SPOTLIGHT ON CLUSTERING & UNSUPERVISED LEARNING

Jen Schellinck[1,4,5,6,7], Patrick Boily[1,2,3]

Abstract
Supervised learning methods can be presented in a formalism which generalizes statistical and regression analysis, and their performance are easy to evaluate; consequently, they have been studied extensively and often form the backbone of machine learning training. On the other hand, apart from a select few classical models, unsupervised learning tasks are not usually presented with quite the same depth, primarily due to the vagueness which infect their core – a number of the important concepts are ambiguously defined; the validation of the results is sometimes elusive, and the actionable applications of the outcomes are not always clear. The interest in such methods and tasks (clustering and segmentation, association rules mining, link profiling, etc.) is mounting, however, with the increased focus on artificial intelligence and machine learning research. In this document, we describe some of the most commonly-used clustering algorithms, and discuss related issues and challenges.

Keywords
Clustering, k—means, hierarchical clustering, DBSCAN, spectral clustering, expectation-maximization, affinity propagation, fuzzy clustering, cluster ensembles, validation, model selection.

Contributions
Aditya Maheshwari has also contributed to part of this report.

[1] Data Action Lab, Ottawa, Canada
[2] Department of Mathematics and Statistics, University of Ottawa, Ottawa, Canada
[3] Idlewyld Analytics and Consulting Services, Wakefield, Canada
[4] Sysabee, Ottawa, Canada
[5] Institute of Cognitive Science, Carleton University, Ottawa, Canada
[6] AI Guides, Ottawa, Canada
[7] Think Digital, Ottawa, Canada
Email: pboily@uottawa.ca

Contents

1 Overview 1
1.1 Unsupervised Learning . . . 1
1.2 Clustering in General . . . 2
1.3 A Philosophical Approach to Clustering . . . 3
1.4 Case Studies . . . 5
1.5 Outline . . . 6

2 Simple Clustering Algorithms 7
2.1 k-Means and Variants . . . 7
2.2 Hierarchical Clustering . . . 9

3 Evaluation 13
3.1 Clustering Assessment . . . 13
3.2 Model Selection . . . 24

4 Advanced Topics 26
4.1 Density-Based Clustering . . . 26
4.2 Spectral Clustering . . . 30
4.3 Probability-Based Clustering . . . 41
4.4 Affinity Propagation . . . 45
4.5 Fuzzy Clustering . . . 48
4.6 Cluster Ensembles . . . 50

1. Overview

We introduced some of the basic notions of unsupervised learning in *Machine Learning 101* [10]; here, we review some of these concepts in the context of clustering, discuss the problems of validation and model selection, and present some simple and sophisticated clustering algorithms.

1.1 Unsupervised Learning

In supervised learning, we differentiate between a dataset's **response variables** $Y_1,\ldots,Y_m$ and its **predictor variables** $X_1,\ldots,X_p$. Which variables are predictors and which are responses depend on the context – for some questions, a given variable could be a predictor, for others, a response.

Unsupervised learning tasks do away with the responses altogether, which means that **prediction** is off the table; note that variables that would have been deemed response variables in a supervised learning framework are not necessarily removed from the dataset during the analysis – they are simply not viewed as an outcome to predict, and the predictor variables are just observation **features**.

In unsupervised learning, the objective is to **identify** and **uncover interesting insights** about the dataset and the system that it represents [11,25], such as:

- informative ways of visualizing the dataset (often associated with **dimension reduction** [31]);
- highlighting subgroups among the dataset's variables or observations (clustering), or
- finding links between variables (association rules mining, link profiling, etc.), say.

1.2 Clustering in General

Clustering consists of a large family of algorithms and methods used to discover so-called **latent groups** in the datasets – natural groups that exist but have not been identified or labeled as such.

Clustering is a **subjective** analytical task; unlike classification and regression, clustering analysis does not have as "simple" a goal as predicting a response for a new observation based on historical data patterns, and there is no "solution key" against which to compare analysis results.

Applications:

- finding subgroups of breast and/or prostate cancer patients based on their gene expression measurements or their socio-demographic characteristics in order to better understand the disease and potential treatment side-effects;
- grouping products in an online shop based on ratings and reviews assigned by customers, or grouping customers based on their purchasing history, in order to make product recommendations;
- finding documents that apply to search queries, and finding similar queries to those entered by a user to increase the odds of finding the documents they are really looking for;
- identifying population segments to test various incentives for vaccination;
- etc.

In each of these cases, the **number** of these latent groups is unknown (and can in fact be taken as a true unknown of the problem).

The **subjectivity** of unsupervised learning tasks may seem to be an insurmountable flaw: analysts attempting to find latent groups in a dataset, say, may obtain a different number of such groups, or assign different observations to their groups if their numbers are identical, without one of them being necessarily "wrong" (although it is conceivable that some of them could produce **sub-optimal** groups, see Section 3 for a detailed discussion on this topic).

In spite of this, clustering is a popular analytical task, in part because it is much easier (and cheaper), typically, to obtain **unlabeled data** than it is to obtain labeled data (against which supervised methods could be evaluated).

A **cluster** is a subset of observations that all have something in common – they are **similar**, according to some measure of similarity. Furthermore, a cluster's observations should be **dissimilar** to other clusters' observations.

Clusters do not necessarily need to be distinct (as in so-called **hard** clustering) – in some cases, it might be sufficient to quantify the likelihood or the degree to which an observation belongs to a cluster (**soft** clustering).

The choice of a **similarity/dissimilarity measure** is entirely **subjective**; there are contexts for which **proximity** could be used as a decent proxy for similarity, and others where it could not. Even in the former case, a **distance measure** (metric) has to be selected, and infinitely many choices are available to analysts.

Without **domain-specific considerations** (this requires thorough data and context understanding), the choice of measure is arbitrary; but understanding the data and the context does not guarantee that all reasonable analysts would agree on such a measure.

For instance, in any group of human beings, which of

> age, ethnic background, gender, postal code, sexual orientation, linguistic abilities, mathematical skills, career, social class, political affiliation, operating system preferences, educational achievements, hockey club fandom, etc.

is responsible for separating its members into "us" vs. "them" groups? Is it some combination of these characteristics? Are the groups fixed? Is everybody in the "us" group based on age also in the "us" group based on "gender"?

We could bypass the problem by creating more groups; given an age group and gender, we could create the clusters: "same age group and gender" (us), "same age group, different gender" (them$_1$), "different age group, same gender' (them$_2$)', "different age group and gender" (them$_3$).

It is clear how the process can be expanded to include more combinations of feature levels, but at the price of introducing an ever increasing number of clusters – how many "them" groups are too many for analysts or human brains to process?

Clustering algorithms are designed to try to model various aspects of this problem, but the latter's complexity gives rise to an enormous number of algorithms: at least 100 have been published, as of January 2022 [53]. Most of these belong to one of six main families [2]:

- **partitional** (k−means and variants, CLARA, etc.);
- **hierarchical** (AGNES, DIANA, BIRCH, etc.);
- **density-based** (DBSCAN, DENCLUE, OPTICS, etc.);
- **connectivity-based** (spectral and variants, etc.);
- **grid-based** (GRIDCLUS, STING and variants, etc.);
- **model-based** (mixture models, latent Dirichlet allocation, expectation-maximization, etc.).

As is the case for all analytical methods, some modifications are required when dealing with "**Big Datasets**", for **high-dimensional data**, or for specific **types of datasets**, such as stream data, network data, categorical data, text and multimedia data, time series data, and so on. **Ensemble methods**, which combine various clustering results, can also prove useful.

Distance, Similarity, and Dissimilarity Measures Although the choice of how to interpret and compute **similarity** between observations is, to all intents and purposes, completely up to the analysts, all such measures must satisfy certain properties: they must take on

- large values for similar objects, and
- small (or even negative) values for dissimilar objects.

Dissimilarity measures function in the opposite manner.

The **kernel functions**[1] of machine learning [9] are examples of similarity (or dissimilarity) measures, most notably the **Gaussian** (or radial) kernel

$$K_\gamma(\mathbf{x},\mathbf{y}) = \exp(-\gamma\|\mathbf{x}-\mathbf{y}\|_2^2),$$

for a given $\gamma > 0$, for which points near one another (in the $\|\cdot\|_2$ sense) have a similarity measure $w = K(\mathbf{x},\mathbf{y}) \approx 1$ (and thus are **similar**), and points far from one another have a similarity measure near 0 (and thus are dissimilar).

Some similarity measures are derived from **distance (metrics)** $d : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_0^+$, which are functions with special properties:

1. $d(\mathbf{x},\mathbf{y}) = 0 \iff \mathbf{x} = \mathbf{y}$;
2. $d(\mathbf{x},\mathbf{y}) \geq 0$ for all $\mathbf{x},\mathbf{y} \in \mathbb{R}^p$;
3. $d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x})$ for all $\mathbf{x},\mathbf{y} \in \mathbb{R}^p$;
4. $d(\mathbf{x},\mathbf{y}) \leq d(\mathbf{x},\mathbf{z}) + d(\mathbf{z},\mathbf{y})$ for all $\mathbf{x},\mathbf{y} \in \mathbb{R}^p$.

In effect, distances are positive-definite symmetric functions $\mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_0^+$ satisfying the **Triangle Inequality**. Commonly used distances include the:

- **Euclidean** distance $d_2(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_2$;
- **Manhattan** distance $d_1(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_1$;
- **supremum** distance $d_\infty(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_\infty$;
- more general **Minkowski** distance $d_p(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_p$, for $p \geq 1$, of which the first three examples are special cases;
- and more esoteric distances such as the **Jaccard** distance for binary vectors, the **Hamming** distance for categorical vectors, the **Canberra** distance for ranked lists, the **cosine** distance for text data, **mixed** distances for mixed variables, and so on [10, 12, 18, 21].

[1]Formally, a **kernel** is a symmetric (semi-)positive definite operator $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_0^+$. By analogy with positive definite square matrices, this means that $\sum_{i,j=1}^N c_i c_j K(\mathbf{x}_i,\mathbf{x}_j) \geq 0$ for all $\mathbf{x}_i \in \mathbb{R}^p$ and all $c_j \in \mathbb{R}_+$, and $K(\mathbf{x},\mathbf{w}) = K(\mathbf{w},\mathbf{x})$ for all $\mathbf{x},\mathbf{w} \in \mathbb{R}^p$.

Given a distance d, a common construction is to define the associated similarity measures

$$w = \ell - d, \quad w = \exp(-kd^2), \quad \text{or} \quad w = \frac{1}{1+d}.$$

Note that there are similarity measures that cannot be derived from distance measures, however.

Data Transformations Prior to Clustering Prior to clustering, it is crucial that the data be scaled (and potentially centered) so that none of the variables unduly influence the outcomes, or, as the expression prosaically puts it, so that we do not have to compare apples with oranges – if age in years and height in cm are dataset variables, a 10-unit difference in age is likely to be more significant (in real terms) than a 10-unit difference in height.

Putting everything on a $(\min, \max)$ scale, for instance, guarantees that relative differences (relative to the distributions of each variables), and not absolute distances, play the important role. However, there are many ways to scale the data, and the scaling approach may have an effect on the clustering results (as we are sure you will not be surprised to find out, by this point – that is the way, with clustering: out of the frying pan and into the fire).

Common Difficulties There are issues related to clustering other than the vagueness we have already discussed:

- in many instances, the underlying assumption is that **nearness of observations** (in whatever metric) is linked with **object similarity**, and that **large distances** are linked with **dissimilarity**;
- the **lack of a clear-cut definition** of what a cluster actually is makes it difficult to validate clustering results;
- various clustering algorithms are **non-deterministic**;
- the number of clusters cannot usually be known before the analysis;
- even when a cluster scheme has been accepted as valid, a **cluster description** might be difficult to come by;
- most methods will find clusters in the data even if there are none;
- once clusters have been found, it is tempting to try to "explain" them, but that is a job for domain experts.

1.3 A Philosophical Approach to Clustering

In the context of artificial general intelligence,[2] clustering provides a basic way for **intelligences** to structure their experience of the world.

Clustering techniques can allow such machines to identify **object instances** in the world around them and then, on the basis of this identification, to identify or define **types of objects** by grouping together the object instances they have discovered.

[2]Think free-ranging robots, roughly speaking.

With this in mind, we can view creating **concepts** as the fundamental purpose of identifying groupings of similar datapoints; these concepts allow an intelligent agent (whether machine or person) to:

- work in **shorthand** when dealing with objects (i.e., it is easier to deal with 10 'cats' than with 10 unique objects), and
- make assumptions about the object instances in a cluster associated with the concept (if an object is a cat, then that object probably likes fish).

If the existence of some "**ground truth**" about what **should** be clustered together (and by extension what should be counted as a concept) can be presupposed, then regardless of what is currently known about that truth, neither the choice of algorithms nor of clustering algorithm parameters is wholly **subjective**:[3] choosing one algorithm over another (or one set of parameters over another) may lead to a "better" or "worse" reflection of the underlying ground truth.

But what counts as a ground truth? There are, of course, debates about this in philosophical circles. Suppose that **natural kinds** exist, that is to say, suppose that there is a **privileged** and **objectively essential** way in which objects are **properly grouped** in nature.

This assumption is very commonly made in the sciences, where uncovering or discovering **universal truths** about natural kinds of objects is a major objective. In such a case, natural kinds can count as a ground truth, with some clusterings more closely reflecting this reality than others.

The fact that the ground truth is not known by the clustering agent does not mean that it does not exist, or that it cannot be sought out using various techniques.

This is arguably what scientists do when they are using the **scientific method**; they do not know, *a priori*, which of their hypotheses are true or which are false, but they nonetheless engage in various techniques to try to get a better sense of what is true and false.

Even if the existence of natural kinds is rejected, it can still be the case that, relative to a particular circumstance, some clustering results are of **higher quality** than others. Or, stated in terms of goals, some clustering results could achieve a stated goal to a greater or lesser extent than others.

This does appear to be more subjective, in the sense that the goal, and the success of the outcome relative to the goal, are both defined by an individual or individuals, rather than being **independent** of them.

Outside of clustering, it is not unusual for people to create **contextual definitions** of what counts as 'good'. Consider as an example the concept of a 'good meal'. What qualifies as a good meal when camping in the backcountry is not the same as what counts as a good meal when staying at a four-star resort. Context matters.

But this does not mean that it is impossible to have a bad meal under either of these circumstances, or that just anything counts as a good meal – recall that we do not use subjective in the stultifying post-modernist sense that there are no constraints whatsoever and everything is a social construct.[4]

Nonetheless, such situations are difficult to pin down or define in a rigorous fashion. Even if there were some more abstract or subjective sense in which something could be said to be **common to all types** of good meals – or to all types of good clustering results, in this analogy – it is difficult to imagine how this could be stated with any precision. This is, frustratingly, a typically human approach to dealing with the world.

However, since the underlying objective of machine learning and artificial intelligence is to create machines with abilities similar to those of humans, perhaps it is worth trying to capture this less than rigorous approach within the context of machine learning.

Given this inherent lack of rigour, are there applied situations where clustering is **useful**? More concretely, suppose we desire to cluster furniture, based on data about the furniture. We could make **measurements** of various kinds on physical objects, either selected randomly or haphazardly; perhaps, rather than working with the furniture itself, we could use a website catalogue in which each page showcases a particular type of furniture available for sale.

In this scenario, there may be one grouping (created by tagging and linking pages, for instance) of the furniture pages that allows users to **visit the website with maximum efficiency**, and another that helps the store **maximize its sales**.

Moreover, if we believe that natural kinds exist (which, as noted, is debated by philosophers but is a common assumption in science), there might also be one grouping of the furniture that best matches the underlying furniture natural kinds..

When considering outcomes relative to a particular situation, the most appropriate strategy for a particular clustering will depend, broadly speaking, on two considerations:

- the **chosen goal** it is intended to support, and
- the **underlying structure** of the data itself.

[3] In the psychological sense of the term, where it has the connotation of "coming from a person's experience"; philosophically it tends to indicate that whatever it is that is being talked about does not exist separate from such an experience.

[4] This might be a smidgen of a straw-man definition of "post-modernist subjectivity", but not that much of one; all things being equal, we lean more toward the objective side of things, in nature and in thoughts.

The first can be explicitly known and stated, but the second will likely not be known in advance, which leads to numerous **technical issues** when applying clustering algorithms.

A multitude of clustering algorithms can be applied to problems like the website furniture problems described above, and for each of those, many different parameter settings exist. Suppose six different clustering process are carried out in the case of the furniture website example and they generate six different clustering outcomes.

Presumably, some will be more effective than others, if the objective is to get people to spend a maximum amount of money on the online store. If the objective is to allow customers to make their purchases most quickly, the "optimal" clustering might not be the one that leads customers to spend the most money.

It is difficult to say ahead of time which of the six groupings will be the most effective one, in each of these cases. However, it might be possible to carry out **A/B testing** to determine which one is the most effective, once they have been generated. But can the A/B testing step be avoided?

If the applied goal (e.g. the goal to group furniture pages in order to maximize profits) can be operationalized more directly in terms of similarity and difference, then it can be more directly tied to clustering approaches.

If we think that the best way to increase sales is to have loose clusters where people are forced to browse a certain amount, while being exposed to somewhat similar (but still interesting) pieces of furniture to find what they want, it might be possible to select a clustering approach to generate clusters with this desirable property.

Returning for a moment to the less applied general artificial intelligence scenario introduced earlier, if the fundamental functionality of clustering is viewed as creating concepts, then it would seem to make sense to operationalize this goal in terms of creating groupings where the **observations** in a group are similar to each other and different from those in other groups.

In this context, a poor grouping would *de facto* be one where multiple observations are similar to those in other clusters, or very different from those in their own cluster.

This could happen if a clustering process runs into technical difficulties, but it can also happen if there is no such **strongly grouped structure** in the data itself.

To eliminate the possibility that the problem is not linked with the chosen clustering procedure, one strategy is to use **multiple clustering techniques**, as well as **multiple parameter settings** for each technique.

If the issue remains, then we could conclude that it is likely that there is no good clustering structure in the data and by extension, in the objects being represented by the data.

1.4 Case Studies

Interested readers can get more information on clustering, as well as examples of applications, in [1, 2, 5, 10, 13, 15, 20, 22, 24–26, 28, 33, 35–37, 39, 41–44, 46, 52, 53]. In the rest of this section, we will present a few case studies that showcase the range of possible clustering applications.

Case Study 1: Environmental Studies. In this case study [44] ,an affinity propagation (AP) algorithm was applied to find similar characteristics in emissions among 30 provinces in China. The clustering results of CO2 emissions showed that the 30 provinces were divided into five clusters in 1997 and seven clusters in 2012 based on four indicators (generation structure, energy intensity, GDP per capita and electricity intensity). The conclusion was that provincial emissions reduction target and supporting policies for power industry should be customised and consistent with the actual situations considering the similarity and differences in emission characteristics.

Case Study 2: Medical Diagnosis. In this study [52], a semi-supervised affinity propagation algorithm was used to cluster ECG beats in order to detect arrhythmias. The authors found that the resulting clusters exhibited a high degree of precision with respect to grouping ECGs displaying arrhythmias.

Case Study 3: Object Recognition. When working to detect moving objects in videos, it can prove difficult to separate objects of interest from background images and noise (e.g., separate an imaging of a moving car from a park scene with leaves rustling in the background) because the lighting and other video parameters tends to vary from video to video. As a result of these challenges, the process is difficult to automate – typically, a person must tune the object identification algorithm (e.g. MSRM) before it can be applied. In this study [35], researchers used a combination of clustering algorithms, including EM clustering (which they refer to as collaborative clustering) to bypass this step. The collaborative clustering strategy successfully identified 'rough' segments, which could then be used to bootstrap MSRM without human intervention.

Case Study 4: Medical Diagnosis. Mild cognitive impairments (MCI) are a known to be a risk for factor for development of Alzheimer's Disease. MCI are accompanied by changes in brain structure. But which changes indicate that people will go on to develop Alzheimer's? In this study [39], a number of different data science techniques were applied to MRI data to investigate this question: Support Vector Machines, Bayesian Statistics, Voting Feature Intervals, Feature Extraction and (last but not least) DBSCAN. DBSCAN was used once voxels that provide high information about the classification of the image were identified using entropy based measures. DBSCAN then grouped pixels with similar spatial and information levels to determine which parts of the brain are the most important for the diagnosis.

Case Study 5: Disaster Relief. This research paper [43] presented a hybrid fuzzy clustering-optimization approach to emergency logistics distribution during disasters. The approach was grounded using an existing emergency logistics co-distribution conceptual framework. The proposed methodology involved two mechanisms: disaster-affected area grouping, and relief co-distribution. The approach was validated using numerical studies based on data collected during a large-scale earthquake occurring in Taiwan.

Case Study 6: Urban Planning. This study [24] used a fuzzy c—means (FCM) algorithm to identify housing submarkets in the Buffalo-Niagara Falls region. The study focused on refining approaches to selecting appropriate parameters of fuzzy clustering and characterizing the relationship between the clusters produced. Clustering results were validated in terms of hedonic prediction accuracy (prediction of the demand for houses by house purchasers).

Case Study 7: Food Science. In this study [46], clustering was used to investigate coffee preferences. First, a group of taste testers (a panel) were divided into four preference clusters, based on their ratings of 12 regular coffee (RC) samples with various blend ratios of coffee beans. Then, the taste testers tried 88 additional RC samples. Models were created to predict the preference scores of the new samples for each clustered group, using a fuzzy neural network. At the same time a genetic algorithm was used to predict the optimum blends for each cluster. These predicted optimum blends were then also tested on each cluster. The results of this were consistent with what the models predicted and with how the tasters in each cluster rated the new blends. The researchers thus suggest that this approach could be used for the development of coffee products.

Case Study 8: Traffic Safety. This study [33] investigated analysis strategies for identifying black spots – locations where traffic accidents frequently occur – using a dataset of accidents occurring in Denizli, Türkiye. First, fuzzy clustering methods were used to identify and define black spots based on accident density. Then, the safety levels of black spots were determined using a Shannon Entropy approach based on accident types and effective factors related to accident occurrence. Finally, the safety levels of accident locations were classified using both fuzzy logic and crisp approaches to classification, based on the calculated entropy values. The results of the analysis were used to create a series of recommendations to improve traffic safety.

Case Study 9: The Livehoods Project. When we think of similarity at the urban level, we typically think in terms of neighbourhoods. Is there some other way to identify similar parts of a city? In this study [13], the authors study the social dynamics of urban living spaces with the help of clustering algorithms. The researchers aims to draw the boundaries of livehoods, areas of similar character within a city, by using spectral clustering. Unlike static administrative neighborhoods, the livehoods are defined based on the habits of people who live there (Pittsburgh, PA). In total, 9 livehoods have been identified and validated by 27 Pittsburgh residents.

Case Study 10: Comparative Mythology. Studying myths from different cultures can help us understand their similarities and possibly shared origins, as many myths have splintered off and evolved from common sources. In this study [15], the author uses hierarchical clustering to trace the evolution of myths, which are broken down into common story elements. They are then categorized based on the presence/absence of these elements, and clustered based on this categorization. The results show certain myths clustering together – could this suggest a possible common origin for these myths?

Case Study 11: Speech Separation. In this study [5], the authors combined prior relevant knowledge and spectral clustering to separate two different speakers (each giving their speech and voice signal) from a one-microphone blind source. The result is an optimized segmenter for spectrograms of speech mixtures.

Case Study 12: Sensor Validation. In practice, the current status and environment factors may affect sensors' performance. If the sensors are widely distributed, it is impractical to bring a calibrating device to test each sensor individually. In this study [28], researchers use peer sensors to detect badly performing sensors.

Other Possible Domain-Specific Use Cases

- **Geography/Interdisciplinary:** detect hotspots (clusters) on maps, revealing multiple events occurring in the same location over time (e.g., earthquakes, crimes, etc.);
- **Business:** detect similar business cycles across countries, decades, etc.;
- **Tourism:** add fuzzy clustering to recommender engine algorithms to improve trip recommendations;
- **Tourism:** use fuzzy clustering to detect types of tourists, relative to a particular destination;
- **Construction/Quality Control:** use unsupervised learning to detect patterns in audio feedback indicating damage;
- etc.

