prevents the algorithm from not converging.

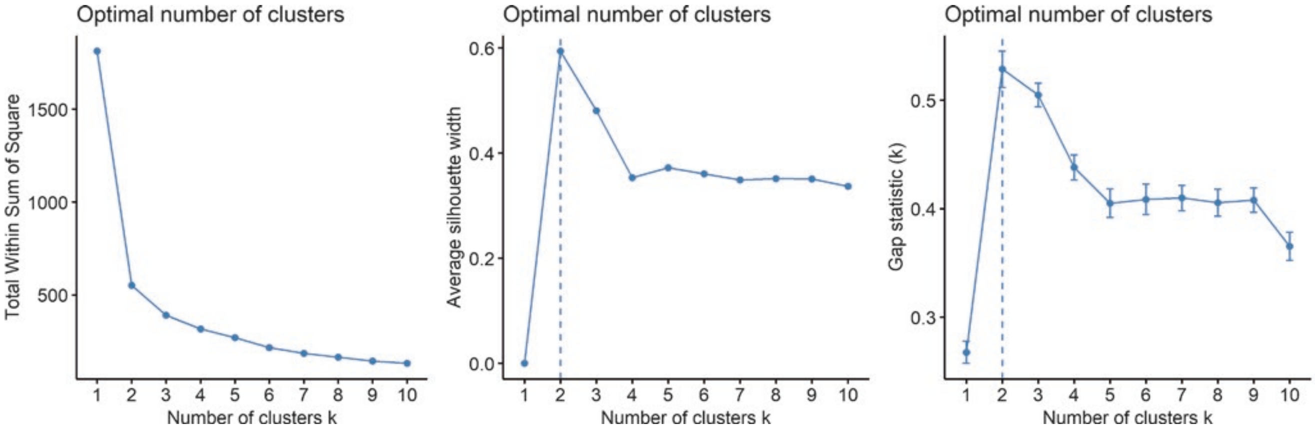

Fig. 12.6 *k*-means clustering partitions, from 2 to 7 clusters

Using the same dataset since the beginning of the chapter we perform a *k*-means clustering with the Hartigan–Wong algorithm, 5 random starts, and from 2 to 7 clusters. Figure 12.6 presents the results of the clustering.

To determine the optimal number of clusters, the same methods described before apply: the elbow method, the silhouette and the Gap method. Again, as with the hierarchical clustering approach the optimal amount is revealed to be two by all accounts. The results of both algorithms are strikingly similar. Figure 12.7 presents the results of the optimization process.

What are the advantages and disadvantages of centroid-based clustering?

Advantages:

- Simpler algorithm to implement.
- Computationally efficient.
- It has been shown to produce results with high external validity.
- Adapts and recognizes well clusters with distinct functional forms and relative sizes.

Disadvantages:

- It does not identify clusters with non-convex shapes.
- It has difficulties identifying clusters of different size.
- It is not completely suited to clustering exercises of high dimensionality, due to Euclidean distance causing the algorithm to converge almost immediately.

12.4 Density-Based Clustering

Compared to the previous two methods of clustering, density-based clustering does not impose a hierarchy or partitions the space. It rather choses clusters based on the defined areas higher statistical density than the rest. Different from before, all observations are not assigned a cluster, points outside the optimized clusters are considered to be noise.

The most used clustering method based on this principle is the density-based spatial clustering of applications with noise (DBSCAN) (Fig. 12.8). Developed in 1996 by Ester, Kriegel, Sander, and Xu, and it is a non-parametric algorithm [15]. The intuition of the algorithm is straightforward. The model uses what is called *minPts*, a threshold on the number of neighboring points, within a radius *e*. Points with more neighboring points than the threshold are considered as a core point, analogous to a centroid. The objective of the algorithm is then to find separated areas of high-density vs. areas of low density.