1.5 Outline

We start by describing **two simple clustering algorithms** in Section 2 (k—means and its variants, and hierarchical clustering); we will then discuss the problems of **validation and model selection** in a general clustering setting in Section 3. Finally, we provide details for some **sophisticated clustering algorithms** in Section 4, which have been selected to provide analysts with a decent idea of the technicalities involved in applications.

2. Simple Clustering Algorithms

We start with two of the simplest clustering algorithms: k**—means** and **hierarchical clustering.**[5]

2.1 k-Means and Variants

One potential objective is to achieve minimal **within-cluster variation –** observations within a cluster should be very similar to one another, and the total variation over all clusters should also be small.

Assume that there are k clusters in the (scaled) dataset

$$\mathbf{X}_{n\times p}=\begin{bmatrix}\mathbf{x}_1\\ \vdots\\ \mathbf{x}_n\end{bmatrix}.$$

Let $C_1,\ldots,C_k$ denote the set of indices in each cluster, so that

$$\{1,\ldots,n\}=C_1\sqcup\cdots\sqcup C_k \quad (\textbf{hard}\text{ clustering});$$

we use the notation $\mathbf{x}_i\in C_\ell$ to indicate that observation i lies in cluster ℓ. The within cluster variation $\text{WCV}(C_\ell)$ measures the amount by which the observations in C_ℓ differ from one another.

The approach is partition-based; we look for a **partition** $\{C_\ell^*\}_{\ell=1}^k$ such that the total within cluster variation is minimized:

$$\{C_\ell^*\}=\underset{\{C_\ell\}}{\text{argmin}}\left\{\sum_{\ell=1}^k \text{WCV}(C_\ell)\right\}.$$

The first challenge is that there are numerous ways to define $\text{WCV}(C_\ell)$, and that they do not necessarily lead to the same results (as one would expect from clustering); most definitions, however, fall in line with expressions such as

$$\text{WCV}(C_\ell)=\frac{1}{(|C_\ell|-g)^\mu}\sum_{\mathbf{x}_i,\mathbf{x}_j\in C_\ell}\text{variation}(\mathbf{x}_i,\mathbf{x}_j),$$

where $\text{variation}(\mathbf{x},\mathbf{x})=0$.

Common choices for the **variation** include

$$\text{variation}(\mathbf{x}_i,\mathbf{x}_j)=\|\mathbf{x}_i-\mathbf{x}_j\|_2^2=\sum_{m=1}^p(x_{i,m}-x_{j,m})^2$$

or

$$\text{variation}(\mathbf{x}_i,\mathbf{x}_j)=\|\mathbf{x}_i-\mathbf{x}_j\|_1=\sum_{m=1}^p|x_{i,m}-x_{j,m}|;$$

these are used because of the ease of vectorizing the distance measurements, and not necessarily because they make the most sense in context. With these choices, if all observations $\mathbf{x}$ within a cluster C are near one another, we would expect $\text{WCV}(C)$ to be small. The values of the parameter μ can be adjusted to influence the cluster sizes.

[5]In this section, we borrow heavily from [25].

Traditionally, we use $\mu=0$ (or $\mu=1$) and $g=0$, and the partition problem reduces to

$$\{C_\ell^*\}=\underset{\{C_\ell\}}{\text{argmin}}\left\{\sum_{\ell=1}^k\frac{1}{|C_\ell|^\mu}\sum_{\mathbf{x}_i,\mathbf{x}_j\in C_\ell}\text{variation}(\mathbf{x}_i,\mathbf{x}_j)\right\}.$$

As an optimization problem, obtaining $\{C_\ell^*\}_{\ell=1}^k$ is NP-difficult due to the **combinatorial explosion of possible partitions** $\{C_\ell\}_{\ell=1}^k$ when n is large.[6]

Algorithm There is a way to obtain a reasonably close partition (hopefully) without having to go through all possible partitions:

1. **randomly assign** a cluster number $\{1,\ldots,k\}$ to each observation in the dataset;
2. for each C_ℓ, compute the cluster centroid;
3. assign each observation to the cluster whose centroid is **nearest** to the observation;
4. repeat steps 2-3 until the clusters are **stable**.

Three choices need to be made in order for the algorithm to run:

- the **number of clusters** k in step 1;
- the **centroid computation measure** in step 2;
- the **distance metric** used in step 3.

The most common choice of centroid measure for numerical data is to use the vector of means along each feature of the observations in each cluster (hence, k**—means**); using other centrality measures yield different methods (such as k**—medians**). For categorical data, the algorithm becomes k**—modes**.

The distance used in step 3 is usually aligned with the centroid measure of step 2 (and with the choice of a variation function in the problem statement): Euclidean for k—means, Manhattan for k—medians, Hamming for k—modes.

Variants of these approaches may use a different random initialization step: the first iteration centroids may be selected randomly from the list of observations, say.[7] Other variants indicate how to process computations in parallel (for Big Data) or for data streams (with an updating rule).

The algorithm can be shown to converge to a **stable cluster assignment**, but there is no guarantee that this assignment is the **global minimizer** of the objective function; indeed, different initial conditions can find different **local minima**, which is to say, different **clustering schemes.**

[6]Computing the number of such partitions in general cannot be done by elementary means, but it can be shown that the number is bounded above by n^k.

[7]Unfortunately, the clustering results depend very strongly on the initial randomization – a "poor" selection can yield arbitrarily "bad" (sub-optimal) results; k—means++ selects the initial centroids so as to maximize the chance that they will be well-spread in the dataset (which also speeds up the run-time).

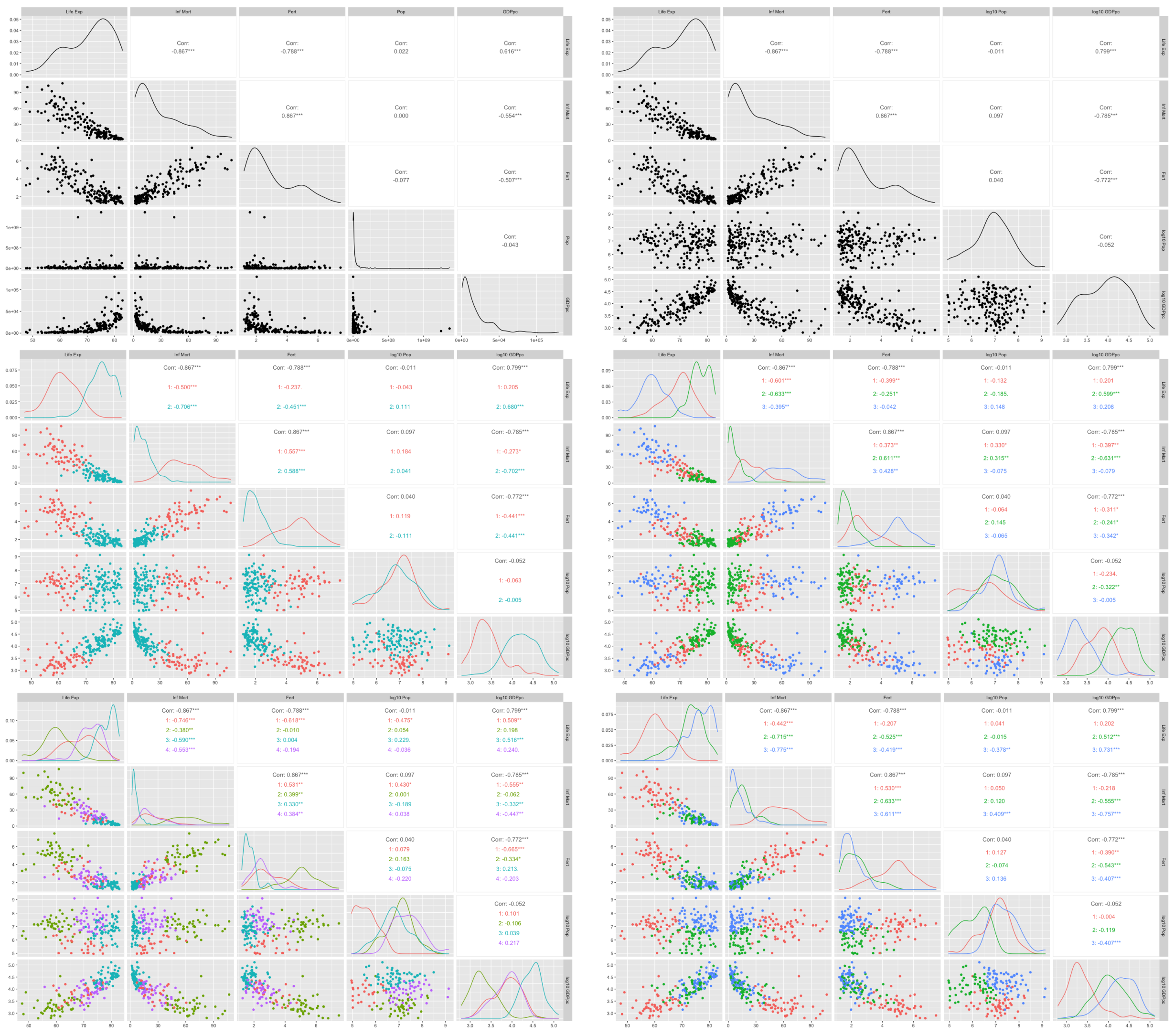


Figure 1. A scatter plot matrix of the original Gapminder 2011 data (top, left), with two outliers in the population charts (India and China); the same, but with the logarithm of the population and of the GDP per capita (top, right). Realizations of 2—means (middle, left), 3—means (middle, right; bottom, right), and 4—means (bottom, left) are also displayed.

Example We have worked with the Gapminder dataset in [9]; we will use it again to illustrate some of the notions in this module. The 2011 data contains observations on $n = 184$ countries, for the following variables:

- life expectancy (in years);
- infant mortality rate (per 1000 births);
- fertility rate (in children per woman);
- population (we use the logarithm for clustering), and
- GDP per capita (same).

A scatter plot of the original dataset is shown in Figure 1 (top row). Due to outlying observations in the population variable, we will be working instead with the logarithm of the population (and the logarithm of GDP per capita).

We run k—means for $k = 2, 3, 4$ and obtain the results shown in Figure 1:

- **2—means:** there are 64 observations in cluster 1 (red), and 120 observations in cluster 2 (blue);
- **3—means:** there are 53, 84, and 47 observations in clusters 1, 2, and 3, respectively;
- **4—means:** there are 26, 47, 61, and 50 observations in clusters 1, 2, 3, and 4, respectively.

The colours (cluster labels) are not used by the clustering algorithm – they are its **outputs** (the value is irrelevant).

The last chart shows the result of a different initialization for $k = 3$, leading to a different cluster assignment.

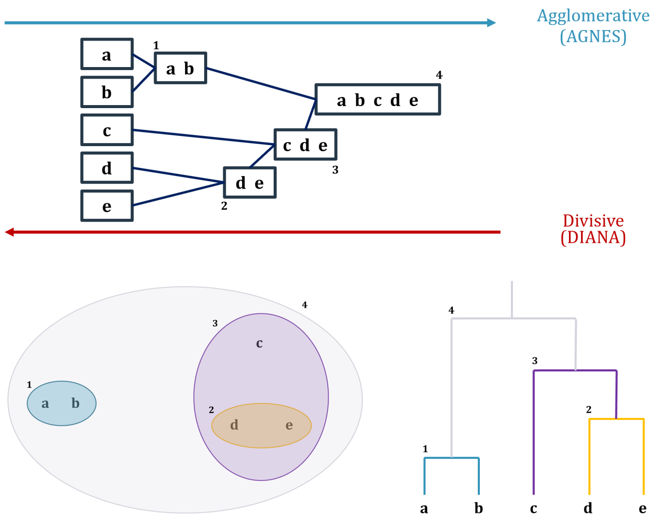


Figure 2. Illustration of AGNES and DIANA (top); corresponding hierarchical structure and dendrogram (bottom).

2.2 Hierarchical Clustering

One of the issues surrounding the use of k—means (and its variants) is that nothing in the result of a single run of the algorithm indicates if the choice of k was a good one.[8]

This can only be achieved by repeatedly running the algorithm over a range of "reasonable" values of k (to account for initialization variability), and by comparing the outputs using some of the methods discussed in Section 3. This process can be memory-extensive.

Hierarchical clustering (HC) can sidestep this difficulty altogether by building a deterministic (for a choice of parameters) **global clustering structure** from which we can select a specific number of clusters after the algorithm has converged; the advantage of this approach is that if we want to use a different number of clusters, we do not need to re-run the clustering algorithm – we simply read off the new clusters from the global clustering structure.

There are two main conceptual approaches:

- **bottom-up/agglomerative** (AGNES) – initially, each observation sits in its own separate cluster, which are then merged (in pairs) as the hierarchy is climbed, leading (after the last merge) to a large cluster containing all observations;
- **top-down/divisive** (DIANA) – initially, all observations lie in the same cluster, which is split (and re-split) in pairs as the hierarchy is traversed downward, leading (after the last split) to each observation siting in its own separate cluster.

They are illustrated in Figure 2.

[8]The results might look good on a 2-dimensional representation of the data, but could it look better?

In theory, the two approaches are equivalent (they produce the same hierarchical structure given a similarity metric and a **linkage strategy** (more on this later); in practice, we use AGNES over DIANA for anything but small datasets as the former approach runs in polynomial time (with respect to the number of observations), whereas the latter runs in exponential time.

With AGNES, the **clustering dendrogram** is built starting from the leaves, and combining clusters by pairs, up to the root, as in Figure 3.

Algorithm We build the global structure as follows:

1. each observation is assigned to **its own cluster** (there are n clusters, initially);
2. the two clusters that are the least dissimilar are merged into a **supra-cluster**;
3. repeat step 2 until **all of the observations** belong to a **single** large merged clusters (with n observations).

Three parameters need to be made in the order for the algorithm to run:

- the choice of a **linkage strategy** in steps 2 and 3;
- the **dissimilarity measure** used in step 2;
- the **dissimilarity threshold** required to "cut" the dendrogram into clusters.

If Figure 3, the dataset is split into $n = 50$ clusters; observations 13 and 34 are then found to be most similar, and merged into a single cluster, and the 50 observations are grouped into 49 clusters. The next two observations which are most similar are 14 and 37, which are themselves merged, so that there are 48 clusters at that level.

The process is continued until all observations are merged into a single cluster, leading to the global clustering structure (**clustering dendrogram**) for the dataset.

In order to obtain actual clusters (as opposed to the global structure), we cut the dendrogram at the selected dissimilarity level, with the resulting groups of observations yielding the dataset clusters (5, in the example of Figure 3). Increasing the dissimilarity threshold decreases the number of clusters, and *vice-versa*.

Linkage Strategy In the first AGNES stage, we compare all pairs of observations to determine which two are least dissimilar, and how these are merged into a cluster.[9]

In the second stage, we must also compare each of the non-merged observations with a **cluster** of two observations to determine their dissimilarity (the other dissimilarities have been computed in the first stage and do not need to be computed anew).

In subsequent steps, we might also need to compare two clusters with any number of observations for overall similarity. How can this be achieved?

[9]With n observations, there are $1 + \cdots + (n-1) = \frac{(n-1)n}{2}$ such pairs.

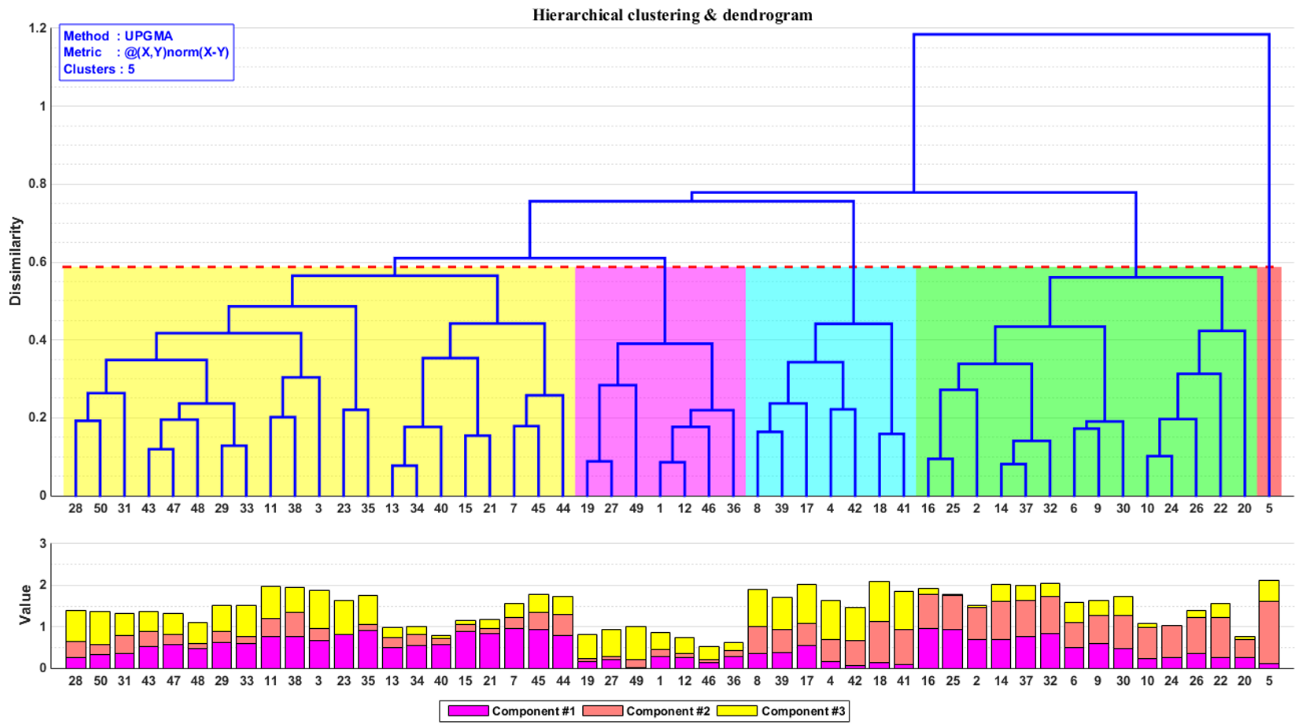


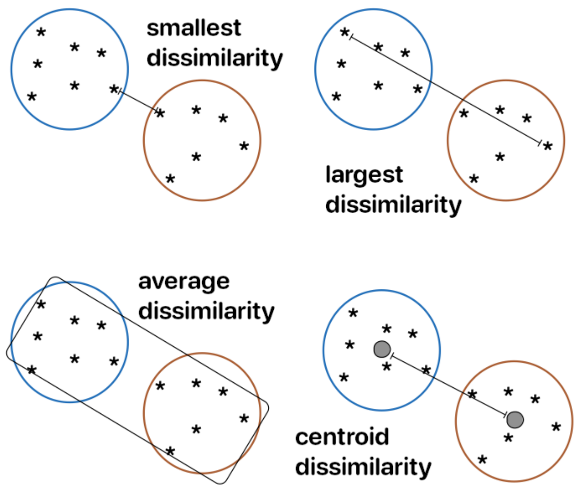
Figure 3. Cluster dendrogram for the hiearchical cluster structure of a dataset with 50 observations and 3 variables, with average linkage (UPGMA) and using the Euclidean distance as the dissimilarity measure [author unknown]. The dendrogram is cut at a dissimilarity level ≈ 0.6 so that 5 clusters emerge (ordered and coloured in red, magenta, blue, green, and red); the observations profiles are shown in the stacked bar chart and provide potential descriptions of the clusters (magenta = small total height, with mostly dominant 3rd components, say). Based on the profiles, we might also have elected to cut a slightly lower dissimilarity level (≈ 0.55), so that the yellow and green clusters would have been further split into two clusters apiece (between observations 35 and 13, and 30 and 10, perhaps?).

Let A and B be clusters in the data, with $|A| = n_A$, $|B| = n_B$. The **dissimilarity** between A and B can be computed in multiple ways:

- **complete linkage** – the **largest** dissimilarity among all pairwise dissimilarities between the observations in A and those in B ($n_A n_B$ computations);
- **single linkage** – the **smallest** dissimilarity among all pairwise dissimilarities as in complete linkage;
- **average linkage** – the **average** dissimilarity among all pairwise dissimilarities as in complete (or single) linkage;
- **centroid linkage** – the dissimilarity between the **centroids** of A and B (found using whatever method is appropriate for the context);
- **Ward's method** (and its variants) uses any reasonable **objective function** which reflects domain knowledge [4, 54];
- etc.

The choice of a **linkage strategy** (and of a dissimilarity measure) affects not only how clusters are compared and merged, but also the **topology** (shape) of the resulting dendrogram (are the clusters tight, loose, blob-like, etc.).

The various linkage strategies are illustrated below, for Euclidean dissimilarity.



Example We show the results of hierarchical clustering on the Gapminder 2011 data, using complete linkage and Eulidean dissimilarity (see Figure 4) and Ward D linkage and Manhattan dissimilarity (see Figure 5). In each case, we consider $k = 2, 3, 4$ clusters.

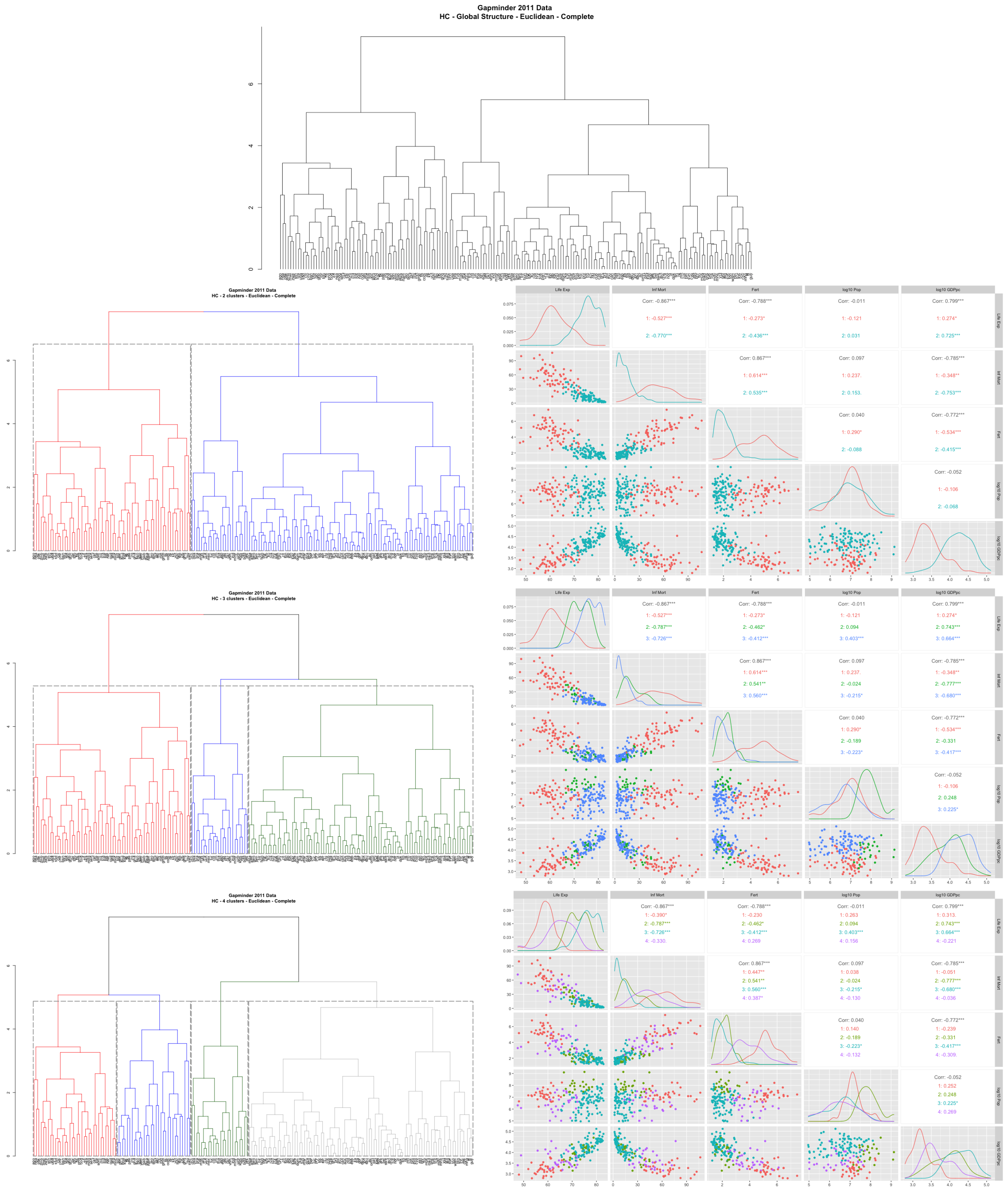


Figure 4. Gapminder 2011 data; hierarchical clustering with complete linkage and Euclidean dissimilarity. Global clustering structure (top); 2 clusters with 66 and 118 observations apiece (2nd row); 3 clusters with 66, 24, and 94 observations apiece (3rd row); 4 clusters with 35, 31, 24, and 94 observations apiece (4th row).