In abstract terms, the DBSCAN algorithm has three steps. Find the points within the *e* radius of every point, and identify core points with a number of observations above the threshold *minPts*. Then, the connected core points are merged, and finally points are assigned either to clusters or to noise.

What are the advantages and disadvantages of density-based clustering?

Advantages:

- It does not require a pre-specified or optimized number of clusters.
- It does recognize non-convex clusters, and even strange shapes such as circles within circles.
- Because density has a noise component, the method is robust with respect to eliminating outliers.
- It only requires two parameters which are independent of the order or functional forms of the underlying data-generating process.

Disadvantages:

- It does not cluster well data with different densities, meaning that if there are two clusters in the dataset, but

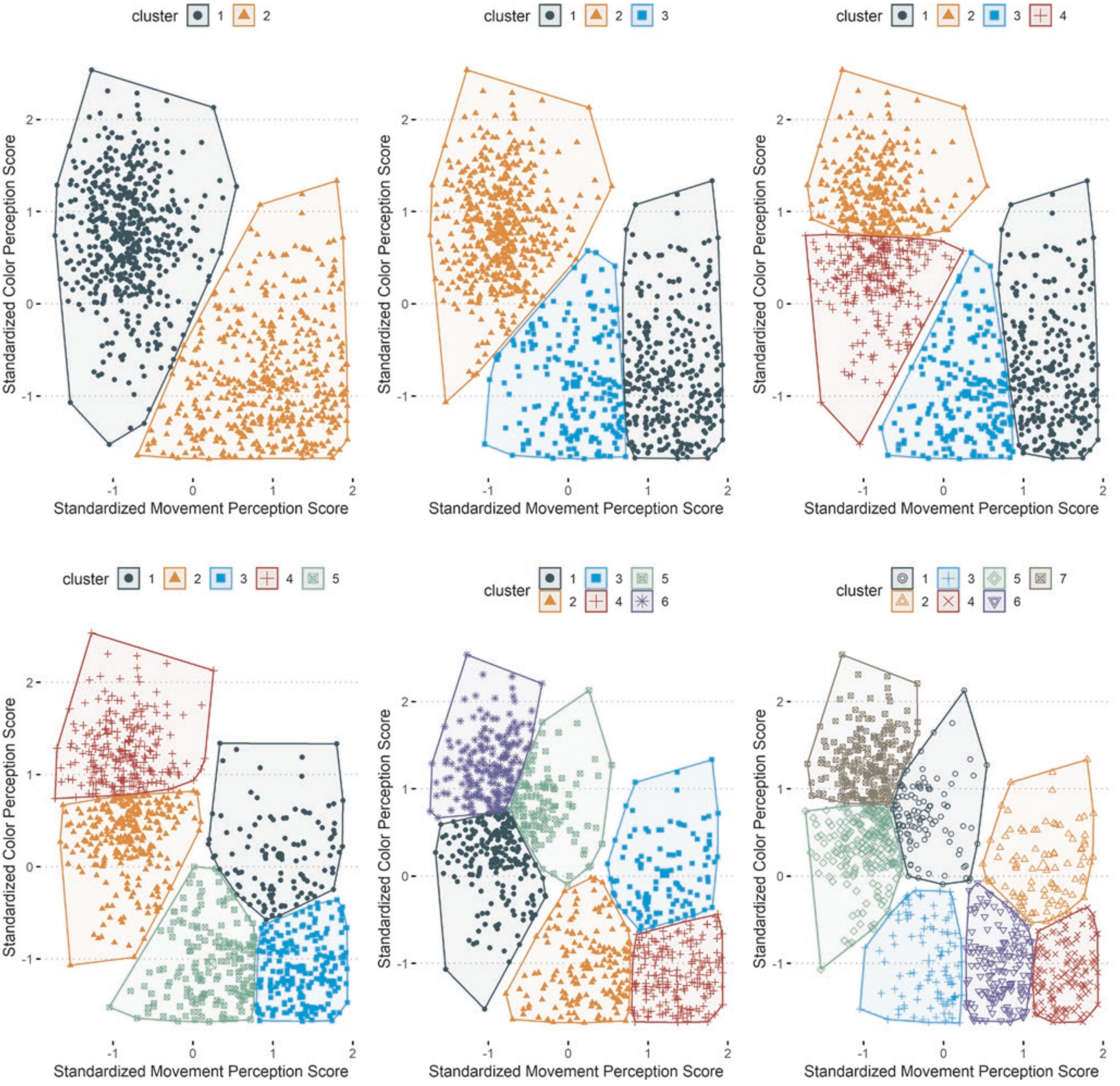

Fig. 12.7 Optimal number of clusters by method

one is highly dense and the other is not, density-based models will have difficulty recognizing them.

- Given some combinations of both parameters in the algorithm, irrelevant tiny clusters might appear.
- It requires the most user supervision of all the algorithms, as the results are highly unstable based on different combinations of parameters.

12.5 Dimensionality Reduction

Until now, all of our examples have been based on two dimensions, *x*- and *y*-axis values. However, in real-life scenarios, it is unlikely that setting investigated has only two. Most problems in clinical science appear within incredibly complex causal networks. Patients, their diseases, and realities are highly dimensional. We have highlighted that clustering algorithms tend to fail when the number of dimensions increases because distance-based metrics tend to be meaningless at high values. The response to this phenomena: to reduce dimensions of your data.

Dimension reduction is the task that transforms high dimensions of data to low dimensions while conserving the most important relations and features of the original. There is an almost infinity of ways to achieve such a purpose, from principal component analysis to uniform manifold approximation and projection algorithms. Let us demonstrate this with another toy example.

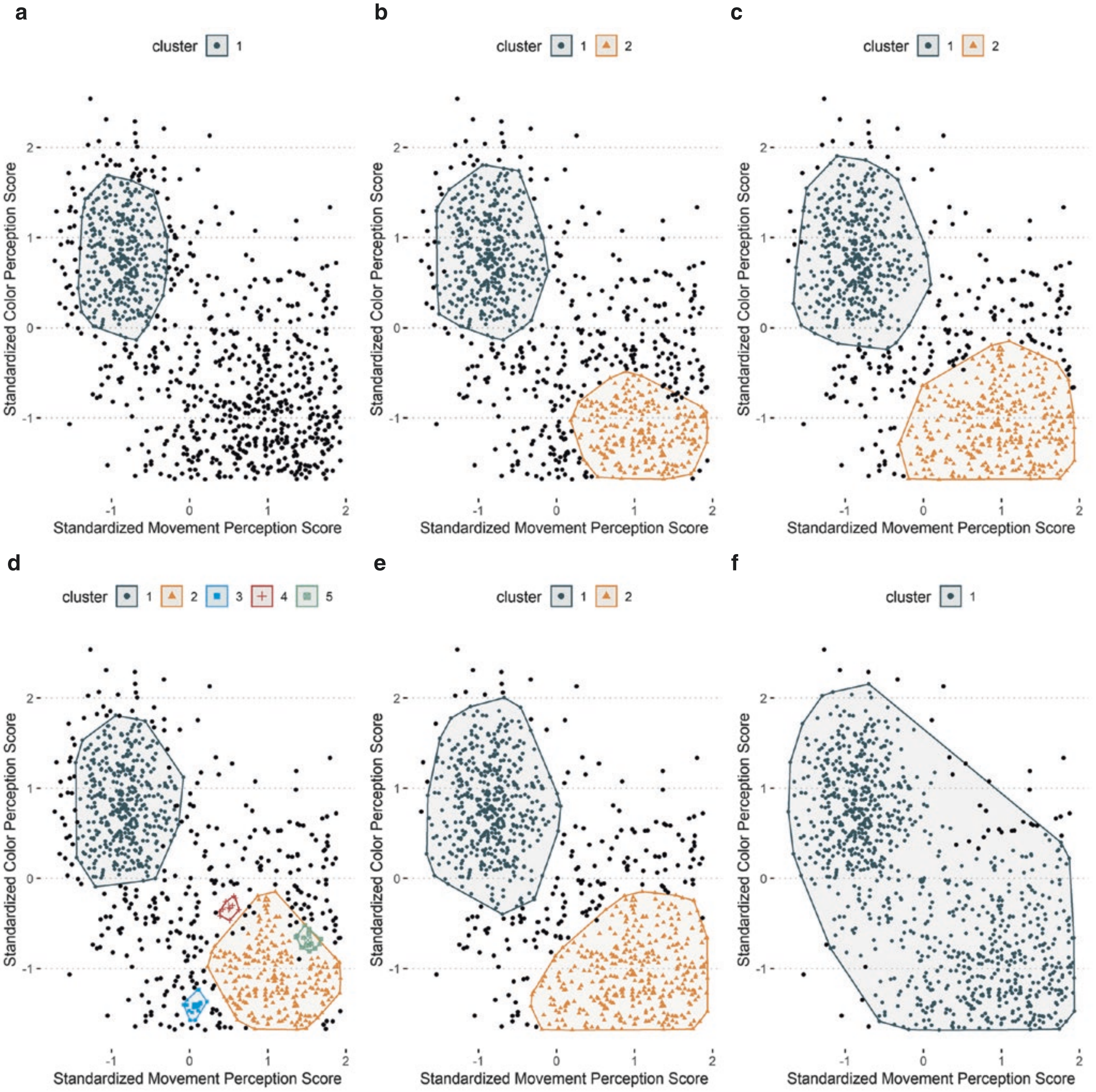

Fig. 12.8 DBSCAN clustering results with varying parameters. *Notes:* (**a**) $e = 0.25$, $minPts = 40$, (**b**) $e = 0.25$, $minPts = 30$, (**c**) $e = 0.25$, $minPts = 20$, (**d**) $e = 0.15$, $minPts = 10$, (**e**) $e = 0.30$, $minPts = 30$, (**f**) $e = 0.35$, $minPts = 30$

We have now performed six additional cognitive tasks on our imaginary sample, resulting in eight variables. However, we want to describe the sample with as little complexity as possible, let us say three components maximum. The first step is to compute the principal components of the dataset. To do so, the covariance matrix of the data has to be estimated, and the eigenvalues and eigenvectors are factored in to diagonalize the elements that form the variance of each respective dimensions. The proportion of explained variance that each eigenvector reflects is calculated by dividing the eigenvalue by the addition of each eigenvector. In our case, the first component explains 36% of the data variance, the second 20%, and the third around 12%. This means that by using the first three components we are resembling 68% of the original dataset, with of 3 out 8 dimensions, or 37.5% of the original data. Figure 12.9 shows the graphical presentation.

We cut the dimensions to three, and now we apply again the optimized hierarchical clustering algorithm of the