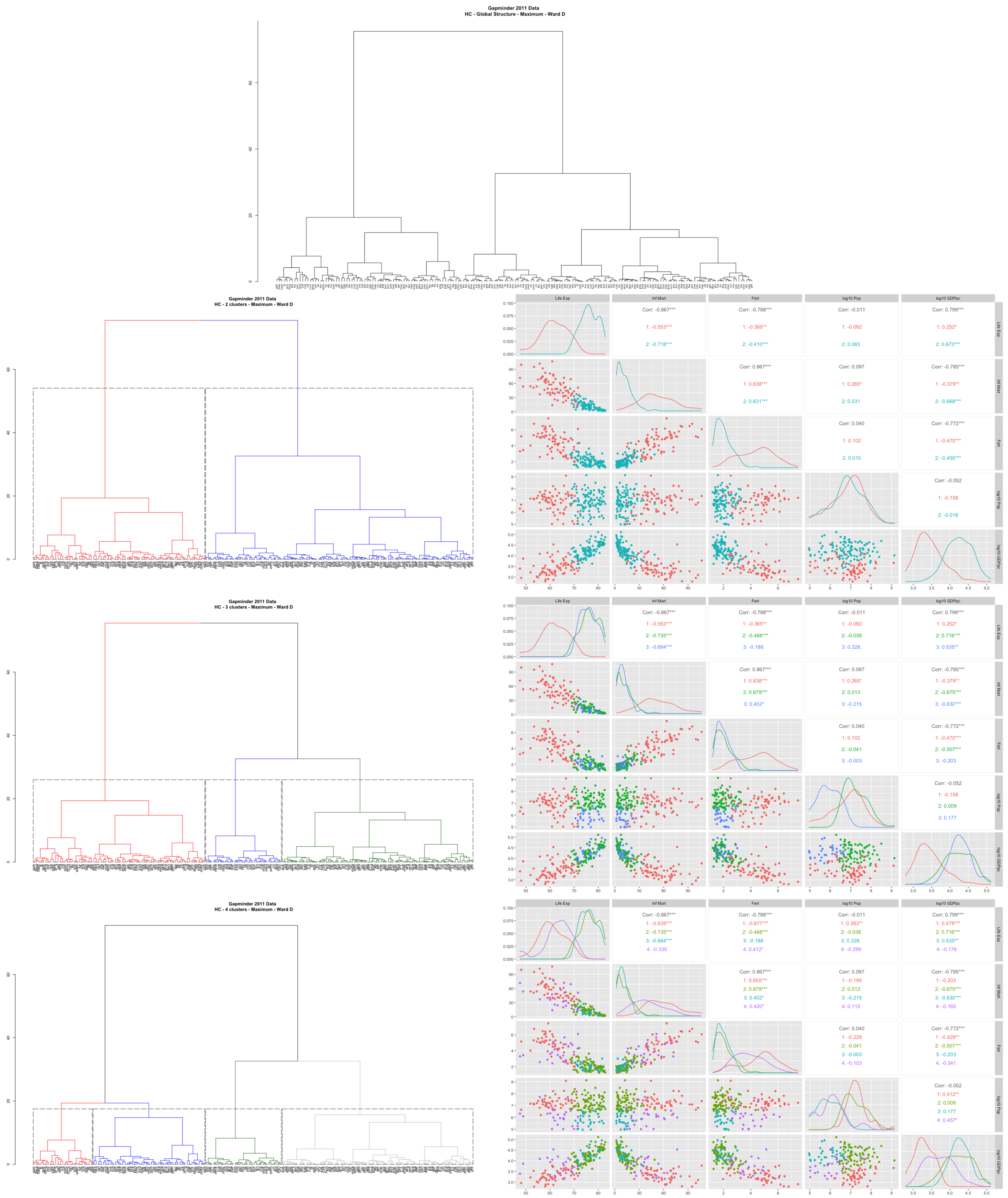


Figure 5. Gapminder 2011 data; hierarchical clustering with Ward *D* linkage and Manhattan dissimilarity. Global clustering structure (top); 2 clusters with 72 and 112 observations apiece (2nd row); 3 clusters with 72, 80, and 32 observations apiece (3rd row); 4 clusters with 47, 25, 80, and 32 observations apiece (4th row).

3. Evaluation

Hierarchical clustering (HC) and k—means (and its variants) both attempt to separate the data into **natural groups**, using different conceptual approaches; k—means tries to minimize within-cluster variation while HC builds a global clustering structure.

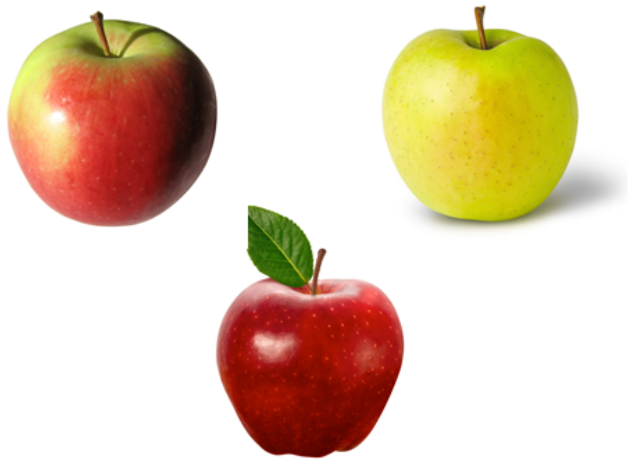
We have discussed some of their shortcomings in Section 2; the fact that they may yield different clustering outcomes depending on the choices made along the way (initialization, similarity/dissimilarity measures, linkage strategy, number of clusters, etc.) reinforces the notion that unsupervised learning is **difficult**.

We will discuss advanced algorithms that sidestep some of these issues in Section 4, but we make an important observation in the meantime: a hallmark of clustering is that whenever a new approach manages to overcome a previously-identified difficulty, it usually does so at the cost of introducing holes in hitherto sound aspects.

We may thus not be able to ever find a "best" clustering approach/outcome (an updated take on the No Free Lunch theorem, perhaps? [55]), but is it possible to identify which of several clustering scheme is "optimal" (or best-suited for an eventual task)?

3.1 Clustering Assessment

In machine learning, clustering is defined as grouping objects based on their over-all similarity (or dissimilarity) to each other.[10] It can be tempting to focus on just one or two attributes (especially for visual or "eyeball" clustering), but keep in mind that even if we were to focus on one or two particular attribute, all of the other attributes must still come along for the ride. For instance, consider the objects shown below.



What is the same about these objects? What is different? Do they belong in the same group? If so, how many?

[10]Note that each object has multiple dimensions, or attributes available for comparison.

As a start, they are all pictorial representations of apples;[11] they all possess stems, and appear to be of similar size.

On the other hand, two of them are (mostly) red while the other is green(ish); one of the stem has a leaf while the other two do not, and one of them is spherical, while the other two are "tapered" at the bottom.

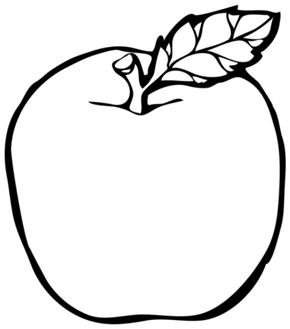
We do recognize them all as apples, but we could make an argument that each of them is unlike the other two (and thus also that each of them is similar to exactly one of the other two).

Fruit Image Dataset In order to appreciate the challenges presented by clustering validation, it will be helpful to relate the concepts to something tangible. We will explore some of these notions through an artificial dataset of 20 fruit images (see Figure 6):

- are there right or wrong groupings of this dataset?
- are there multiple possible 'natural' clusterings?
- could different clusterings be used differently?
- will some clusterings be of (objectively) higher quality than others?

Key Notions At a fundamental level, clustering relies on the notion of **representativeness**; ideally, the essence of **instances** (observations) in a cluster would be faithfully captured by the cluster **concept** (examplar, representative), and differentiated from those of other clusters by the same token.

For instance, the image below is a concept for "apples":



as is the Mirriam-Webster definition:

> "The fleshy, usually rounded red, yellow, or green edible pome fruit of a usually cultivated tree (genus Malus) of the rose family."

Of course, this is not *all* that an apple is, but most people who have seen or eaten an apple at some point in the past will have no trouble recognizing what is being alluded to by the concept, even if the corresponding mental image differs from one person to the next.

[11]While we cannot forget that they are not actual apples, we will assume that this is understood and simply refer to the objects as fruit, or apples.

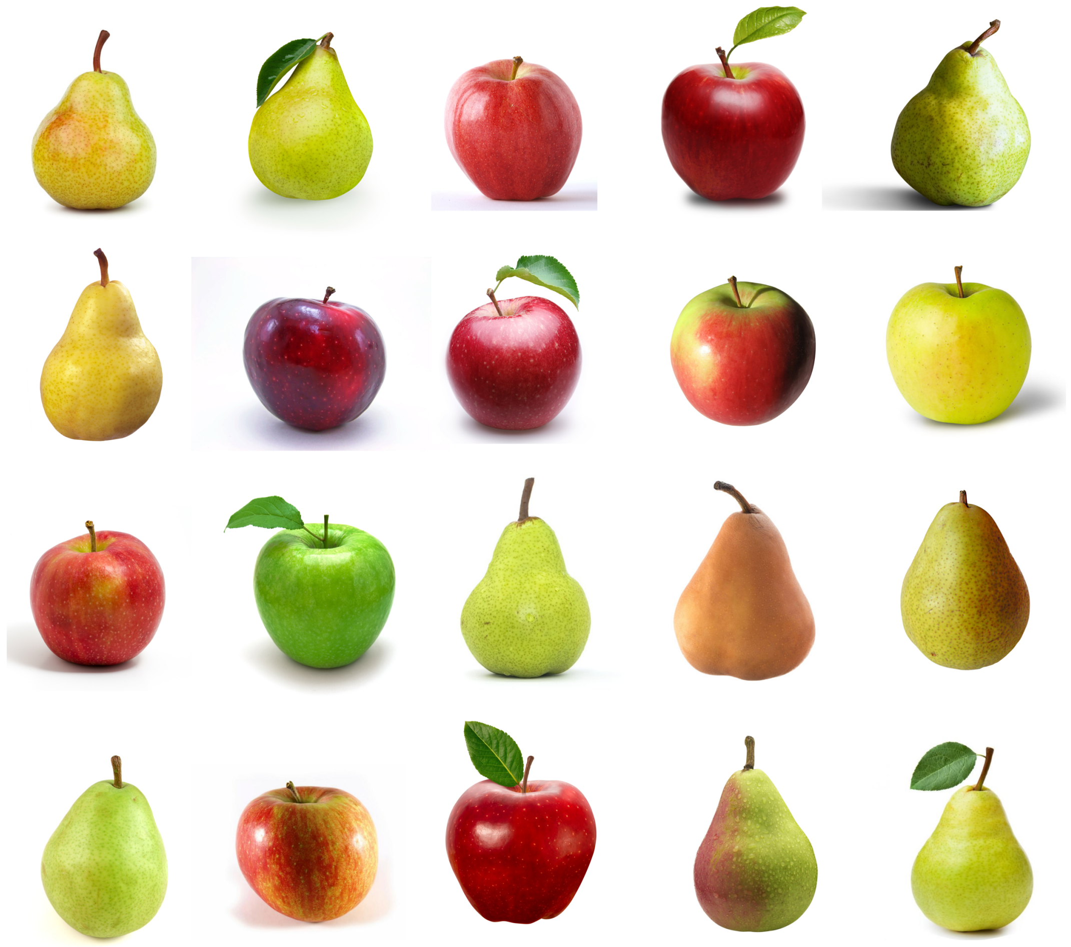


Figure 6. Toy dataset with which the key concepts of clustering validation are illustrated.

The cluster concepts offer a **generalized representation** – they capture "something" of their corresponding cluster's instances. For a given cluster, then, can we clearly identify a concept capturing its (and solely its) essence? If so, does that make the entire clustering scheme a good one?

For machine learning purposes, we use **signature vectors** to represent **instance properties**. Each vector element represents an instance **attribute**; the element's value is the **measured value** of the corresponding object's property (for instance, the colour of the apple).

The apple below, as an example,

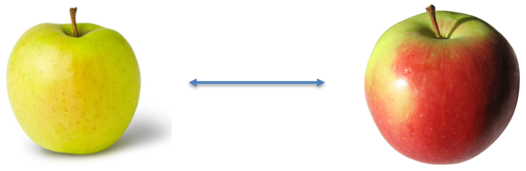
could perhaps be described by the signature vector

$$(12, 9.12, \text{tapered}, \text{golden delicious}),$$

where the components are the instance's colour (ordinal), height (continuous), shape and variety (both categorical).[12]

Signature vectors are used to compare objects (**instance-to-instance relationships**); such comparisons could yield, among others, a measure of **similarity** between instances.

While similarity can be measured against a single **dimension** (component), the comparisons of interest for clustering task require an overall similarity measure, **across all dimensions**. We would compare the two apples below, say,



[12]An important consideration, from a general data science perspective, is whether the signature vector provides a **sufficient description** of the associated object or whether it is too crude to be of use. This is usually difficult to ascertain prior to obtaining analysis results, and comparing them to the "reality" of the underling system (see [8, 11] for details).

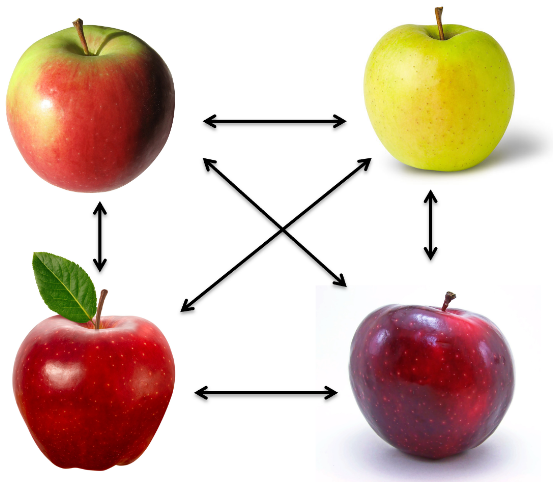
by comparing their signature vectors

$$\mathbf{v}_1 = (12, 9.12, \text{tapered}, \text{golden delicious})$$
$$\mathbf{v}_2 = (2, 10.43, \text{spherical}, \text{macintosh})$$

with the help of some similarity measure $w(\mathbf{v}_1, \mathbf{v}_2)$.[13] As we have discussed in Section 1, the use of a **distance measure** (or metric) is a popular strategy for defining how similar (or dissimilar) two objects are to each other.[14]

In the **clustering framework**, we are often interested in all pairwise similarities between objects, not just in the similarity between two objects, which is to say that pairs of objects are not solely interesting in and of themselves, but also in relation to other pairs of objects.

In a dataset with 4 objects, for instance, we might require the computation of (up to) 6 pairwise similarities (as shown below).



As is the case with objects, **clusters** have **properties**. These could include:

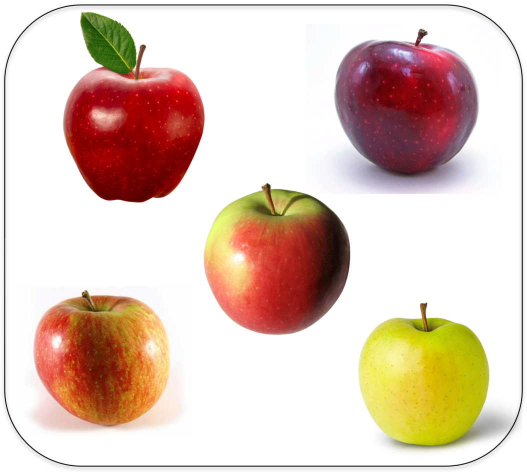
- the number of instances in a cluster;
- the similarity measures across instances within a cluster (minimum, maximum, average, median, standard deviation, mean absolute deviation, variance, etc.);
- the cluster representative, which may be an actual instance, or an amalgamation of multiple instances (**exemplar**).

[13]Various similarity measures may yield various results, in some cases showing the two apples to be similar, in others to be dissimilar.

[14]Importantly, a distance takes into account all of the properties of the objects in question, not just a few of them. Traditionally, only numeric attributes are allowed as input (see [12] for an in-depth discussion of distance metrics), but it is technically possible to convert categorical attributes to numeric ones, or to define **mixed distances**.

While the moniker "distance" harkens back to the notion of Eulidean (physical) distance between points in space, it is important to remember that the measurements refer to the distance between the associated signature vectors, which do not necessarily correspond to their respective physical locations.

How many instances are there in the cluster below, for instance? What pair of observations is most similar? The least similar? What are the similarity values? Which instance is most representative?



We can also define **cluster-to-instance** relationships. A specific instance can be:

- compared to a cluster representative;
- compared to specific instances in a cluster (as in instance-to-instance relationships), such as the most similar instance or the most distant instance.

This allows for **membership** questions: is the green apple similar to the cluster below? Does it belong in the cluster, or is it most likely to belong to another cluster? Or perhaps to no cluster in particular?

Finally, we might be interested in **cluster-to-cluster** relationships, where we compare cluster-level properties, such as:

- number of instances;
- **within-cluster** similarities;
- cluster representatives.

To this, we can also add **between-cluster** (or across-cluster) similarities, as a way to determine if the instances are notably different from one cluster to the next.

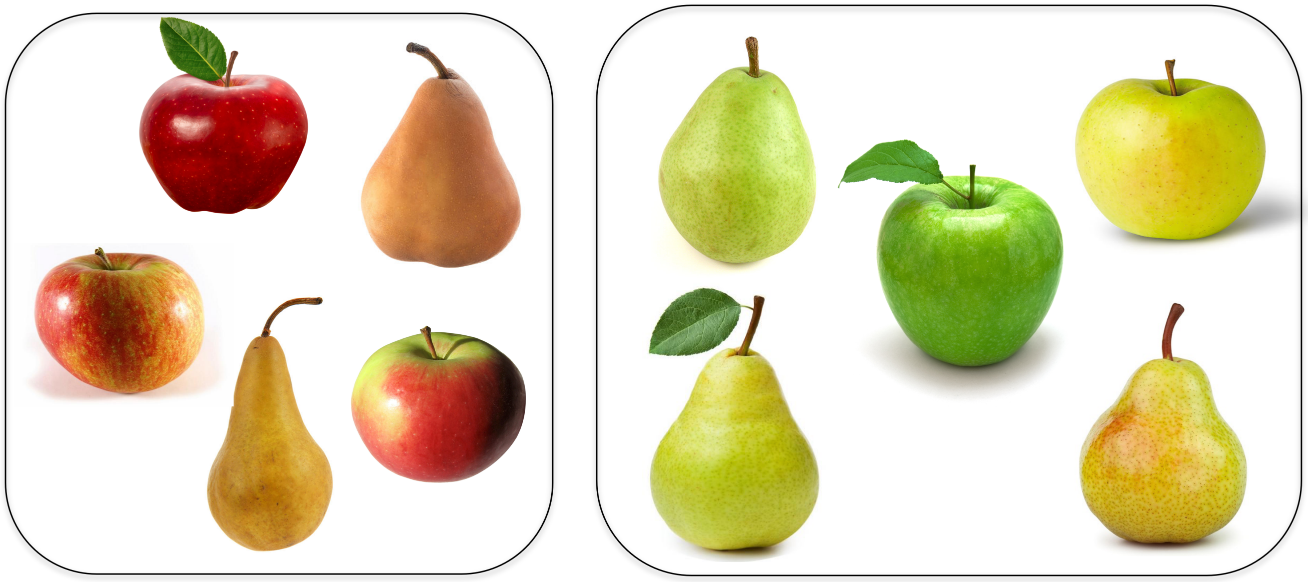
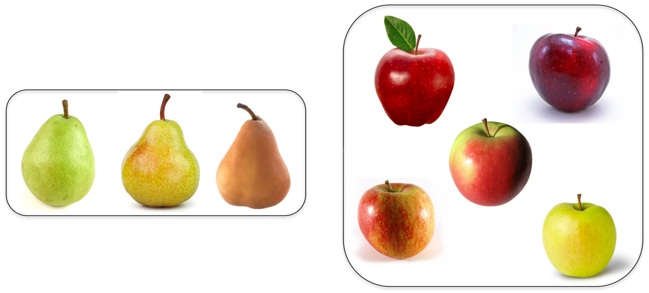


Figure 7. Two clusters in a subset of the toy dataset.

This allows for **validity** questions: are the two clusters below significantly different? Should they be joined into a mega-cluster? Does it make sense to have them as separate clusters in the dataset?



How would we qualify the clustering outcome of Figure 7, for instance, as it relates to colour, height, and shape? Could there be clusterings of higher quality? Of lower quality? How could this be quantified?

Cluster and instance comparisons can be combined in many different ways, which can then be used to generate a vast number of **clustering validation functions**.

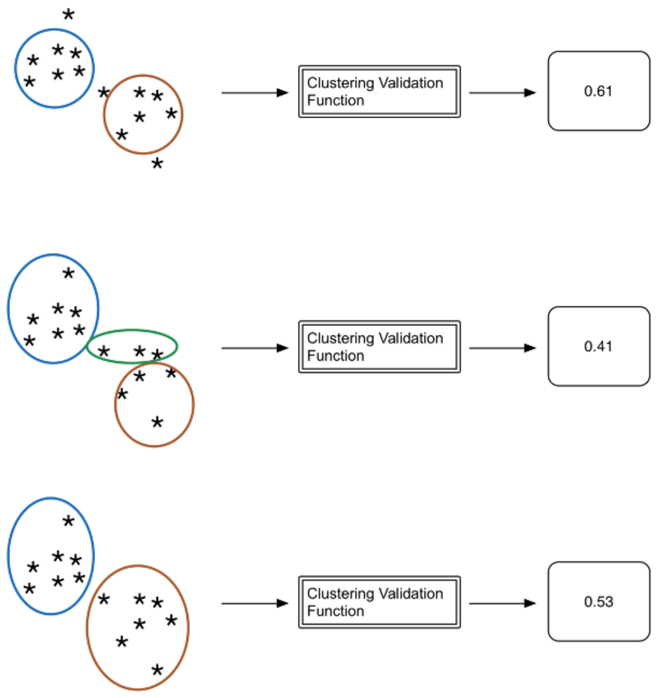
The central cluster validation question becomes: what can these tell us about the quality of a particular clustering outcome relative to some objective criteria about "good" clustering schemes (**internal validation**), to another clustering option (**relative validation**), or to external information (**external validation**)?

Clustering Quality Measures In general, clustering involves two main activities:

- **creating/building** the clusters, and
- **assessing their quality**, individually and as a whole.

From a practical perspective, clustering requires two functions: one which assigns each instance to a cluster,[15], and one which assigns each clustering scheme to a **cluster quality measurement**.[16]

An illustration is provided below, on an artificial dataset containing two variables, with dissimilarity between instances given by the corresponding Euclidean distance:


Clustering Validation Function
0.61
Clustering Validation Function
0.41
Clustering Validation Function
0.53

[15]Or in the case of soft clustering, assign each instance a "probability" of belonging to each cluster.
[16]The similarity matrix is typically required at both stages.

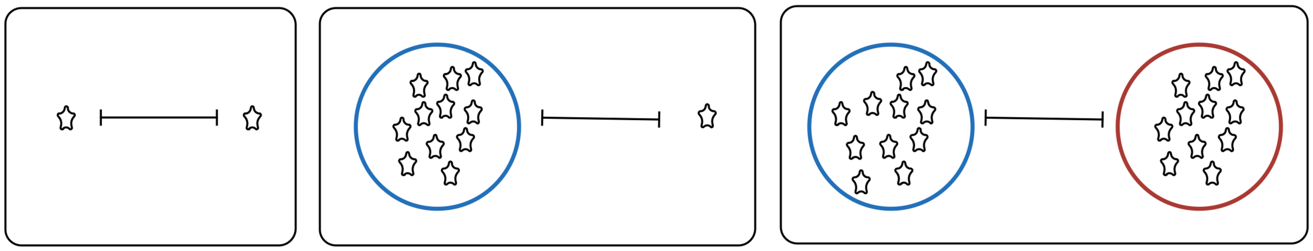


Figure 8. Schematic illustrations of various instance/cluster properties and relationships.

Three different clustering schemes are obtained, and their quality is assessed with the help of some clustering validation function:[17]

- **top** – two clusters are found in the data (with outliers), and the quality of the clustering is assessed as 0.61;
- **middle** – three clusters are found (no outliers), with quality assessment at 0.41;
- **bottom** – two clusters are found (no outliers), with quality assessment at 0.53.

With this choice of clustering validation function, the top scheme would be prefered, followed by the bottom scheme; the middle one brings up the rear.

We have already mentioned the abundance of clustering algorithms [53]; it will come as no surprise that a tremendous number of clustering validation function in practice as well (for much the same reasons as those discussed in Section 1.2).

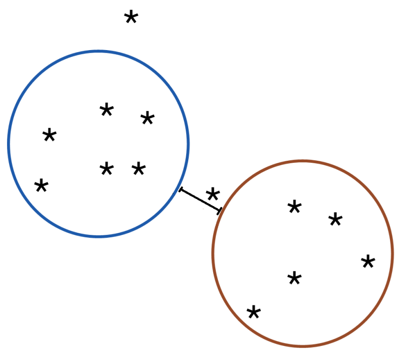
They are, however, all built out of the basic measures relating to instance or cluster properties we have already reviewed, as illustrated schematically in Figure 8:

- instance properties;
- cluster properties;
- instance-to-instance relationship properties;
- cluster-to-instance relationship properties, and
- cluster-to-cluster relationship properties,

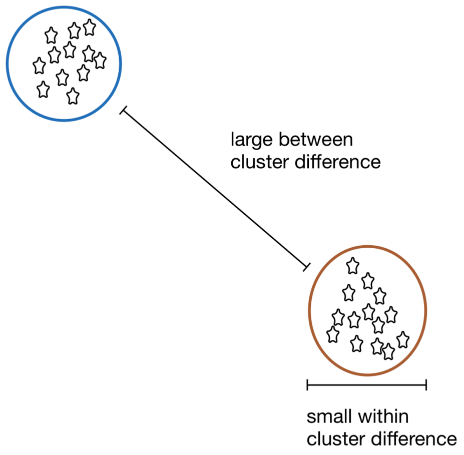
Internal Validation Context is quite relevant to the **quality** of a given clustering result. But what if no context is readily available? **Internal validation** methods attempt to measure cluster quality objectively, without context.[18]

We could elect to validate the clustering outcome using only the properties of the obtained clusters, such as, say, the distance between all clusters: if the average between-cluster distance is large, we might feel that there is some evidence to suggest that the resulting clusters provide a good segmentation of the data into natural groups (as at the top of the column on the right).

[17]The specifics of that function are not germane to the discussion.
[18]"Clustering validation" suggests that there is an ideal clustering result against which to compare the various algorithmic outcomes, and all that is needed is for analysts to determine how much the outcomes depart from the ideal result. "Cluster quality" is a better way to refer to the process.



Alternatively, we might validate cluster quality by tempering the average between-cluster distance against the average within-cluster distance between the instances, which would reward "tight" and "isolated clusters", as opposed to simply "isolated" cones.



There are multiple ways of including both between-cluster and within-cluster similarities in a **cluster quality metric** (CQM): both the **Davies-Bouldin index** and **Dunn's index** do so, to name but two examples.

The broad objectives of clustering remain universal: instances within a cluster should be similar; instances in different clusters should be dissimilar. The problem is that there are many ways for clusters to deviate from this ideal: in specific clustering cases, how do we weigh the "good" aspects (such as high within-cluster similarity) relative to the "bad" ones (such as low between-cluster separation)?

Other internal properties and relationships can also be used and combined in various ways, leading to an embarrassment of riches when it comes to CQMs. The following indices are all available in the R package clusterCrit, for instance [14]:

- Ball-Hall
- Banfeld-Raftery
- C
- Calinski-Harabasz
- **Davies-Bouldin**
- Det Ratio and LogDetRatio
- **Dunn**
- Baker-Hubert Gamma
- GDI
- Gplus
- KsqDetW
- McClain-Rao
- PBM
- Point-Biserial
- Ratkowsky-Lance
- Ray-Turi
- Scott-Symons
- SD
- SDbw
- **Silhouette**
- Tau
- TraceW and TraceWiB
- Wemmert-Gancarski
- **WSS** and LogSSRatio
- Xie-Beni

While it is useful to have access to so many CQMs, we should remain aware that the No-Free Lunch theorem is still in play: none of them is universally superior to any of the others.[19]

In practice, certain approaches (**weightings**) may prove more relevant given the eventual specific analysis goals, but what these could prove to be is not always evident from the context; consequently, we recommend assessing the quality of the clustering outcomes using a variety of CQMs.

Studies have been performed to compare a large number of internal validation measures, relative to one another and against human evaluation; generally speaking, variants of the **silhouette index** performed reasonably well (but see previous NFL comment) [32, 48].

One possible interpretation of these results is that internal validation *via* CQMs may be providing information about clustering across all contexts, and that it is usually easier to identify clustering outcomes that clearly miss the mark than it is to objectively differentiate amongst "good" outcomes, for reasons that are not fully understood yet.

[19]Given that all of them are supposedly provide context-free assessments of clustering quality, that is problematic (although emblematic of unsupervised endeavours).

Details on the CQMs are readily available from a multitude of sources (see [2, 32, 48], among others); we provide more information for 4 commonly-used CQMs:

- (within) sum of squared error;
- Davies-Bouldin;
- Dunn, and
- silhouette;

Let $\mathscr{C} = \{C_1, \ldots, C_K\}$ be the K clusters obtained from a dataset $\mathbf{X}$ *via* some algorithm $\mathscr{A}$. Denote the **centroid** (or some other central representative) of cluster C_k by $\mathbf{c}_k$. The **(within) sum of error** for $\mathscr{C}$ is

$$\text{WSE} = \sum_{k=1}^{K} \sum_{\mathbf{x} \in C_k} \text{dissimilarity}(\mathbf{x}, \mathbf{c}_k).$$

The dissimilarity is often selected to be the **square of the Euclidean distance** (thence "sum of squared error"), but the choice of the Euclidean distance (and of the square exponent) is arbitrary – it would make more sense, in practice, to use a distance metric related to the similarity measure used by $\mathscr{A}$.

Note that WSS decreases with the number of clusters K, and optimality is obtained at **points of diminishing returns** (see Section 3.2 for details); from a validation perspective, it might be easier to interpret the **(within) average error**:

$$\text{WAE} = \operatorname*{avg}_{k=1}^{K} \left\{ \operatorname*{avg}_{\mathbf{x} \in C_k} \{\text{dissimilarity}(\mathbf{x}, \mathbf{c}_k)\} \right\} := \operatorname*{avg}_{k=1}^{K} \{s_k\}.$$

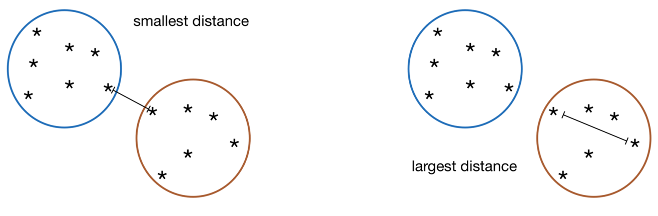
With s_k, $k = 1, \ldots, K$ as above, the **Davies-Bouldin index** (DBI) is defined as

$$\text{DBI} = \frac{1}{K} \sum_{k=1}^{K} \max_{\ell \neq k} \left\{ \frac{s_k + s_\ell}{\text{dissimilarity}(\mathbf{c}_k, \mathbf{c}_\ell)} \right\}.$$

If the clusters $\{C_k\}$ are **tight** and **dissimilar to one another**, we expect the numerators $s_k + s_\ell$ to be small and the denominators dissimilarity$(\mathbf{c}_k, \mathbf{c}_\ell)$ to be large, so that the DBI would be **small**.

Conversely, with clusters that are **loose** or **somewhat similar to one another**, we expect the DBI to be **large**.

There are other ways to assess separation and tightness; **Dunn's index** is the ratio of the **smallest between-cluster dissimilarity** (for pairs of observations not in the same cluster) to the **largest cluster diameter** (within-cluster dissimilarity).



If the clusters $\{C_k\}$ are **tight** and **dissimilar to one another**, we expect the smallest between-cluster dissimilarity to be large and the largest cluster diameter to be small, leasing to a **large** Dunn ratio.

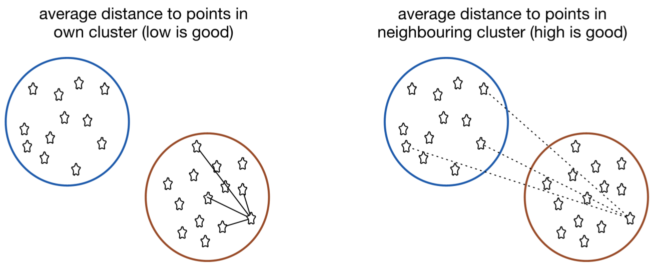
Conversely, with clusters that are **loose** or **somewhat similar to one another**, the Dunn ratio will be **small**.

As is the case with the sum of errors and the DBI, the choice of the dissimilarity (or distance) measure leads to different variants of the Dunn index, but all of them take non-negative values.

The three previous CQMs provide examples of validation metrics that are computed for the entire clustering outcome, the **silhouette metric** instead assigns a score to each observation, and aggregates the scores at a cluster level and at the dataset level: for a dissimilarity d and a point $\mathbf{x}$ in a cluster C, set

$$b(\mathbf{x}) = \min_{C' \neq C}\left\{\operatorname*{avg}_{\mathbf{y}\in C'}\{d(\mathbf{x},\mathbf{y})\}\right\}, \quad a(\mathbf{x}) = \operatorname*{avg}_{\substack{\mathbf{w}\in C\\ \mathbf{w}\neq\mathbf{x}}}\{d(\mathbf{x},\mathbf{w})\}.$$

Small values of $a(\mathbf{x})$ imply that $\mathbf{x}$ is similar to the instances in its cluster, large values of $b(\mathbf{x})$ imply that it is dissimilar to instances in other clusters.


average distance to points in own cluster (low is good)
average distance to points in neighbouring cluster (high is good)

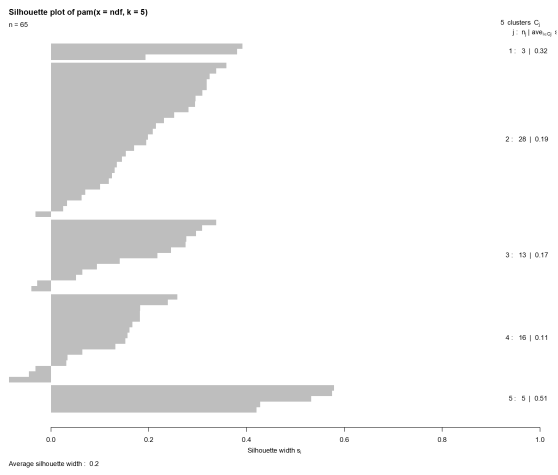
The silhouette metric at $\mathbf{x}$ is

$$\begin{aligned}\text{silhouette}(\mathbf{x}) &= \frac{b(\mathbf{x})-a(\mathbf{x})}{\max\{a(\mathbf{x}),b(\mathbf{x})\}}\\ &= \begin{cases}1-a(\mathbf{x})/b(\mathbf{x}) & \text{if } a(\mathbf{x})<b(\mathbf{x})\\ 0 & \text{if } a(\mathbf{x})=b(\mathbf{x})\\ b(\mathbf{x})/a(\mathbf{x})-1 & \text{if } a(\mathbf{x})>b(\mathbf{x})\end{cases}\end{aligned}$$

Consequently, $-1 \leq \text{silhouette}(\mathbf{x}) \leq 1$ for all $\mathbf{x}$. Thus, when silhouette($\mathbf{x}$) is large (≈ 1), the ratio $a(\mathbf{x})/b(\mathbf{x})$ is small and we interpret $\mathbf{x}$ to be correctly assigned to cluster C (and conversely, when silhouette($\mathbf{x}$) is small (≈ -1), we interpret $\mathbf{x}$'s assignment to C to be "incorrect").

The **silhouette diagram** corresponding to the clustering results displays the silhouette scores for each observations, and averages out the scores in each cluster. The mean over all observations (and the number of instances that have been poorly assigned to a cluster) gives an indication of the appropriateness of the clustering outcome.

In the example below, 65 observations are separated into 5 clusters: 6 observations are **poorly assigned** (those with negative silhouette scores) and the **average silhouette score** over the entire dataset is 0.2, which suggests that the clustering is more successful than unsuccessful, overall.

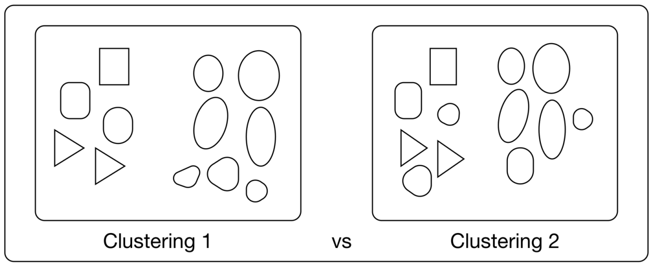


Note, however that it could prove difficult to associate intrinsic meaning to a **lone numerical score**, in general – there could be clustering contexts where 0.2 is considered to be a fantastic silhouette score, and others where it is viewed as an abject failure.

It is in comparison to the scores obtained by different clustering schemes that this score (and those of the other CQMs) can best serve as an **indicator** of a strong (or a poor) clustering outcome.

Relative Validation Obtaining a **single validation measure** for a **single clustering outcome** is not always that useful – how would we really characterize the silhouette score of 0.2 in the previous example? Could the results be better? Is this the best that can be achieved?

One approach to this problem is to compare clustering outcomes across multiple runs to take advantage of non-deterministic algorithms or various parameters' values (see image below, and schematics in Figure 9): if the outcome of different clustering algorithms on the same dataset are "similar", this provides evidence in favour of the resulting clusters being efficient, natural, or valid, in some sense.



Consider, for instance, a dataset with 6 instances, which is clustered in two different manners ($\mathscr{A}$ and $\mathscr{B}$, with $|\mathscr{A}| = 3$ and $|\mathscr{B}| = 2$), as shown below. The clusterings are clearly not identical; how similar are they?

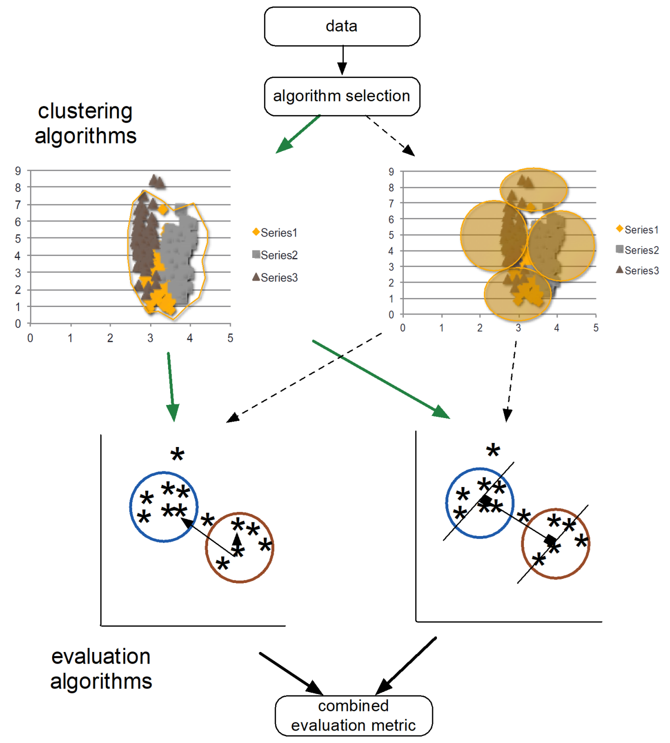
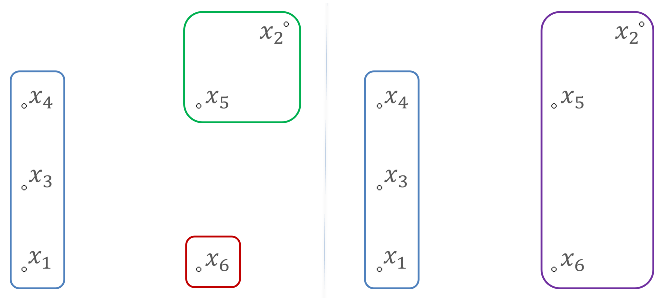


Figure 9. Schematics of relative clustering validation.



One way to approach relative validation for two outcomes is by computing "**correlations**" between cluster outcomes. Intuitively, we would expect the similarity between both clustering schemes to be high, seeing as the two outcomes are not that different from one another.[20]

	P1	P2	P3	P4	P5	P6
P1	1					
P2	0	1				
P3	1	0	1			
P4	1	0	1	1		
P5	0	1	0	0	1	
P6	0	0	0	0	0	1

	P1	P2	P3	P4	P5	P6
P1	1					
P2	0	1				
P3	1	0	1			
P4	1	0	1	1		
P5	0	1	0	0	1	
P6	0	1	0	0	1	1

How can this be quantified? **Correlation measures** include the Rand, Jaccard (see [12]), and Gamma (see [57]) indices.

[20]In $\mathscr{B}$, the two smallest clusters have been merged into a larger cluster.

Let $\mathscr{A} = \{A_1,\ldots,A_k\}$ and $\mathscr{B} = \{B_1,\ldots,B_\ell\}$ be two clusterings of a dataset $\mathbf{X}$. If $\mathbf{X}$ consists of n instances, there are thus

$$\binom{n}{2} = \frac{n(n-1)}{2}$$

pairs of observations in the data.

Denote the number of pairs of observations that are in:

- the same cluster in $\mathscr{A}$ **and** in the same cluster in $\mathscr{B}$ by ss,
- different clusters in $\mathscr{A}$ **and** different clusters in $\mathscr{B}$ by dd;
- the same cluster in $\mathscr{A}$ **but** in different clusters in $\mathscr{B}$ by sd, and
- different clusters in $\mathscr{A}$ **but** in the same cluster in $\mathscr{B}$ by ds.

Thus,

$$\binom{n}{2} = \text{ss} + \text{dd} + \text{sd} + \text{ds},$$

and we define the **Rand index** of $\mathscr{A}$ and $\mathscr{B}$ as

$$\text{RI}(\mathscr{A},\mathscr{B}) = \frac{\text{ss}+\text{dd}}{\text{ss}+\text{dd}+\text{sd}+\text{ds}} = \frac{\text{ss}+\text{dd}}{\binom{n}{2}}.$$

Large values of the index are indicative of similarty of clustering outcomes.[21]

Unfortunately, the Rand index does not satisfy the **constant baseline property**.[22] The **adjusted Rand index** (as well as several other pair-counting, set-matching, and information theoretic measures) relies on the **contingency table** between $\mathscr{A}$ and $\mathscr{B}$, a method to summarize the outcomes of two clustering results on the same dataset:

	B_1	$\cdots$	B_ℓ	Total
A_1	$n_{1,1}$	$\cdots$	$n_{1,\ell}$	μ_1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_k	$n_{k,1}$	$\cdots$	$n_{k,\ell}$	μ_k
Total	ν_1	$\cdots$	ν_ℓ	n

In this array, $n_{i,j} = |A_i \cap B_j|$ represents the number of instances that are found in **both** the cluster $A_i \in \mathscr{A}$ and $B_j \in \mathscr{B}$. The adjusted Rand index ($\in [-1,1]$) is given by

$$\text{ARI}(\mathscr{A},\mathscr{B}) = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{\mu_i}{2}\sum_j\binom{\nu_j}{2}\right]\Big/\binom{n}{2}}{\frac{1}{2}\left[\sum_i\binom{\mu_i}{2}+\sum_j\binom{\nu_j}{2}\right] - \left[\sum_i\binom{\mu_i}{2}\sum_j\binom{\nu_j}{2}\right]\Big/\binom{n}{2}},$$

which can also be re-written as

$$\text{ARI}(\mathscr{A},\mathscr{B}) = \frac{2(\text{ss}\cdot\text{dd} - \text{sd}\cdot\text{ds})}{(\text{ss}+\text{sd})(\text{ds}+\text{dd})+(\text{ss}+\text{ds})(\text{sd}+\text{dd})}.$$

[21]The formula for $\text{RI}(\mathscr{A},\mathscr{B})$ is not unlike the one used in the definition of accuracy, a performance evaluation measure for (binary) classifiers.

[22]In a nutshell, the expected value of $\text{RI}(\mathscr{A},\mathscr{B})$ for independent, random clusterings $\mathscr{A}$ and $\mathscr{B}$ is not 0 [49].

It is straightforward to compute RI and ARI for the two clusterings of the artificial example with 6 instances. Since $n = 6$, there are $6 \cdot 5/2 = 15$ pairs of observations in the data, and we have:

- ss $= 4$ (x_1 and x_3; x_1 and x_4; x_3 and x_4; x_2 and x_5);
- ds $= 2$ (x_2 and x_6; x_5 and x_6);
- sd $= 0$ (none);
- dd $= 9$ (all remaining pairs).

Thus,

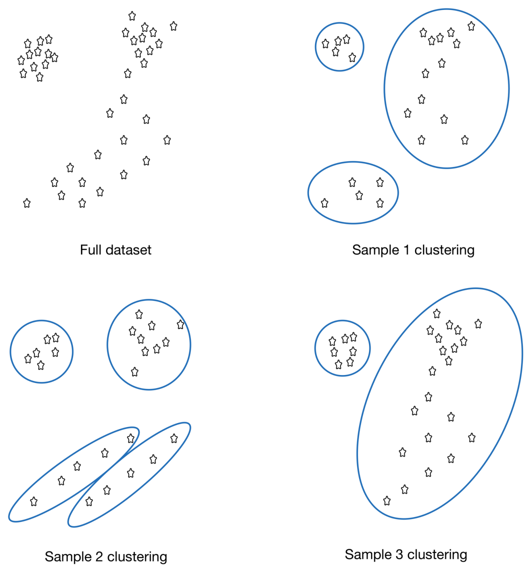
$$\begin{aligned}\text{RI}(\mathscr{A},\mathscr{B}) &= \frac{4+9}{4+9+0+2} = \frac{13}{15} = 0.87\\ \text{ARI}(\mathscr{A},\mathscr{B}) &= \frac{2(4\cdot 9 - 0\cdot 2)}{(4+0)(2+9)+(4+2)(0+9)} = 0.73.\end{aligned}$$

Both of these values are indicative of high similarity between the clustering outcomes $\mathscr{A}$ and $\mathscr{B}$.