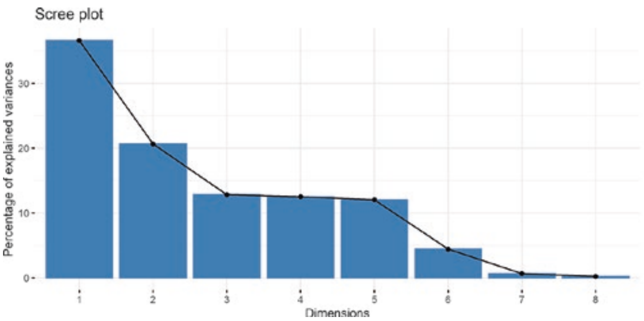

Fig. 12.9 Variance explained by each dimension

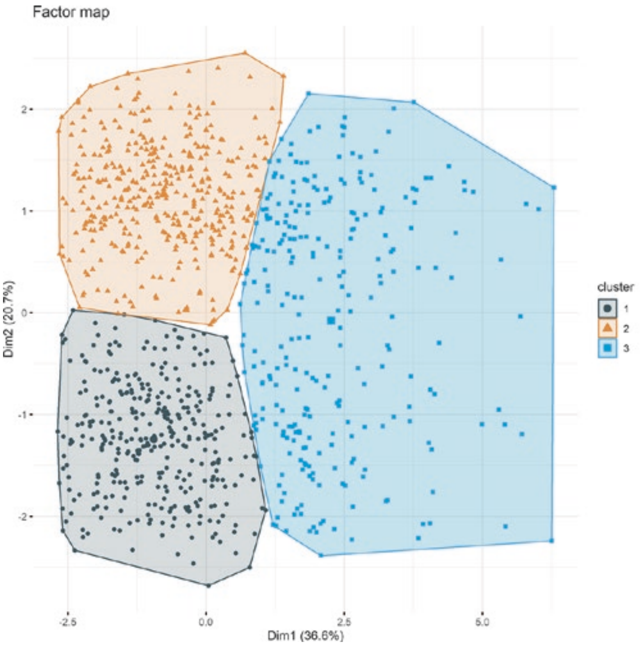

Fig. 12.10 Optimized hierarchical clusters of the simplified dataset

beginning of this chapter resulting in three clusters presented in Fig. 12.10.

12.6 Applications

Adult Spinal Deformity

We previously published one paper, in 2019 [16], using the methods described in this chapter. We did it in the adult spinal deformity (ASD) field. ASD, also known as scoliosis of the adult, is a highly heterogeneous and debilitating condition. Its defining feature is a physical deformation of the spine mainly measured in key angles of its shape. Up to that point, the available classifications of the disease were mainly based in X-ray measurements of the Spine, the Schwab [17] and Lenke [18] classifications. And, while it is true that the spine is a complex structure that entails a lot of features, to us, ignoring non-spine specific patient parameters seemed like an incomplete model of the disease process.

Using a combined data query from both the European Spine Study Group (ESSG) and the International Spine Study Group (ISSG) we set up to simply describe and characterize the potential latent patient clusters. Adding simple quality of life and demographic metrics, we performed a hierarchical clustering modelling to group similar patients from dissimilar ones. We found an optimal of three types of patients, we called them, young coronal patients, old first-timers, and old-revisions. The main descriptive characteristics of the groups were: young coronal patients typified by much younger patients with a coronal spinal deformity and little sagittal malalignment. Old first-timers were patients mostly in their late 50s or early 60s with a more severe deformity mostly related to the lumbar spine and with no previous spinal surgery. Finally, old revision patients were the oldest and the ones with the most severe malalignment, especially in the sagittal plane and who had undergone prior spinal surgery.

However, to us the task seemed incomplete, and on top of a patient-specific clustering exercise we also applied it to surgical techniques. The surgical treatment of scoliosis involves a wide variety of different techniques. The termination levels of fusions and placement of nerve decompressions and vertebral releases and osteotomies result in significant treatment heterogeneity. When we clustered the range of surgical treatments, we found four types of surgeries to be the main clusters.

Finally, by superimposing both the patient and surgery classification, we obtained a descriptive grid of patient and surgery heterogeneity. By doing so, we were able to look at what happened 2 years after surgery when patients within a same cluster where operated on by different surgical clusters. What we obtained was not only a descriptive result in terms of clusters, but also a simple prognostic model associating types of patients and surgeries to outcomes. We observed that, for instance, young coronal patients were the ones with the lowest functional and quality of life improvement, on average, while those young coronal patients receiving more aggressive surgeries were also experiencing higher levels of post-surgical complications. This allowed us to identify simple areas of improvement in terms of patient and surgical selection with a cost-benefit that might not justify more aggressive surgeries.