Cluster **stability** can also be used to assess the appropriateness of the choice of algorithm for the data. Assume that we apply a clustering algorithm $\mathscr{G}$ to a dataset $\mathbf{X}$, resulting in a clustering scheme $\mathscr{C} = \{C_1, \ldots, C_K\}$.

In general, a dataset $\mathbf{X}$ is a **realization** of a (potentially unknown) underlying data generating mechanism related to the situation of interest; a different realization $\mathbf{X}'$, which could arise from the collection of new data, may yield a distinct clustering scheme $\mathscr{C}'$. How **stable** is the clustering outcome of $\mathscr{G}$, relative to the realization of $\mathbf{X}$?

We can get a sense for the underlying stability by sampling multiple row subsets from $\mathbf{X}$. Alternatively, we could also sample from $\mathbf{X}$'s columns, or sample columns from a variety of sampled rows of $\mathbf{X}$; however this is achieved, we have obtained ℓ subsets $\mathbf{X}_1, \ldots, X_\ell \subseteq \mathbf{X}$, on which we apply the algorithm $\mathscr{G}$, with parameters selection $\mathscr{P}$, yielding the corresponding clustering schemes $\mathscr{C}_1, \ldots, \mathscr{C}_\ell$, as below.



Let $\mathscr{W}$ be the similarity matrix for the various clustering schemes:

$$\mathscr{W} = \begin{pmatrix} w(\mathscr{C}_1,\mathscr{C}_1) & \cdots & (\mathscr{C}_1,\mathscr{C}_\ell) \\ \vdots & \ddots & \vdots \\ w(\mathscr{C}_\ell,\mathscr{C}_1) & \cdots & (\mathscr{C}_\ell,\mathscr{C}_\ell) \end{pmatrix};$$

this $\mathscr{W}$ can be used as the basis of a **meta-clustering scheme** in which $\mathscr{C}_1, \ldots, \mathscr{C}_\ell$ play the role of instances, using a graph-based meta-clustering method such as DBSCAN (which we will discuss in Section 4.1).

If the clustering results obtained from $\mathbf{X}$ by applying $\mathscr{G}$ are **stable**, we might expect the meta-clustering results to yield a **single meta-cluster**.

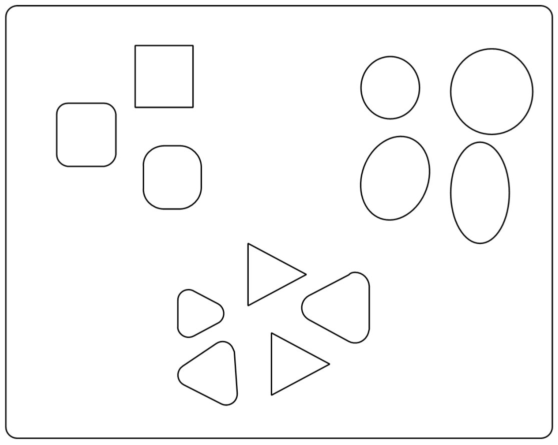
If multiple meta-clusters are found from $\mathscr{W}$, this supports the position that $\mathscr{G}$ does not produce stable clusters in $\mathbf{X}$, although this does not necessarily imply instability as the number of meta-clusters could be an artefact related to the choice of the meta clustering method.

This approach seems simple in theory, but in practice it often simply pushes the issues and challenges of clustering to the meta-clustering activity; a more sophisticated treatment of the problem of cluster stability assessment is presented in [51].

External Validation In everyday applications, we tend to gauge clustering results against some (often implicit) **external expectation** (or standard): we cannot help but bring in outside information to evaluate the clusters.

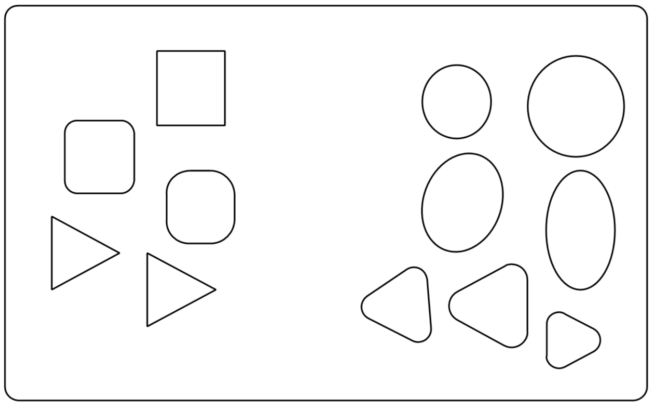
The outside information is typically what is deemed to be the 'correct' groups to which the instances belong, which is starting to look an awful lot like **classification**, a supervised learning approach.

In the example below, for instance, we might be interested in finding **natural groups** in the dataset of shapes, but we might hold the pre-conceived notion that the natural groups are linked to the underlying shape (square, triangle, circle).



Natural Groupings

If the outcome agrees with this (external) notion, we would naturally feel that the clustering was successful; if, instead the outcome seems to pick up something about the sharpness of the shapes' vertices, say (as in the image below),



Clustering Results

we might conclude that the algorithm does not a good job of capturing the essential nature of the objects, or, more rarely, that we need to revisit our pre-conceived notions about the dataset and its natural groups.

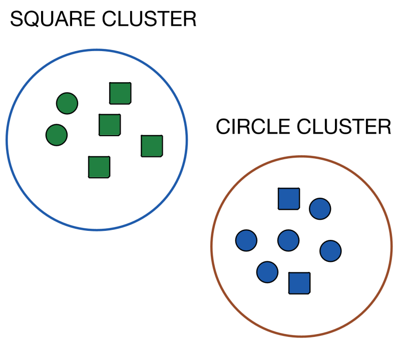
This validation strategy is often used to **build confidence** in the overall approach, based on preliminary or sample results, but it rests on a huge assumption (which often conflicts with the unsupervised learning framework), namely, that the natural groups in the data are **identifiable**, explicitly or implicitly.

Due to the conceptual similarity to classification,[23], **external validation measures** often resemble classification performance measures.

Case in point, the **purity metric**. In the presence of an external classification variable, this metric assign a label to a cluster C according to the most frequent classes of the instances in C, and

$$\text{purity}(C)=\frac{\#\text{ correctly assigned points in }C}{|C|},$$

as in the example below:



[23] Which it is emphatically not, it bears repeating.

The **clustering purity score** for $\mathscr{C}=\{C_1,\ldots,C_K\}$ is obtained as the average of the cluster purity scores, or as a weighted average of purity scores:

$$\text{weighted purity}(\mathscr{C})=\frac{1}{n}\sum_{\ell=1}^K|C_\ell|\cdot\text{purity}(C_\ell),$$

where n represents the number of instances in the data.

In the previous example, the green cluster is labeled as the square cluster (since 4 of its 6 instances are classified as squares), and the blue cluster is labeled as the circle cluster (since 5 of its 7 instances are classified as circles). At the cluster level, the purity scores are thus:

$$\text{purity}(C_\square)=\frac{2}{3},\quad \text{purity}(C_\bigcirc)=\frac{5}{7};$$

the average and weighted purity scores are

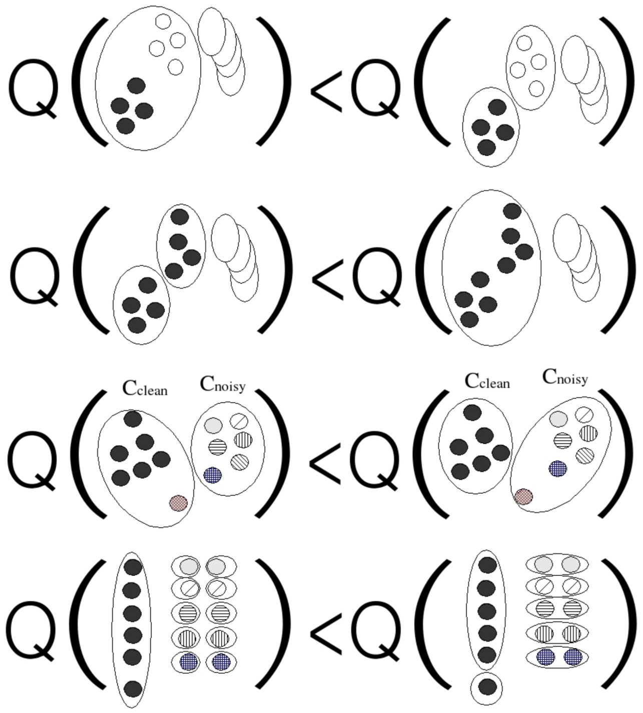
$$\text{average purity}(\mathscr{C})=\frac{1}{2}\left(\frac{2}{3}+\frac{5}{7}\right)=69.0\%$$

$$\text{weighted purity}(\mathscr{C})=\frac{1}{6+7}\left(6\cdot\frac{2}{3}+7\cdot\frac{5}{7}\right)=69.2\%.$$

These two numbers are very nearly identical because the clusters have roughly the same size; if the size variance is large, the two measurements would be quite different.

The purity is an obvious analogue to **accuracy**; other measures based on **precision** and **recall** are also popular [3].

Useful external quality metrics take advantage of the natural classes (if they are aligned with the clustering results), and take into account cluster **homogeneity** (first row, below), **completeness**, (second row), **noisy and outlying data** (third row), and **size vs. quantity** considerations (last row): the preferred behaviour is shown on the right [3].



Example Let us illustrate some of these notions using various k—means and hierarchical clusters of the Gapminder data used in Section 2. In all instances, we use the Euclidean distance on a scaled dataset.

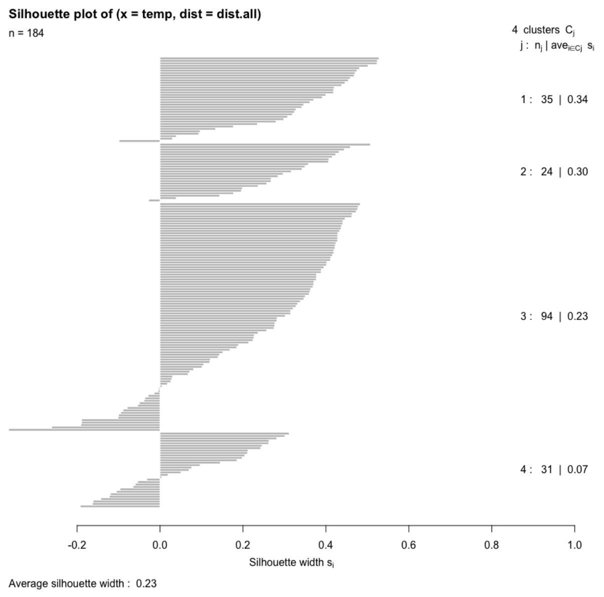
Internal Validation: we use the R packages `cluster`, `fpc`, and `clusterCrit` to compute 3 CQMs: the Dunn index, the average silhouette width, and the Calinski-Harabasz index.[24]

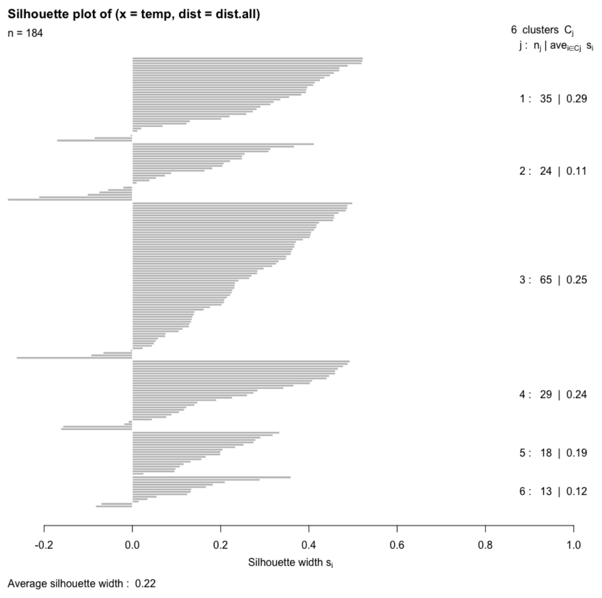
We start by clustering the data using 4—means; we then use hierarchical clustering with complete linkage and 3 clusters. The results are shown below:

method	Dunn	Avg. Sil.	C.-H.
4—means	0.097	0.315	139.0
HC(comp; 3)	0.091	0.274	125.4

Both of the Dunn values are low, although the 4—means result is marginally superior; while the average silhouette widths are also low, they are least positive in both cases, with a slight advantage in favour of 4—means; the Calinski-Harabasz values are not very indicative on their own, but once again, 4—means comes out ahead of HC.

The average silhouette width is an interesting metric. On the one hand, we attempt to gauge the quality of the **entire clustering** with a single number, but the average is a fickle summary measurement and potentially affected by outlying values; on the other, we do have access to a silhouette score **for each observation**, and can thus get a better idea of the performance by studying the **silhouette profile**. We show the silhouette results for hierarchical clustering with complete linkage for 4 (average width= 0.23) and 6 clusters (average width= 0.22).





While the average silhouette width would seem to favour the 4-cluster result, the profile for the 6-cluster result seems more in-line with desirable properties: in both instances, some observations are "mis-clustered" (negative silhouette scores), but these seem to be more broadly distributed in the latter case.[25]

Relative Validation: we compute the Rand index (RI) and the adjusted Rand index (ARI) to compare the outcomes of a single run of 2—means ($\mathscr{A}_2$), 3—means ($\mathscr{A}_3$), and 4—means ($\mathscr{A}_4$), respectively.

There are $\binom{184}{2} = 16836$ pairs of distinct observations in the dataset; the pair types and indices are shown below.

schemes	ss	dd	sd	ds	RI	ARI
$\mathscr{A}_2, \mathscr{A}_3$	5304	7032	3852	648	0.73	0.52
$\mathscr{A}_2, \mathscr{A}_4$	4395	6392	4761	1288	0.64	0.33
$\mathscr{A}_3, \mathscr{A}_4$	3754	8955	2198	1929	0.75	0.46

$\mathscr{A}_2, \mathscr{A}_3$ are relatively similar according to RI, as are $\mathscr{A}_3, \mathscr{A}_4$, but the ARI suggests that $\mathscr{A}_2, \mathscr{A}_3$ are more similar to one another than $\mathscr{A}_3, \mathscr{A}_4$ are; $\mathscr{A}_2, \mathscr{A}_4$ are not as similar, according to both indices, which is not that surprising as the number of clusters in this case jumps from 2 to 4.

External Validation: finally, we will compare the clustering results of hierarchical clustering, for 4 and 8 clusters, with a variety of external grouping: 6 world regions, as determined by the Gapminder Foundation, and OECD/G77 membership (see Figure 10).

Are there any reasons to suspect that the clusters would be aligned with these external classes?

[24] Ratio of the sum of **between-clusters dispersion** to the **inter-cluster dispersion** for all clusters; higher is better.

[25] in the 4-cluster case, half a cluster seems to have been mis-assigned, for instance..

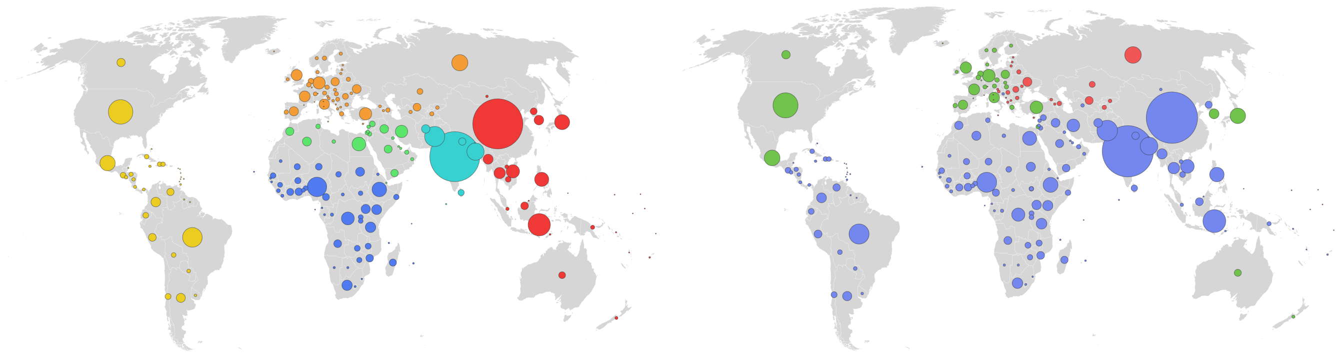


Figure 10. 6 world regions (left): America (yellow, 33 countries), East Asia Pacific (red, 26), Europe Central Asia (orange, 49), Middle East North Africa (green, 20), South Asia (turquoise, 8), Sub Saharan Africa (blue, 48); memberships (right): OECD (green, 30), G77 (purple, 128), other (red, 26); bubble size represents population [40].

For the 6 world regions classes, the clusters labels (the most frequent class per cluster) for HC(4) are

Cluster	Label	Size	Frequency
1	Sub Saharan Africa	35	31
2	East Asia Pacific	24	9
3	Europe Central Asia	94	42
4	Sub Saharan Africa	31	14

yielding a purity of 0.54 and a weighted purity of 0.52 – the overall score is not great, but the Sub Saharan countries are well captured by clusters 1 and 4.

The clusters labels for HC(8), on the other hand, are

Cluster	Label	Size	Frequency
1	Sub Saharan Africa	35	31
2	America	17	6
3	Europe Central Asia	65	34
4	East Asia Pacific	7	3
5	america	29	10
6	Sub Saharan Africa	18	7
7	East Asia Pacific	10	5
8	Sub Saharan Africa	3	3

yielding a purity of 0.55 and a weighted purity of 0.54; which is still not that great? Perhaps the clusters have little to do with geography, in the end.

Are they aligned with OECD/G77/other membership? The labels for HC(8) are

Cluster	Label	Size	Frequency
1	G77	27	27
2	G77	29	22
3	OECD	28	17
4	G77	20	18
5	G77	12	11
6	OECD	23	11
7	G77	25	24
8	G77	20	10

The purity values are 0.77 and 0.76, respectively: these are better values than the previous external validation attempts, but they might not really be meaningful given the preponderance of G77 countries in the data.

Neither of the external classifications seems to be a good gauge of cluster validity.

For the most part, the cluster validation yields middling results. The few algorithms we have tried with the data suggest that there is some low-level grouping at play, but nothing we have seen so far would suggest that the data segments are all that "natural."

While this result is somewhat disappointing, we should note that this is often the case with real-world data: there is no guarantee that natural groups even exist in the data.

However, we have not been directing our choices of algorithms and parameters – everything has been fairly arbitrary. Can anything be done to help with **model selection**?

3.2 Model Selection

How do we pick the number of clusters and the various other parameters (including choice of to use for "optimal" results?

A common approach is to look at all the outcomes obtained from various parameter runs and replicates (for a given algorithm), and to select the parameter values that optimize a set of QCMs, such as Davies-Bouldin, Dunn, CH, etc.

Optimization is, of course, dependent on each QCM's properties: in some cases, we are searching for parameters that maximizes the index, in other cases, those that minimize it, and yet in other cases, for "knees" or "change points" in various associated plots.[26]

[26]The parameter values that optimize a QCM may not optimize others; when they coincide, this reinforces the support for the existence of natural groups; when they do not, they provide a smaller collection of models from which to select, removing some of the arbitrariness discussed above.

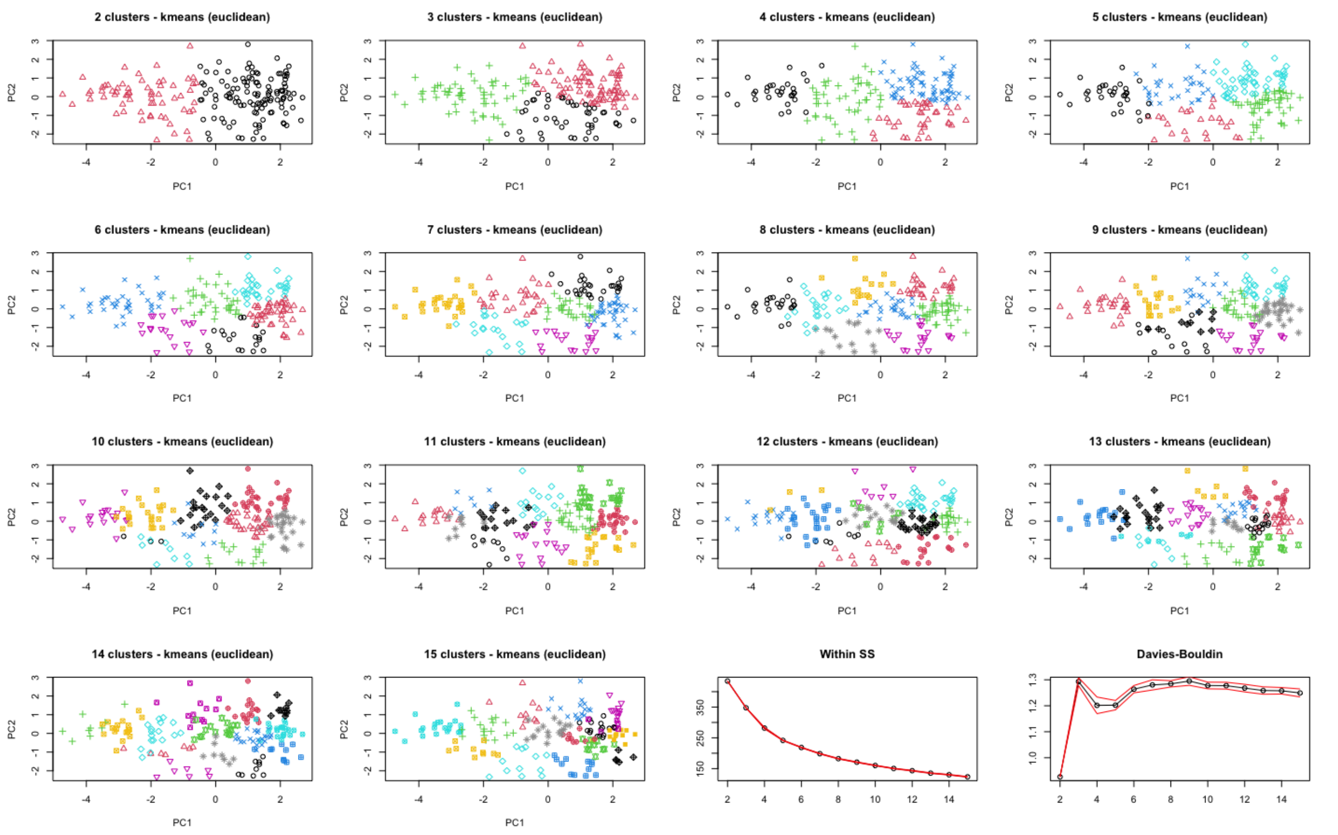


Figure 11. One realization of k—means on the (scaled) Gapminder dataset, for $k = 2, \ldots, 15$. The clusters are displayed on the first 2 principal components of the dataset, which explain 88% of the variation in the data. The average within sum of squares curve and the average Davies-Bouldin curves are also provided, with 95% confidence intervals. The Davies-Bouldin curve suggests that $k = 3$ or $k = 9$ would be good choices in this case.

This can also be done for clustering outcomes arising from different algorithms, although in this case we are not selecting parameter values so much as identifying the model that best describes the natural groups in the data among all results, according to some metric(s).[27]

It is important to remember that there is a lot of diversity in clustering validation techniques. The various types of validation methods do not always give concordant results; this variation within the types can be demoralizing at times, but it can also be leveraged to extract useful information about the data's **underlying structure**.

In general, we should avoid using a single assessment method; it is preferable to seek "signals of agreement" across a **variety of strategies** (both in the choices of clustering algorithms and evaluation methods).

[27]The metrics presented in Section 3.1 all provide frameworks to achieve this. There are additional approaches, such as: seeking the clustering $\mathscr{C} = \{C_1, \ldots, C_k\}$, among a list of such outcomes, which minimizes the **quadratic cost**

$$\Lambda_{\mathbf{W}}(\mathscr{C}) = -\text{trace}\left(Z^{\top}(\mathscr{C})\mathbf{W}Z(\mathscr{C})\right),$$

where $z_{i,\ell} = 1$ is $\mathbf{x}_i \in C_\ell$ and 0 otherwise, associated with a **similarity matrix W**; or methods relying on stability assessment [30, 51]. Model assessment and selection remains a popular research topic.

Finally, remember that clustering results may just be 'ok'... and that is ok too! We can study the situation and decide what is important and what can safely be ignored – as always, **a lot depends on context**.

Example How many clusters k should we seek when clustering the (scaled) Gapminder dataset using Euclidean distance? For each $k = 2, \ldots, 15$, we compute the outcome of $m = 40$ runs of k—means, and average the **within sum of squares** (WSS) and a (modified) **Davies-Bouldin** index (DBI) over the runs. The optimal number of parameters is obtained at a DBI maximum or a WSS "knee"; the results are shown in Figure 11.

The WSS curve does not yield much information, but the DBI curve suggests that both $k = 3$ and $k = 9$ could be good parameter choices. With parsimony considerations in mind, we might elect to use $k = 3$, but if the results are too simple or if there are signs of instability appear (recall that k—means is a stochastic algorithm), $k = 9$ might prove to be a better choice in the end.

4. Advanced Topics

We present representative clustering algorithms from the remaining families. More information is available in [2, 23, 53].

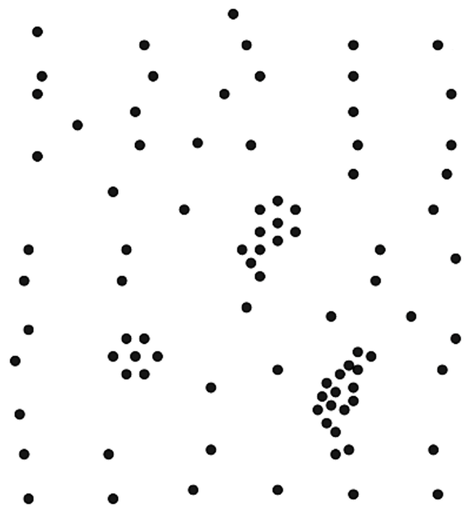
4.1 Density-Based Clustering

The assumptions of the k−means algorithm imply that the clusters that it finds are usually **Gaussian** (blob-like). But this is not always a desired outcome.

In **density-based clustering**, it is the **density** of observations and the **connectivity** of the accompanying clustering network (which we will discuss further in the next section) that determine the number and location of clusters. Popular density-based clustering algorithms include DBSCAN, DENCLUE, OPTICS, CHAMELEON, etc.

Once density has been defined in a meaningful way (which depends on a number of contextual factors), density-based algorithms are fairly straightforward to apply and understand (see [2, 22, 38, 41, 42] for details).

Density How do we measure density? Intuitively, we can recognize areas of **low density** and **high density** in the (artificial) dataset below.



As the saying goes, "birds of a feather flock together"; it should not come as a surprise that areas of higher density could be viewed as **clusters** in the data.

In that context, if $\mathbf{\Psi} \subseteq \mathbb{R}^n$ is an n−dimensional **sub-manifold** of $\mathbb{R}^n$, we could define the **density** of $\mathbf{\Psi}$ around $\mathbf{x}$ by, say,

$$\text{density}_{\mathbf{\Psi}}(\mathbf{x}; d) = \lim_{\varepsilon\to 0^+} \frac{\text{Vol}_n(B_d(\mathbf{x},\varepsilon)\cap\mathbf{\Psi})}{\text{Vol}_n(B_d(\mathbf{x},\varepsilon))},$$

where

$$B_d(\mathbf{x},\varepsilon) = \{\mathbf{y}\in\mathbb{R}^n \mid d(\mathbf{x},\mathbf{y}) < \varepsilon\}$$

and

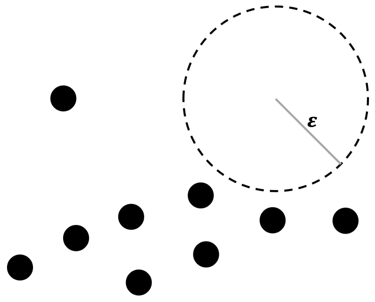
$$\text{Vol}_n(A) = \text{"volume" of } A \text{ in } \mathbf{R}^n.$$

DBSCAN In practice, the dataset $\mathbf{X}$ is usually a **discrete subset** of $\mathbb{R}^n$, and the limit definition above cannot apply.

Density-based spatial clustering of applications with noise (DBSCAN) estimates the density at an observations $\mathbf{x}\in\mathbf{X}$ as follows: we pick a "reasonable" value of $\varepsilon^* > 0$ and set

$$\text{density}_{\mathbf{X}}(\mathbf{x}; d) = |B_d(\mathbf{x},\varepsilon^*)\cap\mathbf{X}|.$$

The outcome depends, of course, on the choice of both ε^* and the distance d.



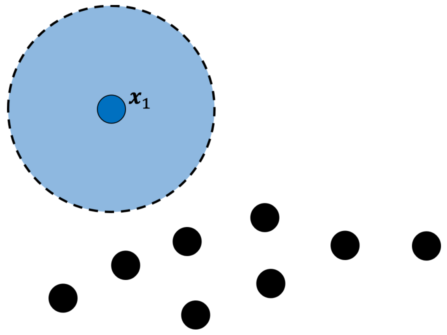
DBSCAN also requires a connectivity parameter: the **minimum number of points** *minPts* in

$$V_{\mathbf{x}} = B_d(\mathbf{x},\varepsilon^*)\cap[\mathbf{X}\setminus\{\mathbf{x}\}]$$

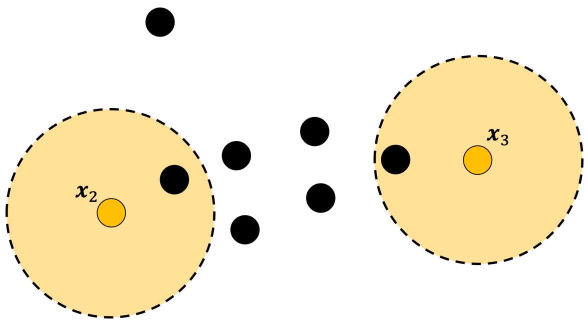
(excluding $\mathbf{x}$). If $|V_{\mathbf{x}}| \geq minPts$, the observations in $V_{\mathbf{x}}$ are said to be **within reach of** (or reachable from) $\mathbf{x}$.

In other words, for a given choice of d, ε^*, and *minPts*, there are three **types of observations** in $\mathbf{X}$:

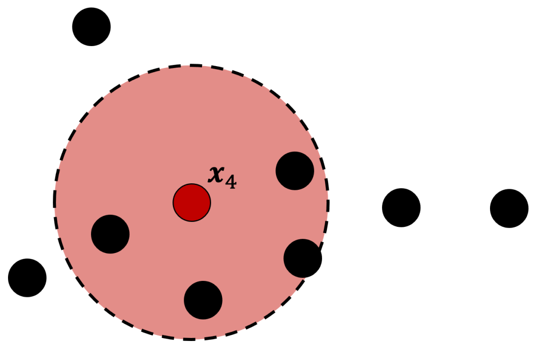
- **outliers** are observations that are not within reach of any of the other observations, such as $\mathbf{x}_1$ below:



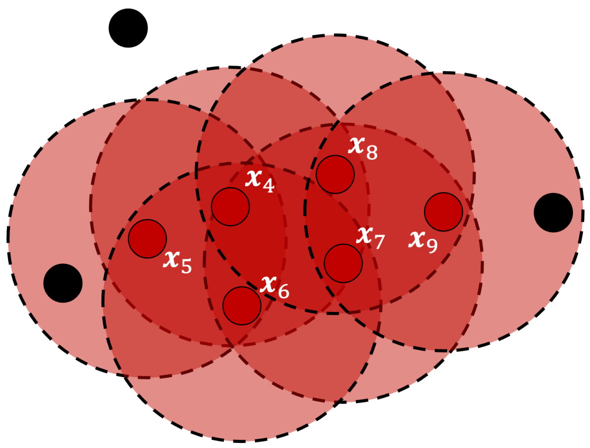
- **reachable** (non-core) **observations** are observations that are within reach of fewer than *minPts* other observations, such as $\mathbf{x}_2$ and $\mathbf{x}_3$ below (with $minPts = 3$):



- **core observations** are within reach of at least *minPts* other observations, such as $\mathbf{x}_4$ below (with $minPts = 3$):



There are other such points: $\mathbf{x}_5$, $\mathbf{x}_6$, $\mathbf{x}_7$, $\mathbf{x}_8$, and $\mathbf{x}_9$.



Reachability is **not a symmetric relation**: no observation is reachable from a non-core point (a non-core point may be reachable, but nothing can be reached from it).

We can build a new symmetric relation on non-outlying observations on the basis of reachability, however:

$$\mathbf{p},\mathbf{q} \in \mathbf{X} \setminus \{\text{outliers}(\mathbf{X})\}$$

are said to be **density-connected** for $\varepsilon^* > 0$ and d if there is an observation $\mathbf{o} \in \mathbf{X}$ such that $\mathbf{p},\mathbf{q} \in V_{\mathbf{o}}$, with $|V_{\mathbf{o}}| \geq minPts$.

The same $\mathbf{p},\mathbf{q}$ are said to be **density-connected in a path** if either they are density-connected or if there is a sequence of observations

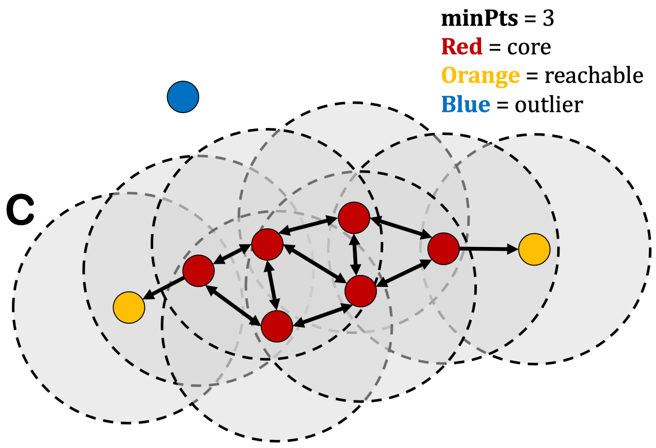
$$\mathbf{p} = \mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_{k-1}, \mathbf{r}_k = \mathbf{q}$$

such that $\mathbf{r}_{i-1}, \mathbf{r}_i$ is density-connected for all $i = 1, \ldots, k$.

That the latter is a relation on $\mathbf{X} \setminus \{\text{outliers}(\mathbf{X})\}$ is clear:

- it is **reflexive** as every $\mathbf{x} \in \mathbf{X} \setminus \{\text{outliers}(\mathbf{X})\}$ is either reachable or a core observation, so that $\exists \mathbf{o}_{\mathbf{x}} \in \mathbf{X}$ with $\mathbf{x} \in V_{\mathbf{o}_{\mathbf{x}}}$ and $|V_{\mathbf{o}_{\mathbf{x}}}| \geq$ minPts, and so $\mathbf{x}$ is density-connected to itself;
- it is **symmetric** and **transitive** by construction.

DBSCAN clusters are, essentially, composed of observations that are density-connected in a path.



In the image above, arrows represent density-connection.[28]

Algorithm DBSCAN clusters are grown according to the following algorithms:

1. Select an observation at random, from the list of not previously selected observations that have not been assigned to a cluster yet.
2. Determine the selected observation's type (outlier, non-core, core).
3. If the observation is an outlier or a non-core point, assign it to the **noise cluster**.
4. Else, build its network of density-connected paths.
5. Assign all observations in the network to a **stand-alone cluster**.
6. Repeat steps 1 to 5 until all points have assigned to a cluster.

All points within a cluster are **mutually density-connected in a path**. If a point is reachable from any point of the cluster, it is part of the cluster as well. An illustration of the DBSCAN algorithm is provided in Figure 12.

The observations in the noise cluster are typically identified as **outliers**, making DBSCAN a reasonable unsupervised learning approach for anomaly detection [12].

Note that clusters, by definition, must contain **at least** one core point. Small groups of observations that are not density-connected to any core points will then also be assigned to the noise cluster.

A non-core point that has been assigned to the noise cluster may end up being assigned to a stand-alone cluster at a later stage (but the opposite cannot occur).

It is possible for two clusters to **share** non-core points, in which case the points in question are randomly assigned (the random order of selection in step 1 may affect the results); consequently, some clusters may end up containing **fewer than** *minPts* observations.

[28] Each orange observation is within reach of a red one, but no observation can be reached from the orange points.

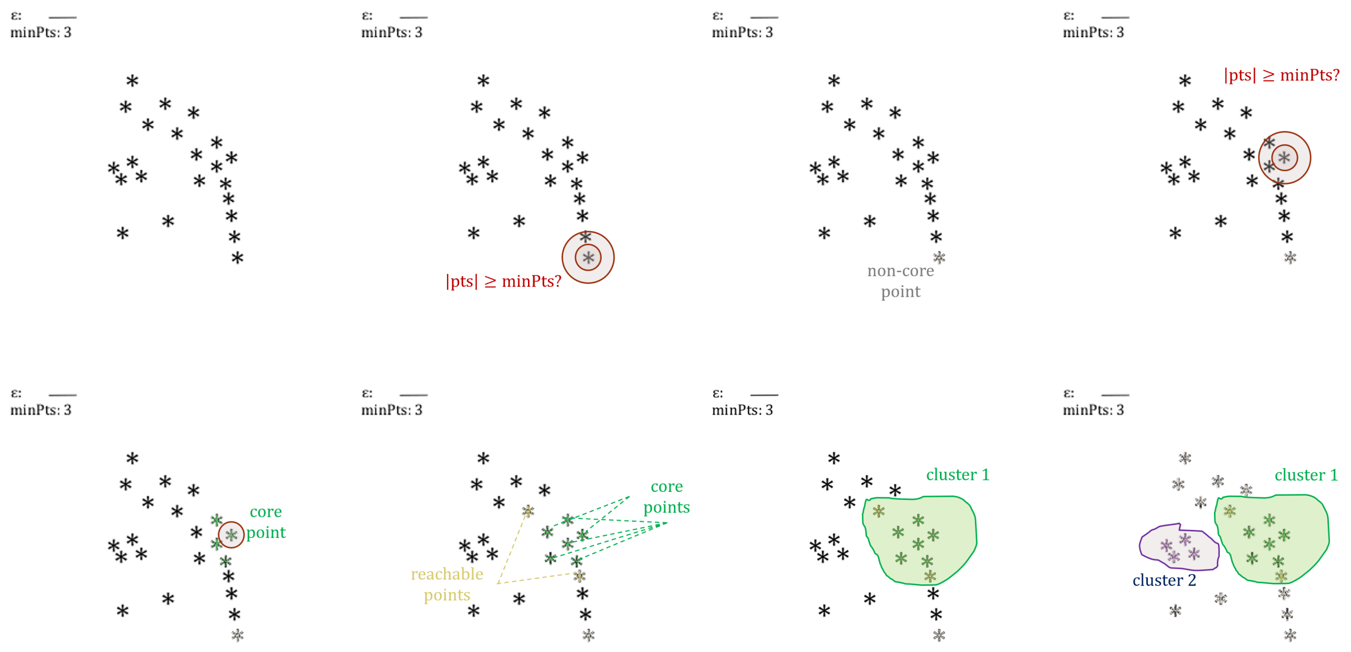


Figure 12. Illustration of DBSCAN on an artificial dataset (top, left). The parameters ε and *minPts* are shown in each display. We select a point at random (second image, top row); it is not a core point as its $\varepsilon-$neighbourhood does not contain more than *minPts* observations (excluding the selected point itself); it is assigned to the noise cluster. We select another point at random (top, right); that one is core point, as its $\varepsilon-$neighbourhood contains 4 observations. All its density-connected observations are shown in green (bottom, left). Its network of density-connected paths is shown in green, for the core observations, and in light green, for the reachable observations (bottom row, second image); they make up cluster 1 (bottom row, third image). Continuing on this way, we obtain 2 clusters and noisy observations (bottom, right).

Advantages and Limitations The main advantages of DBSCAN are that:

- there is no need to specify the **number** of clusters to find in the data;
- clusters of **arbitrary shapes** can be discovered;
- observations that are "noisy"/outlying are not forced into a cluster;
- the clusters are **robust** with respect to outliers, and
- it only requires two parameters ($\varepsilon^* > 0$ and *minPts*) to run properly, which can be set by domain experts if the data is well understood.

In general, it is suggested to use $minPts \geq p+1$, with larger values being preferable for **noisy** data sets, or $minPts \geq 2p$ for large datasets or sets with duplicates.

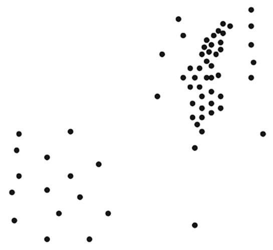
The choice of $\varepsilon^* > 0$ should take into account that if it is too small, a large portion of the observations will be **assigned to the noise cluster**; but if it is too large, a majority of observations will be found in the **same cluster**. Small values are preferable, but how small is too small?

The parameter choices have a large impact on the DBSCAN results, as does the choice of the distance function, which should take place **before** ε^* is selected to avoid **data dredging** and "begging the question".

Given that DBSCAN can handle globular clusters as well as non-globular clusters, why would we not always use it?

One important reason relates to **computational efficiency**. For a dataset $\mathbf{X}$ with n observations, the basic $k-$means algorithm has order $O(nk)$, whereas the most efficient versions of DBSCAN algorithm has order $O(n \log n)$. Thus, when n increases, the DBSCAN run time increases faster than the $k-$means run time.

Another reason is that DBSCAN works well when the density of clusters is assumed to be **constant**.



Most of us would agree that there are **two clusters** in the image above (a loose one in the bottom/left corner, and a tight one in the top/right corner), as well as some **outliers** (around the tight cluster), but no combination of $\varepsilon^* > 0$ and *minPts* can allow DBSCAN to discover this structure: either it finds no outliers, or it only finds the one tight cluster.

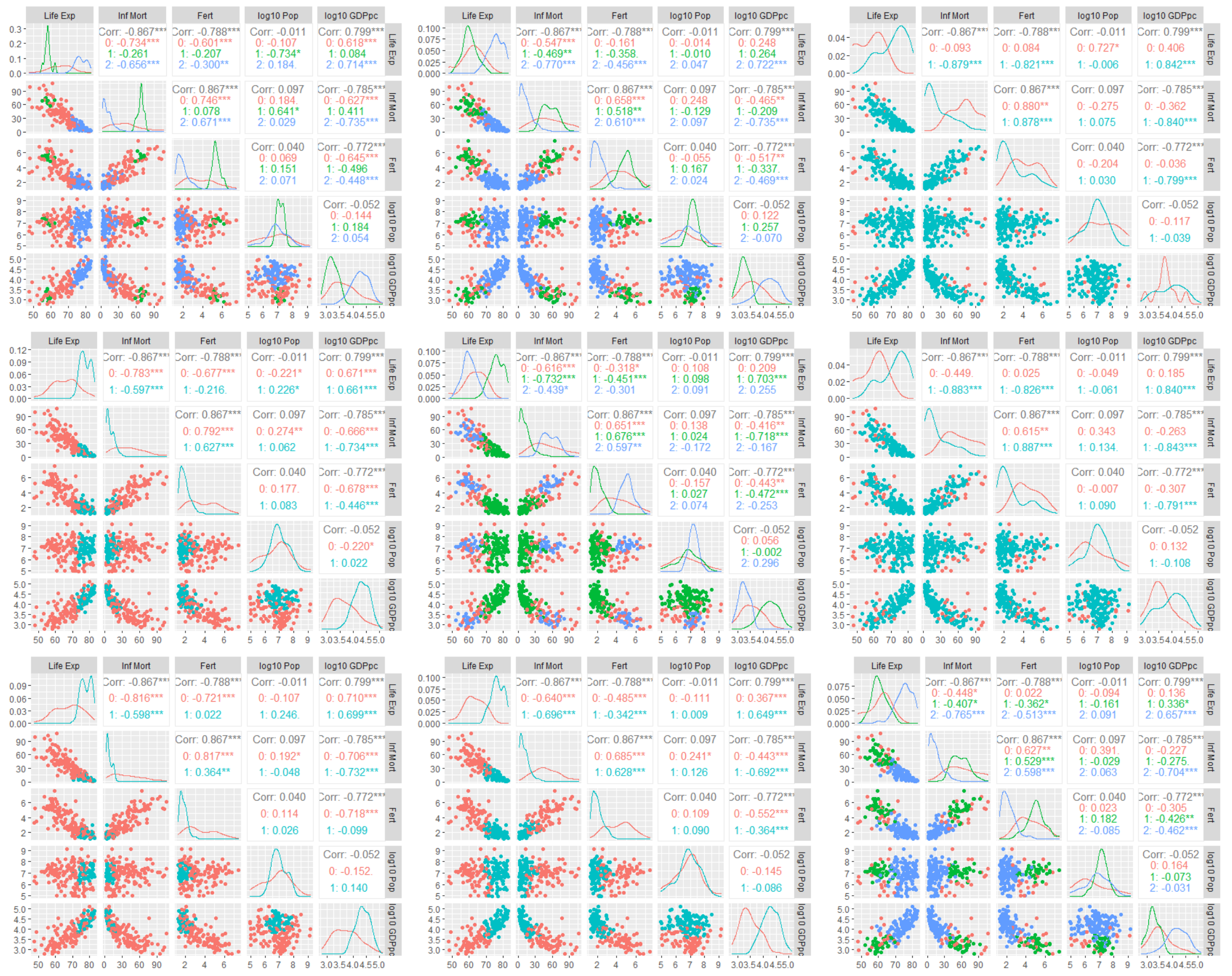


Figure 13. A scatter plot matrix of the realizations of DBSCAN on the 2011 Gapminder dataset. We select parameters from $\varepsilon^* \in \{0.75, 1, 1.25\}$ (left, centre, right columns) and *minPoints* $\in \{6, 10, 15\}$ (top, middle, bottom row). All clustering outcomes are obtained on a scaled dataset, using Euclidean distance. The noisy observations are always shown in red.

Example We show the results of DBSCAN on the Gapminder 2011 data, using Euclidean similarity (see Figure 13) for 9 combinations of parameters

$$\varepsilon^* = \{0.75, 1, 1.25\} \times \{6, 10, 15\} = minPts.$$

The noisy observations are shown in red: one immediate insight is that the number of outlying observations **decreases** as ε^* increases, which is as expected. Another insight is that the number of noisy observations **increases** as *minPts* increases, which is again not surprising.

If we compare the shape of the DBSCAN clusters with those of the k—means and HC clusters, we notice that the option of identifying observations as noisy – coupled with the "right" combination of parameters – creates "reasonable" clusters, that is to say, clusters for which do not have to stretch our ideas about what clusters ought to look like: the problematic observations (like China and India) are simply explained away as outliers.

The various runs find either 1 or 2 stand-alone clusters (as well as noisy observations), but that can change if we use different parameter values.