Sepsis

One of our favorite examples of a successful application of *k*-means clustering is an article by Seymour and coauthors published in 2019 in JAMA [19]. They developed and validated clinical phenotypes for sepsis and model the potential benefit and harm of treatments with data from an external randomized controlled trial.

Sepsis is highly heterogeneous condition defined by an unregulated immune response to an infection that leads to acute organ failure. Given the multidimensional array of clinical symptoms and biological features, the authors used a variety of variables that ranged from demographic, vital signs, markers of inflammation, to markers of organ dysfunction. Out of more than 50 potential candidate variables a total of 29 were selected and observations were clustered according to the consensus *k*-means clustering method. A total of four phenotypes, alpha, beta, gamma, and omega were found to be optimal using diverse measures of optimization. When looking at the outcomes at any time during hospitalization they found startling differences. The first phenotype, alpha, had only a 2% in-hospital mortality with only 25% of patients admitted to the ICU, while the last cluster or phenotype had a 32% in-hospital mortality and 85% ICU admissions rates. Compared to the standard classification of sepsis, the proposed "phenotypic" classification was fairly constant across the other scheme, highlighting that the human-proposed classification did not capture relevant variance.

The authors then, go a step further, analogous to the above-mentioned cost/benefit grid. They used external RCT data to estimate differential treatment effects across phenotypes. First, they assign the observations of three RCTs (ACCESS, PROWESS, and ProCESS) to each of their derived clusters. After, they vary the proportion of patients from each cluster in each trial to simulate scenarios and their causal effects. They find that out of the three interventions used in the RCTs, according to which phenotype they are applied, the effects varied remarkably: from total benefit to an extremely high likelihood of harm.

Common Pitfalls and Proposed Solutions

The three most common pitfalls in clustering research relate to (a) the use of high-dimensional data, (b) the lack of comparison of results across clustering methods, and (c) determining whether the results are meaningful. Geometry behaves irregularly in high-dimensional settings, hence measures of distance are rendered non-useful. Sparsity and the identification of relevant variables in the problem tend to be hidden under large numbers of irrelevant ones. We recommend to thoroughly inspect data in the pre-implementation stage and to make sure that each included feature has a potential meaningful implication. As we have discussed in this chapter, different methods can produce different results, hence judging one clustering configuration without comparing it to potential others can render the external validity of the results null. We recommend applying, at least, three different optimized algorithms to assess the robustness of the results. The determination of the usefulness of the results is perhaps the most crucial part, and where we researchers tend to use follow-up data or third-party linked results. It is imperative to pair any good clustering exercise with expert knowledge on the underlying data-generating process.

12.7 Conclusions

Any clustering task involves investigator-related choices, and many of them are critical to the validity of results, both internally and externally. In the present chapter we have introduced, with examples, a few of the most relevant unsupervised learning techniques for the practicing clinical neuroscience researcher. We have not extensively covered all potential algorithms or methods, as that would require a series of books in itself, but we have provided a few visual examples and applications that we hope successfully aid other researchers in the use of these tools. Moreover, the full capacities of data will only be achieved if everyone learns to pair the right research question with the appropriate tools. Clustering methods are the most important tool for data discovery and description, and its integration with both predictive and causal objectives is crucial to maximize its potential, as alone, it still is a descriptive method.

Our experience reveals that the advantages of using formal unsupervised learning algorithms are superior to standard supervised classification methods for the description of phenotypes or clusters. Not only that, but given their potential for heterogeneous treatment effects, they will be a cornerstone for trial design by selecting populations with expected effect sizes well below or above the mean.

In short, clustering is perhaps, more than other machine learning techniques, the most underused and underappreciated, and should be strongly considered in questioning scientific paradigms regarding classification of features.