We can also determine if the cluster observations are core or non-core observations. For instance, in the realization with $\varepsilon^* = 1$ and $minPts = 6$, we have

	noise	**cluster 1**	**cluster 2**
outlier	34	–	–
reachable	–	10	17
core	–	20	103
total	**34**	**30**	**120**

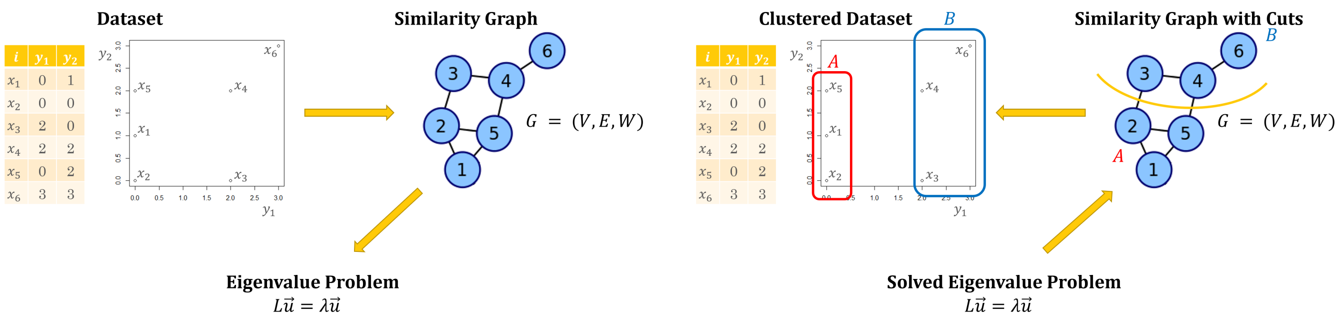

Dataset
Similarity Graph
$G = (V, E, W)$
Clustered Dataset
Similarity Graph with Cuts
A
B
Eigenvalue Problem
$L\vec{u} = \lambda\vec{u}$
Solved Eigenvalue Problem
$L\vec{u} = \lambda\vec{u}$

Figure 14. Schematics of spectral clustering. We extract the similarity graph of a dataset, which gives rise to an eigenvalue problem (left). The eigenenvalue problem is then solved, which suggests an 'optimal' graph cut, which in turns leads to data clusters (right).

4.2 Spectral Clustering

At a fairly coarse level, clustering algorithms are divided among those focusing on **compactness** and those focusing on **connectivity**.

Compactness methods are usually variants of k **Nearest Neighbours** (kNN) methods [9], and are effective when there are distinct clumps in the data. We can make specific assumptions about the distribution of the different clusters ahead of time (see the next two sections), but compact methods struggle to achieve meaningful results in scenarios where groups are not **linearly separable**.

In cases where we have little to no knowledge of the dataset, making assumptions about the distributions of clusters can lead to invalid clustering schemes; in such cases, connectivity-based methods have been shown to work reasonably well [23, 34].

Connectivity methods, such as DBSCAN, focus on dividing observations into groups based on their **similarity graphs**; observations that are quite different in their features (and as such would be differentiated using compactness methods) may end up in the same cluster if there is a chain of sufficiently similar observations linking them.

Connectivity methods require fewer initial assumptions, but their use can be harder to justify mathematically. The validity of such methods can only be determined *post hoc*.

Spectral clustering is a connectivity method that has become quite popular in practice; in a nutshell, we transform the dataset into its **similarity graph** and convert the latter into an **eigenvalue problem**. We then solve the eigenvalue problem, convert the solution into a **graph cut**, and then translate the cut back into dataset **clusters** (see Figure 14).

Before we start delving into the spectral clustering algorithm, we must discuss a few concepts relating to graphs and linear algebra.[29]

[29]These concepts are covered in just enough depth to provide an intuition about the algorithm.

Graphs and Cuts A **graph** is an object which connects **nodes** (or **vertices**) together through **edges**. [30] The edges have **weights** and can also be **directed**.[31] In certain cases, we may assume that all edge weights are identical and bidirectional, which is equivalent to saying that the edges just represent that a relationship exist.

The link with clustering is that once a similarity measure w has been selected, a dataset can be represented by a **similarity graph** $G = (V, E, W)$:

1. observations $\mathbf{x}$ correspond to **vertices** $v \in V$;
2. if $i \neq j$, vertices $v_i, v_j \in V$ are connected by an **edge** $e_{i,j} = 1$ if the **similarity weight** $w_{i,j} = w(\mathbf{x}_i, \mathbf{x}_j) > \tau$ for a predetermined **threshold** $\tau \in [0, 1)$, and by no edge ($e_{i,j} = 0$) otherwise;
3. the edges ($e_{i,j}$) m the **adjacency matrix** E;
4. the weights ($w_{i,j}$) form the **similarity matrix** W;
5. the **(diagonal) degree matrix** D provides information about the number of edges attached to a vertex: $d_{i,i} = \sum_{j=1}^{n} e_{i,j}$.

Note that, by convention, $w_{i,i} = 0$ for all i. For instance, we could use the Gower similarity measure

$$w(\mathbf{x}_i, \mathbf{x}_j) = 1 - \frac{1}{p}\sum_{k=1}^{p} \frac{|x_{i,k} - x_{j,k}|}{\text{range of } k\text{th feature in } \mathbf{X}}$$

for the dataset found in Figure 14; the ranges of X_1 and X_2 are both $r_1 = r_2 = 3$, so that

$$w_{3,4} = w_{4,3} = w(\mathbf{x}_3, \mathbf{x}_4) = 1 - \frac{1}{2}\left(\frac{|x_{3,1} - x_{4,1}|}{r_1} + \frac{|x_{3,2} - x_{4,2}|}{r_2}\right)$$
$$= 1 - \frac{1}{2}\left(\frac{|2-2|}{3} + \frac{|0-2|}{3}\right) = 1 - \frac{1}{2}\cdot\frac{2}{3} = \frac{2}{3};$$

[30]Airports (vertices) and flight paths (edges) form a graph in transportation networks, as do people (vertices) and relationships (edges) in social networks.

[31]In the transportation network example, the edges can be weighted according to flight frequency and/or directed according to their origin and destination, say; in the social network example, they could be weighted according to frequency of communication and/or directed according to who follows who on some app.

the similarity matrix as a whole is

$$W=\begin{pmatrix} 0 & 5/6 & 1/2 & 1/2 & 5/6 & 1/6 \\ 5/6 & 0 & 2/3 & 1/3 & 2/3 & 0 \\ 1/2 & 2/3 & 0 & \mathbf{2/3} & 1/3 & 1/3 \\ 1/2 & 1/3 & \mathbf{2/3} & 0 & 2/3 & 2/3 \\ 5/6 & 2/3 & 1/3 & 2/3 & 0 & 1/3 \\ 1/6 & 0 & 1/3 & 2/3 & 1/3 & 0 \end{pmatrix}.$$

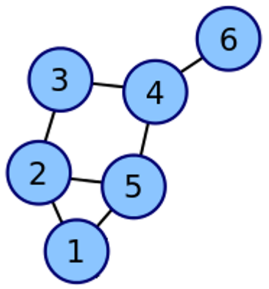
If we use a threshold $\tau = 0.6$, then the adjacency matrix is

$$E=\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix},$$

and the degree matrix is

$$D=\begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

The degree matrix can also be read directly from the similarity graph (which depends on the threshold τ):



A **graph cut** is the process by which we remove edges from the graph and separate the vertices into into groups (or **sub-graphs**).

The clustering task is to separate the nodes into multiple groups by **minimizing the total weight of the edges** we have to break in the process (i.e., making sure that the groups are as dissimilar as possible). This is also known as the **minimum cut problem** (Min Cut).[32]

This task is NP-Hard, meaning there is no theoretically guaranteed efficient way to do so, in comparison to simply testing every possible cut and finding the minimum weight.

[32]This cannot be the whole story, however, as we can minimize the total weight of the edges by simply not cutting any edges. Indeed, there are other approaches:

- **Normalized Cut** (actually used in practice)
- Ratio Cut
- Min-Max Cut

This is problematic: for datasets with n observations, the number of cuts is **bounded below** by 2^n (when we only consider 2–cuts); when n is relatively small, the overall number of cuts to consider remains manageable, but for nearly all reasonable datasets, this becomes an exercise in futility.

The spectral clustering approach generalizes the Min Cut problem (or any of the other problems) by imposing some properties on the similarity graph to ensure that we can get approximate the true Min Cut solution in a **computationally efficient** manner.[33]

Formally, the Min Cut problem involves finding a **partition** $\{A_1,\ldots,A_k\}$ of G which minimizes the objective function

$$\text{Cut}(A_1,\ldots,A_k)=\frac{1}{2}\sum_{i=1}^{k}\mathcal{W}(A_i,\overline{A_i})$$

where

$$\mathcal{W}(A,B):=\sum_{i\in A,j\in B} w_{i,j}$$

and $\overline{A}$ is the complement of A. The factor $\frac{1}{2}$ is used to remove double-counted edges.

Normalized Cut The spectral clustering approach solves the **Normalized Cut** (NCut) problem, which is similar to the Min Cut problem except that we are minimizing the weight of edges escaping a cluster relative to the total weight of the cluster.[34]

In the NCut problem, the **objective function** is

$$J_{\text{NCut}}(A,B)=\text{Cut}(A,B)\left(\frac{1}{\text{Vol}(A)}+\frac{1}{\text{Vol}(B)}\right),$$

where

$$\text{Vol}(C)=\sum_{i\in C} w_{i,*};$$

in a first pass, we seek to **minimize** J_{NCut} against the set of **all possible partitions** (A,B) of G. The procedure can be repeated as necessary on the cluster sub-graphs.

Intuitively, J_{NCut} is small when the observations **within** each sub-graph are **similar** to one another ($\text{Vol}(A),\text{Vol}(B)$ are large) and the observations **across** are dissimilar to one another ($\text{Cut}(A,B)$ is small).

On the plus side, NCut takes into consideration the **size of the partitioned groups** and **intra-group variance**, and tends to avoid isolating vertices, but it is not any easier to solve than the Min Cut problem. So why do we even bring it up in the first place?

As it happens, we can provide an approximation to the NCut solution using **purely algebraic means**.

[33]The spectral Min Cut solution is not guaranteed to be the true Min Cut solution, but it might be close enough to be an acceptable approximation.

[34]For more information about this abstraction, which actually relates a variant of Kernel PCA to spectral clustering, consult [7].

Similarity and Degree Matrices, Revisited There are different ways to construct a **graph** representing the relationships between the dataset's observations. We can use:

- **fully connected** graphs, where all vertices having non-zero similarities are connected to each other;
- $r-$**neighbourhood** graphs, where each vertex is only connected to the vertices falling inside a ball of radius r (according to some distance metric d), where r has to be tuned in order to catch the local structure of data;
- k **nearest neighbours** graphs (and variants), where each vertex is connected to its k nearest neighbours (again, according to some distance metric d), with k pre-selected, and
- **mixtures of** $r-$**neighbourhood and** k**NN** graphs, to better capture sparsity in the data.

The similarity measure w is usually picked from a list that includes: Gaussian (most common), cosine, fuzzy, Euclidean, Gower, etc.

The similarity matrix W is symmetric and has zeros along the diagonal; its non-diagonal entries represent the **similarity strength** between the corresponding graph vertices (and so the corresponding observations in the dataset).

We have discussed previously how to build the adjacency matrix E from W and a threshold $\tau \in [0,1)$.

The only component of a graph that similarity matrices do not directly capture are the **degrees** of each vertex, the number of edges that enter it (we are viewing the similarity graph as **undirected**). The diagonal of the degree matrix D holds that information for each vertex.

We can combine W and D (or E and D) to create a matrix L known as the **Laplacian**, which has properties linked to the topology of the similarity graph.

Laplacian The **Laplacian** of a graph is defined by

$$L_0 = D - \Theta, \quad \Theta \in \{E, W\};$$

The **symmetric Laplacian** by

$$L_S = D^{-1/2} L D^{-1/2} = \mathbf{I}_n - D^{-1/2}\Theta D^{-1/2},$$

and the **asymmetric Laplacian** by

$$L_A = D^{-1} L = \mathbf{I}_n - D^{-1}\Theta.$$

In all cases, the **off-diagonal entries** are non-positive, and the **diagonal entries** contain the degree of each node.

The Laplacians have the following useful properties:

- L_0, L_S are symmetric; L_A is not necessarily so;[35]
- all their eigenvalues are real and non-negative;
- every row and column adds up to 0, which means that $\lambda_0 = 0$ is the smallest eigenvalue of each Laplacian (hence they are singular and cannot be inverted);
- the number of connected components in the graph is the **dimension of the nullspace** of the Laplacian associated to $\lambda_0 = 0$ (which may provide a first approximation to the number of clusters in **X**), and
- the second smallest eigenvalue gives the graph's **sparsest cut**.[36]

[35]The product of symmetric matrices is not necessarily symmetric.

Spectral Clustering Algorithm In the case of two clusters, the objective function J_{NCut} is minimized when finding the eigenvector **f** corresponding to the smallest **positive** eigenvalue of L_S, also known as the **spectral gap**.

The clustering in the original data is recovered by sending $\mathbf{x}_i$ to A when $f_i > 0$ and $\mathbf{x}_j$ to B otherwise. This deterministic algorithm is a special case of the **spectral clustering algorithm** [50].

To divide **X** into k clusters, we follow the steps below:

1. form a similarity matrix W and a degree matrix D, using a threshold $\tau \in [0,1)$;
2. construct a Laplacian matrix L_ξ, using $\Theta = W$;
3. compute the first k eigenvectors $\{\mathbf{u}_1, ..., \mathbf{u}_k\}$ of L_ξ corresponding to the k **smallest positive** eigenvalues of L_ξ;
4. construct the $n \times k$ matrix **U** containing the vectors $\{\mathbf{u}_1, ..., \mathbf{u}_k\}$ as **columns**;
5. normalize the rows of **U** into a matrix **Y** with rows $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ having unit length;
6. cluster the rows of **Y** into k clusters;
7. assign $\mathbf{x}_i$ to cluster j of **X** if $\mathbf{y}_i$ was assigned to cluster j in the preceding step.

Spectral clusters for the dataset of Figure 14 using the Laplacian and symmetric Laplacian are shown in Figure 15.

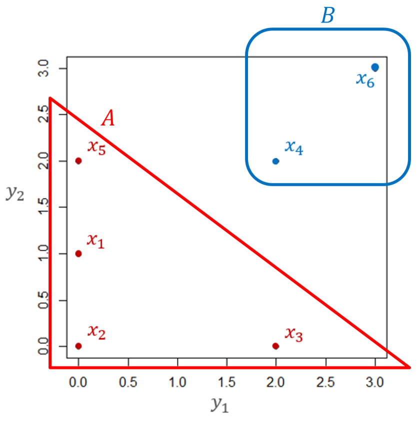
From an experimental perspective, spectral clustering provides an attractive approach because it is easy to implement and reasonably fast, especially for sparse data sets: it is a **graph partitioning problem** that makes no initial assumptions on the form of the data clusters.

Spectral clustering has variants, which depend on the many choices that can be made at various points in the process:

1. **pre-processing** (choice of: number of cluster k, similarity measure w, threshold τ);
2. **spectral representation** (choice of Laplacian);
3. **clustering algorithm** (choice of compact based, potentially non-deterministic algorithm to unleash on the rows of the representation **Y**).

[36]This is not the same as the minimum cut which represents the cut that minimizes the number of edges separating two vertices, but instead represents the minimum ratio of edges across the cut divided by the number of vertices in the smaller half of the partition.

i	f_i	A/B
x_1	-0.39	A
x_2	-0.30	A
x_3	-0.18	A
x_4	0.03	B
x_5	-0.16	A
x_6	0.84	B



i	f_i	A/B
x_1	0.36	B
x_2	0.54	B
x_3	-0.03	A
x_4	-0.43	A
x_5	0.14	B
x_6	-0.61	A

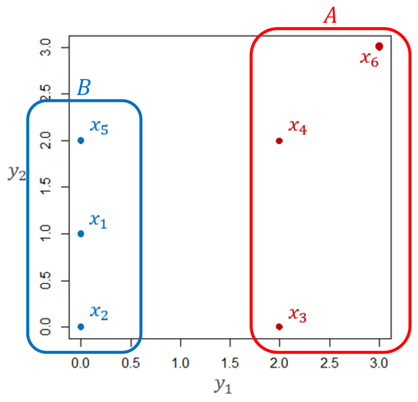


Figure 15. Two clusters for the dataset of Figure 14: simple Laplacian (left); symmetric Laplacian (right).

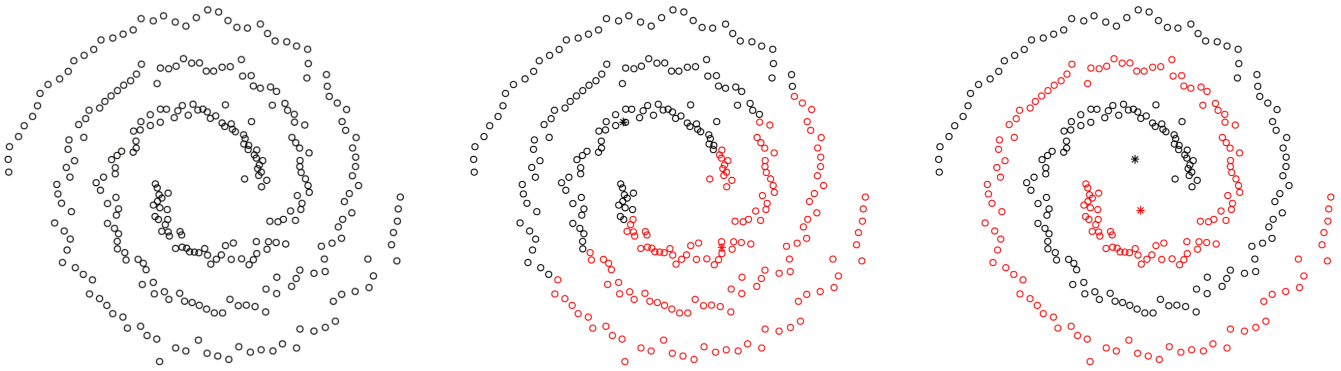


Figure 16. Comparing 2—means (middle) and spectral clustering with $k = 2$ (right) on the spirals dataset (left).

The **NJW algorithm** uses L_S for the spectral representation and k—means as a clustering approach. It can be interpreted as **kernalized** k**—means**: if we select a kernel which transforms the points to their mapped value in the Laplacian of the graph, then we (almost directl) obtain spectral clustering [7].[37] In Figure 16, the different outcomes of k—means and NJW are illustrated on the spirals dataset.

Practical Details and Limitations The most obvious practical detail in the implementation of spectral clustering is related to the construction of the similarity graph. In general, there is virtually no theoretical justification for determining what type of clustering approach to use; even when an approach has been selected, it can be quite difficult to choose appropriate parameter values.

In spectral clustering, there are considerations in favour of using **sparse similarity/adjacency matrix**: we seek to strike a balance between a Laplacian which is too densely connected, and one for which almost all of the observations are seen as dissimilar to one another.

Another issue relates to the computational challenge of **finding the eigenvalues** of the Laplacian. This can be done relatively efficiently if the matrix is sparse enough, however, which suggests using a relatively-high threshold τ; there are methods which help spectral clustering automatically tune for the best parameter values (including τ), but they take up a significant amount of resources [50].

Spectral clustering methods are extremely effective because they do not require assumptions about **distributions** and **centers**, are fairly **easy to implement**, and are **transparent** and **interpretable**.

However, they suffer from some of the same drawbacks as other clustering methods, namely when it comes to:

- selecting **initial parameter values**,
- run-times that **do not scale** with larger datasets, and
- determining optimal ways to visualize the results.

As in all clustering scenarios, analysts are faced with decisions at various levels of the process; they must be prepared to run multiple algorithms, in multiple configurations, in order to get a sense for the data structure (some strategies specific to spectral clustering are presented in [50]).

[37] DBSCAN can also fit within that framework, by picking a similarity method based on the radius that allows the graph separate into different components. Then the multiplicity of $\lambda_0 = 0$ in the Laplacian gives the number of graph components, and these can be further clustered, as above.

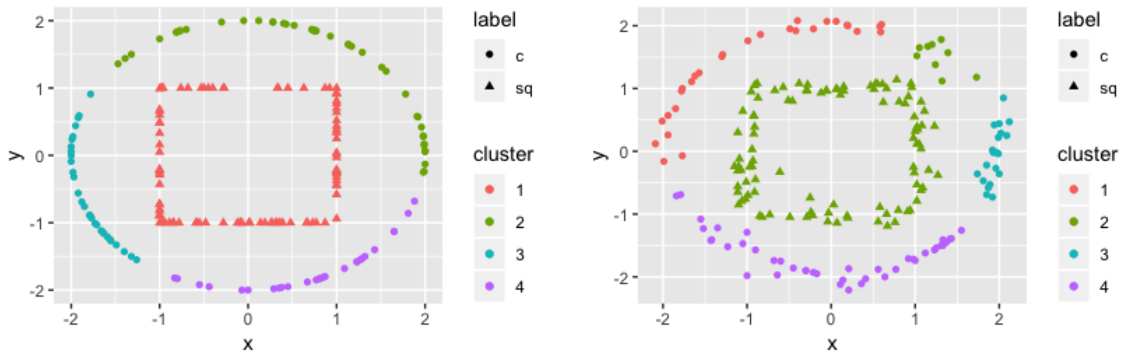


Figure 17. Comparison of spectral clustering on perfect and noisy data.

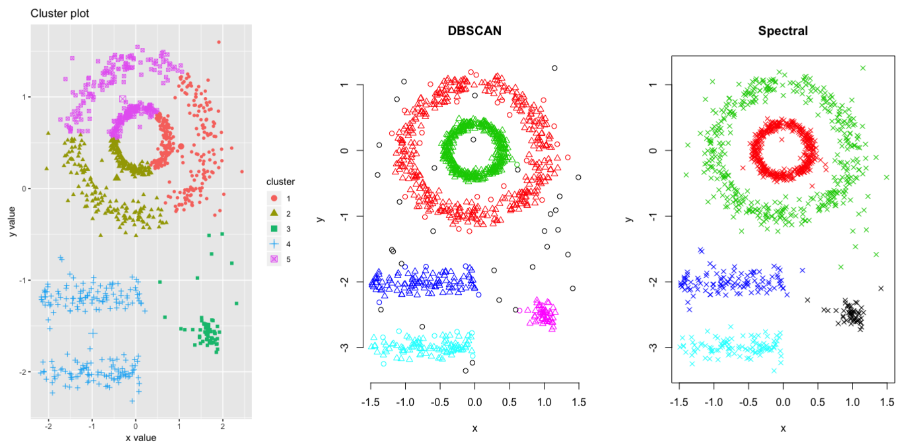


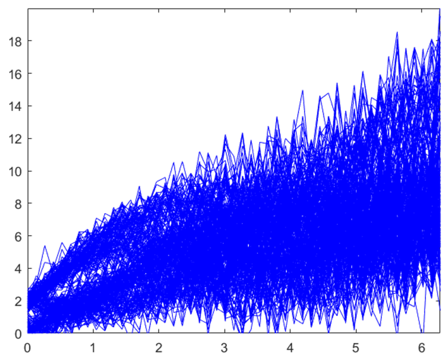
Figure 18. Comparison of Kmeans, DBSCAN, and spectral clustering on random shapes.

Examples In a first example, we look at an artificial dataset (a square within a circle), and a noisy version of the same. We run the NJW algorithm and look for 4 clusters; the impact of the noise can be seen in Figure 17.

In the second example, we compare the clustering outcome on a classic dataset containing multiple 2D shapes, using 5−means, DBSCAN, and NJW with $k = 5$ clusters. The shapes are relatively disconnected, so both spectral clustering and DBSCAN perform similarly , as expected – the only difference is that DBSCAN identifies observations that are far away from everything as outliers, whereas NJW assigns all of these points to one of the clusters; the outcomes are shown in Figure 18.

In the third example, spectral clustering is used to segment greyscale images into different segments based on contrasting colours [47]. Figure 19 shows instances with high contrast, with fairly decent segmentation performance using NCut, Self-Tuning SC [58], and a proposed SC algorithm [47]; Figure 20 shows other instances with less contrast (resulting in a poorer segmentation with the same methods); Figure 21 shows the comparison in segmentations using the proposed SC algorithm when the same image is presented at different resolutions.

In the fourth example, consider a dataset of $n = 250$ times series, with $N = 60$ entries each.