For further reading we recommend the books by Trevor Hastie, Robert Tibshirani & Jerome Friedman. "The elements of statistical learning: data mining, inference, and prediction" Springer Science & Business Media, 2009, and M. Emre Celebi & Kemal Aydin. "Unsupervise learning algorithms" Berlin: Springer International Publishing, 2016.

Conflict of Interest Declaration Dr. Serra-Burriel reports receiving grant funding from the European Commission H2020 program and European Commission EiT Health program.

Dr. Ames reports receiving royalties from Stryker, Biomet Zimmer Spine, DePuy Synthes, NuVasive, Next Orthosurgical, K2M, and Medicrea; being a consultant to DePuy Synthes, Medtronic, Medicrea, and K2M; receiving research support from Titan Spine, DePuy Synthes, and ISSG; being on the editorial board of Operative Neurosurgery;

receiving grant funding from SRS; being on the executive committee of ISSG; and being a director of Global Spine Analytics.

None in relation to the present work.

References

1. Storrs KR, Fleming RW. Unsupervised learning predicts human perception and misperception of gloss. bioRxiv. 2020. https://doi.org/10.1101/2020.04.07.026120.
2. Driver HE, Kroeber AL. Quantitative expression of cultural relationships. Berkeley: University of California Press; 1932.
3. Sánchez-Hernández G, Chiclana F, Agell N, Aguado JC. Ranking and selection of unsupervised learning marketing segmentation. Knowl Based Syst. 2013;44:20–33.
4. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16:321–32. https://doi.org/10.1038/nrg3920.
5. Denny M, Spirling A. Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. Polit Anal. 2017;26(2):168–89.
6. Wang L. Discovering phase transitions with unsupervised learning. Phys Rev B. 2016;94:195105.
7. Sonnewald M, Dutkiewicz S, Hill C, Forget G. Elucidating ecological complexity: unsupervised learning determines global marine eco-provinces. Sci Adv. 2020;6:eaay4740.
8. Syakur MA, Khotimah BK, Rochman EMS, Satoto BD. Integration K-means clustering method and elbow method for identification of the best customer profile cluster. In: IOP conference series: materials science and engineering. 2018.
9. Kodinariya TM, Makwana PR. Review on determining number of cluster in K-means clustering. Int J Adv Res Comput Sci Manag Stud. 2013;1:90–5.
10. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc Ser B Stat Methodol. 2001;63:411–23.
11. Fichet B, Piccolo D, Verde R, Vichi M. Studies in classification, data analysis, and knowledge organization. In: Knowledge organization. 2011.
12. Lloyd S. Least squares quantization in PCM. IEEE Trans Inf Theory. 1982;28:129–37.
13. MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1: statistics. Berkeley: University of California Press; 1967. p. 281–97. https://projecteuclid.org/euclid.bsmsp/1200512992.
14. Hartigan JA, Wong MA. Algorithm AS 136: a k-means clustering algorithm. J R Stat Soc Ser C (Appl Stat). 1979;28:100–8.
15. Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd. 1996. p. 226–31.
16. Ames CP, Smith JS, Pellisé F, Kelly M, Alanay A, Acaroğlu E, et al. Artificial intelligence based hierarchical clustering of patient types and intervention categories in adult spinal deformity surgery: towards a new classification scheme that predicts quality and value. Spine (Phila Pa 1976). 2019;44:915–26.
17. Terran J, Schwab F, Shaffrey CI, Smith JS, Devos P, Ames CP, et al. The SRS-Schwab adult spinal deformity classification: assessment and clinical correlations based on a prospective operative and nonoperative cohort. Neurosurgery. 2013;73(4):559–68.
18. Lenke LG. The Lenke classification system of operative adolescent idiopathic scoliosis. Neurosurg Clin N Am. 2007;18(2):199–206.
19. Seymour CW, Kennedy JN, Wang S, Chang C-CH, Elliott CF, Xu Z, et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. JAMA. 2019;321:2003–17. https://doi.org/10.1001/jama.2019.5791.