The distance d between 2 time series is the **average absolute gap between series**:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{60}\sum_{\ell=1}^{60} |x_{i,\ell} - x_{j,\ell}|.$$

We build the **Gaussian similarity** measure

$$w(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{d^2(\mathbf{x}_i, \mathbf{x}_j)}{2\sigma^2}\right).$$

We use the following parameter values

$$\sigma^2 = 300, \quad \tau = 0.9, \quad k = 5.$$

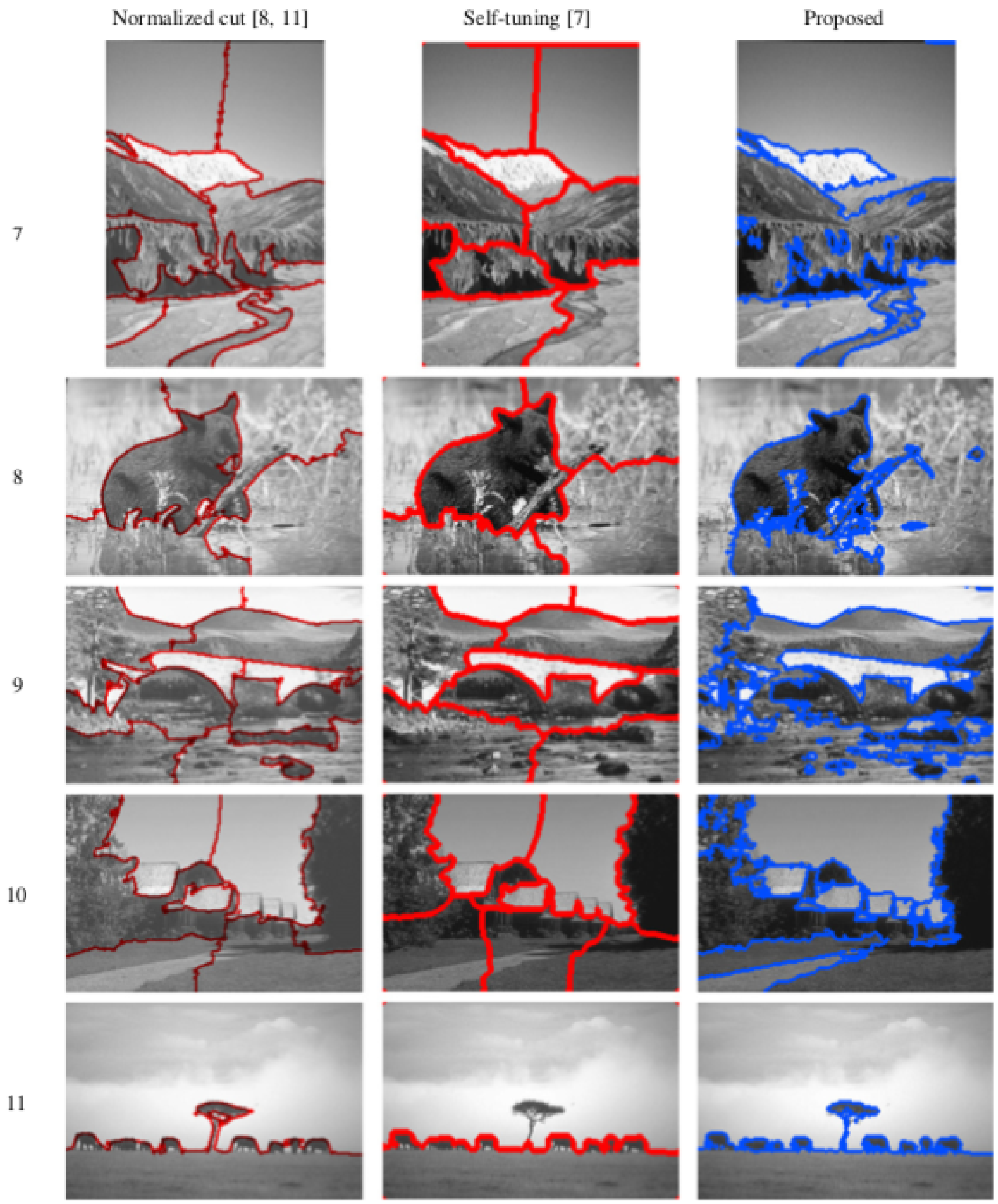


Figure 19. High contrast image segmentation with spectral clustering [47].

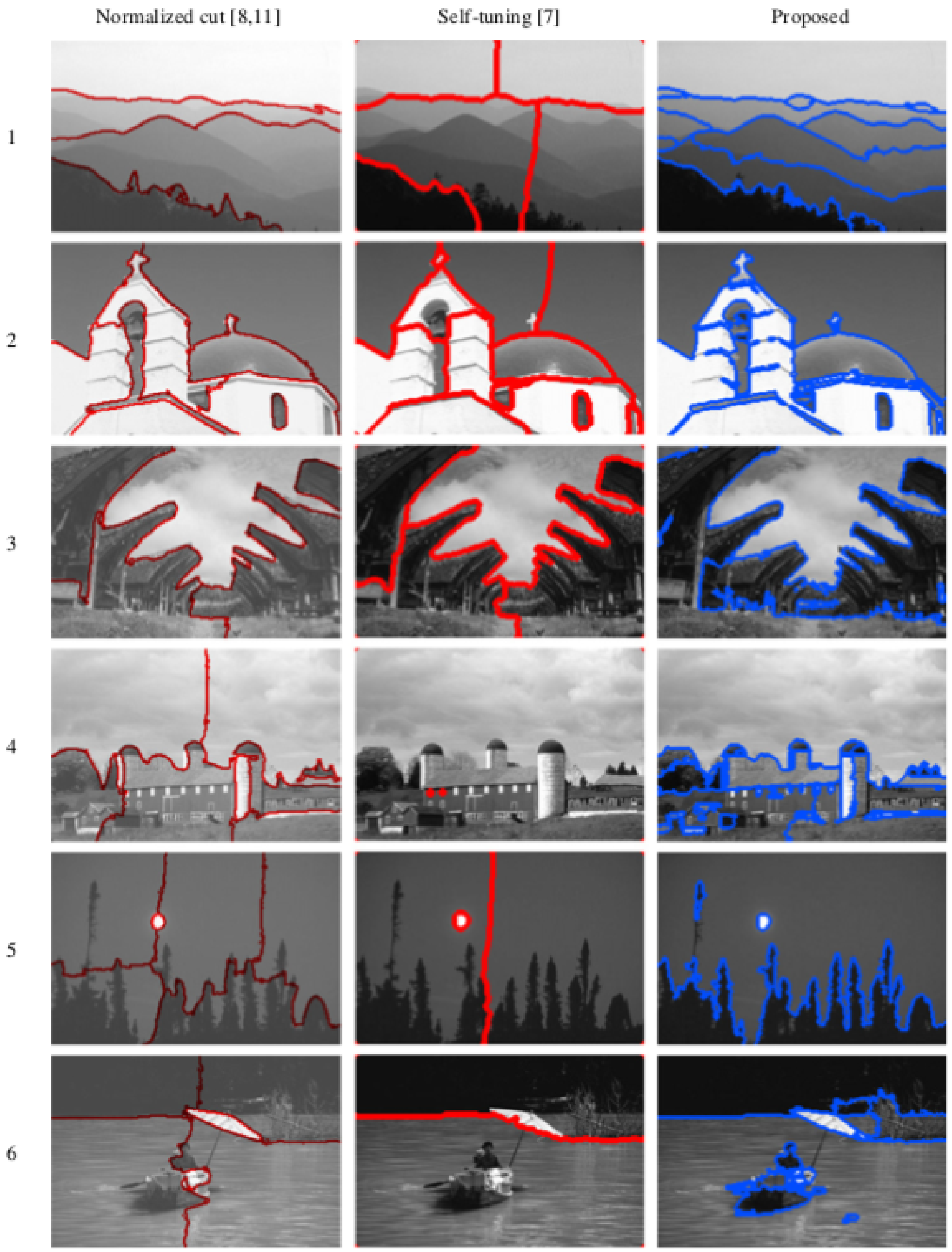


Figure 20. Low contrast image segmentation with spectral clustering [47].

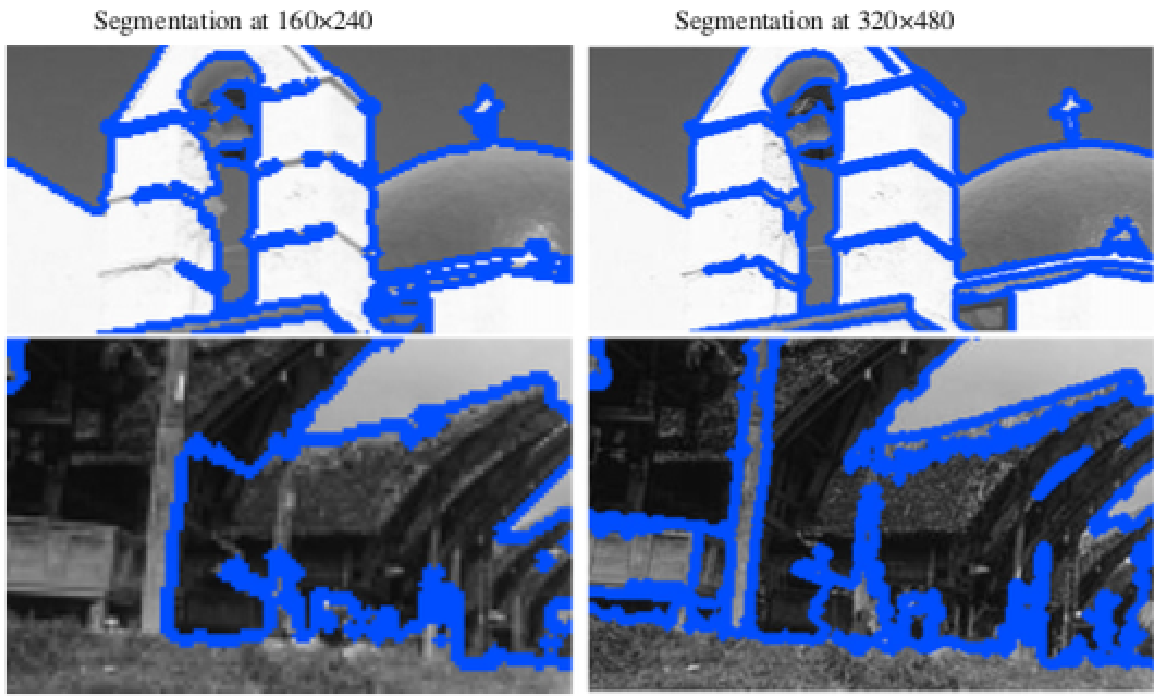
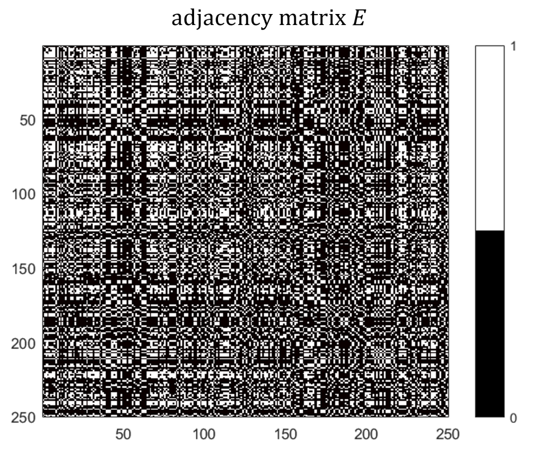
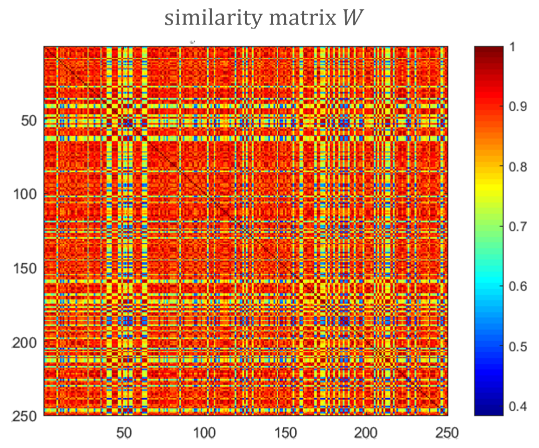
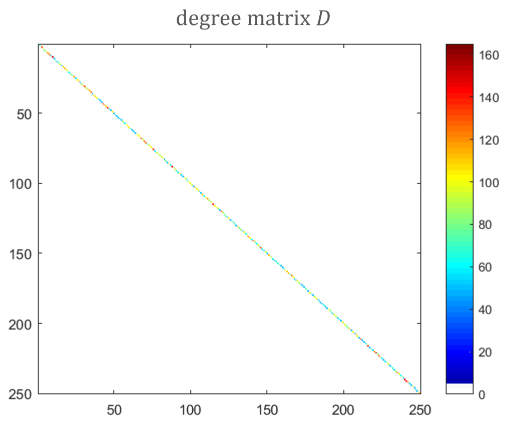


Figure 21. Spectral clustering image segmentation of images at different resolutions [47].

The corresponding adjacency, similarity, and degree matrices E (mostly sparse), W, D are shown below.





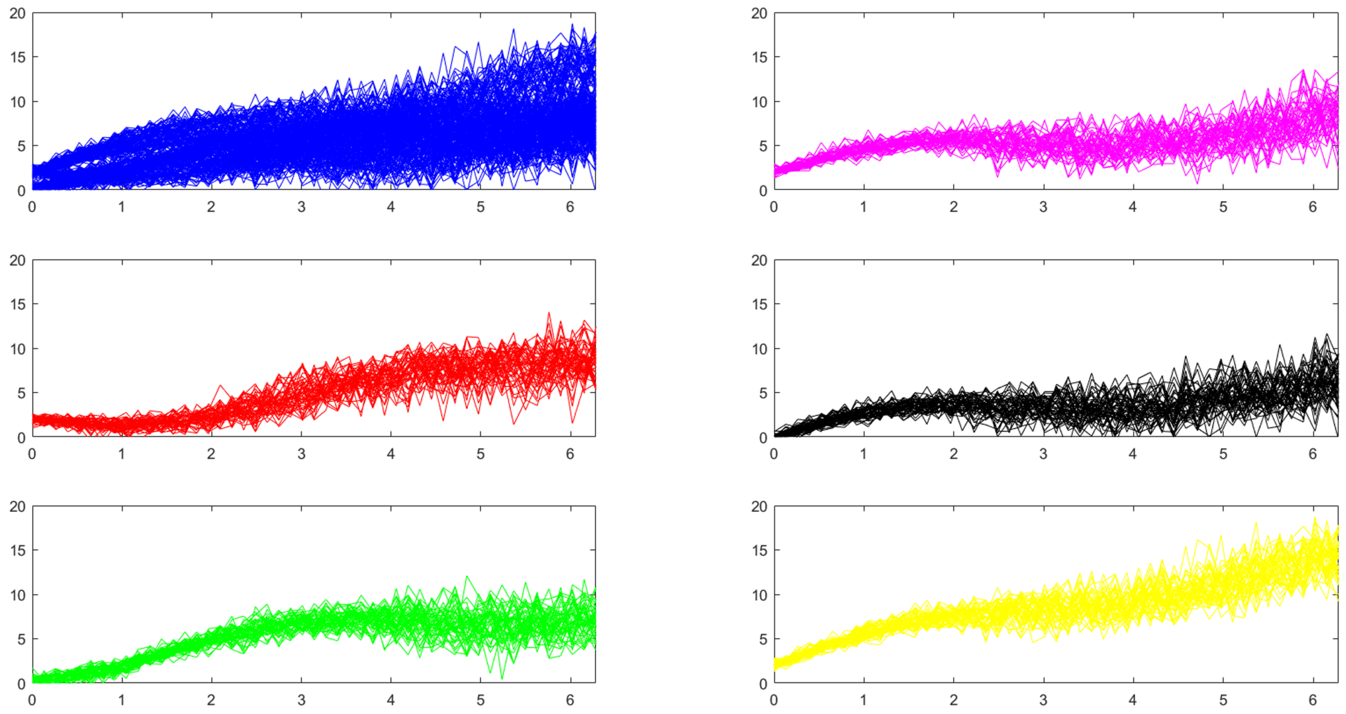


The spectral clustering results are quite appealing, as can be seen in the first realization of the NJW algorithm with $k = 5$ clusters (see top half of Figure 22). Note however that not every run of the algorithm yields an outcome that we would consider meaningful (see bottom half of Figure 22).

In the last example, we once again revisit the 2011 Gapminder dataset. We use the `kernlab` implementation of the NJW algorithm found in `specc()`, with the default settings. The results for one run with each of $k = 2$ to $k = 7$ clusters are shown in Figures 23 and 24.

We can take stock of our attempts to cluster the Gapminder data: in none of the k−means, hierarchical, DBSCAN, and spectral runs have we found what one might call **natural groups**. Perhaps we have not hit on the right method yet, but it could also mean that the task is futile and no such groups exist in the first place.

First realization



Second realization

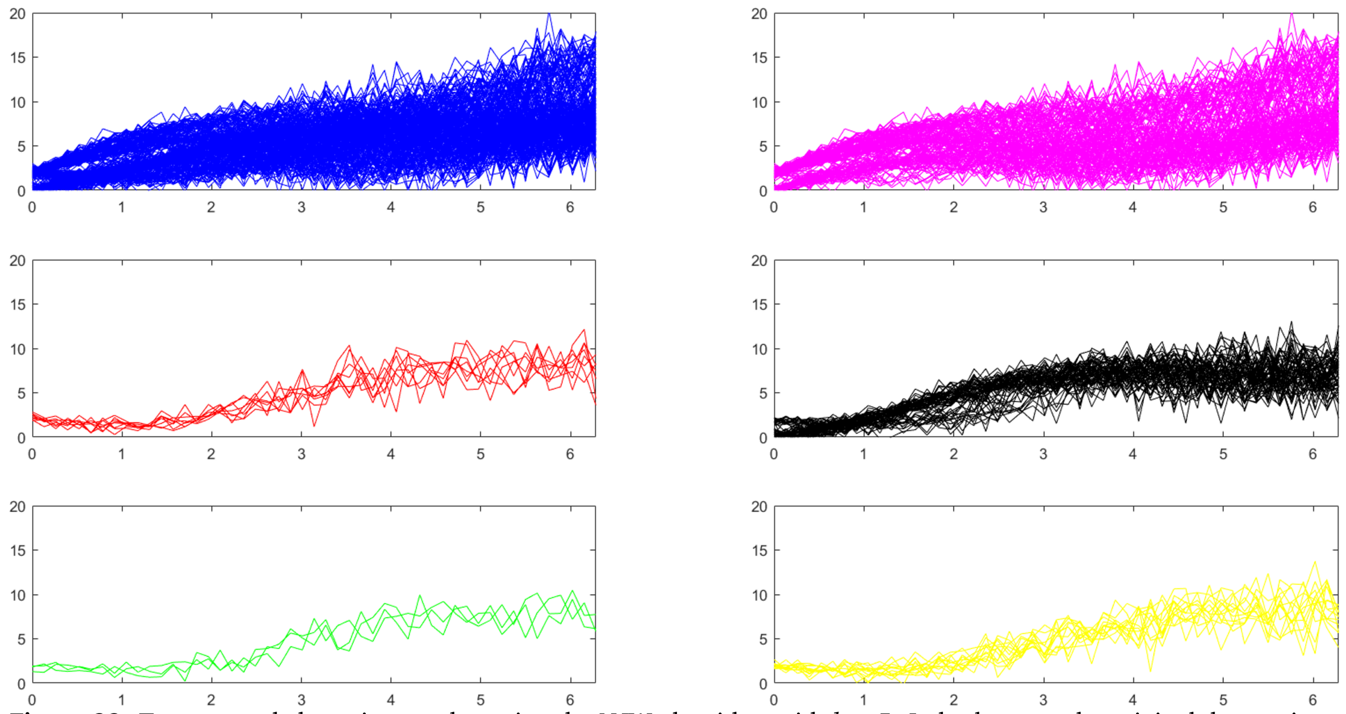


Figure 22. Two spectral clustering results, using the NJW algorithm with $k = 5$. In both cases, the original dataset is shown in blue. In the first case, we see that the NJW algorithm has captured 5 clusters with different times series characteristics, which is an encouraging result. The results are not-deterministic, however: the k−means portion of the algorithm can lead to different clusters, not all of which are of equal quality (as can be seen in the second case).

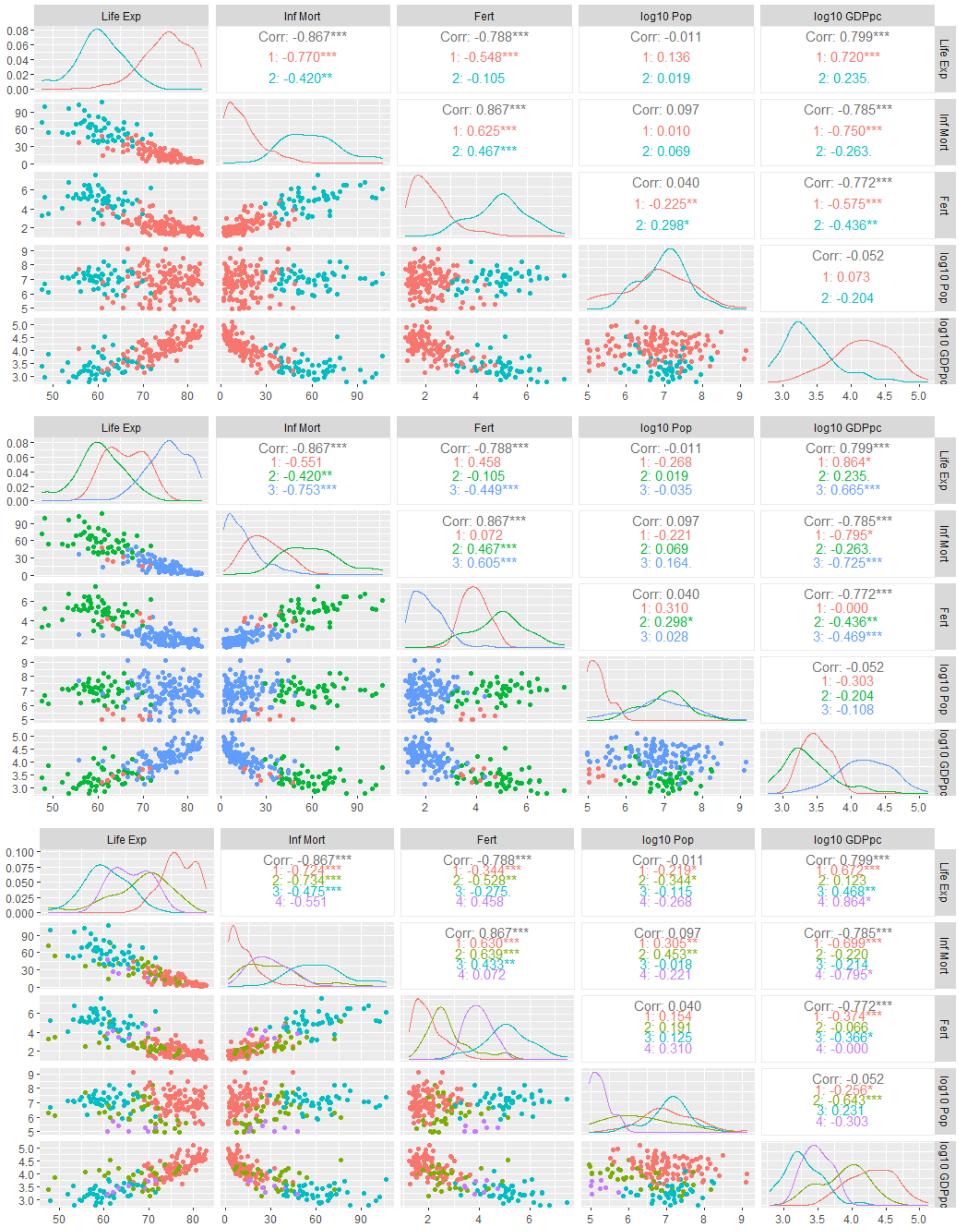


Figure 23. Realizations of the NJW algorithm on the Gapminder data using the default `specc()` settings, for $k = 2, 3, 4$.

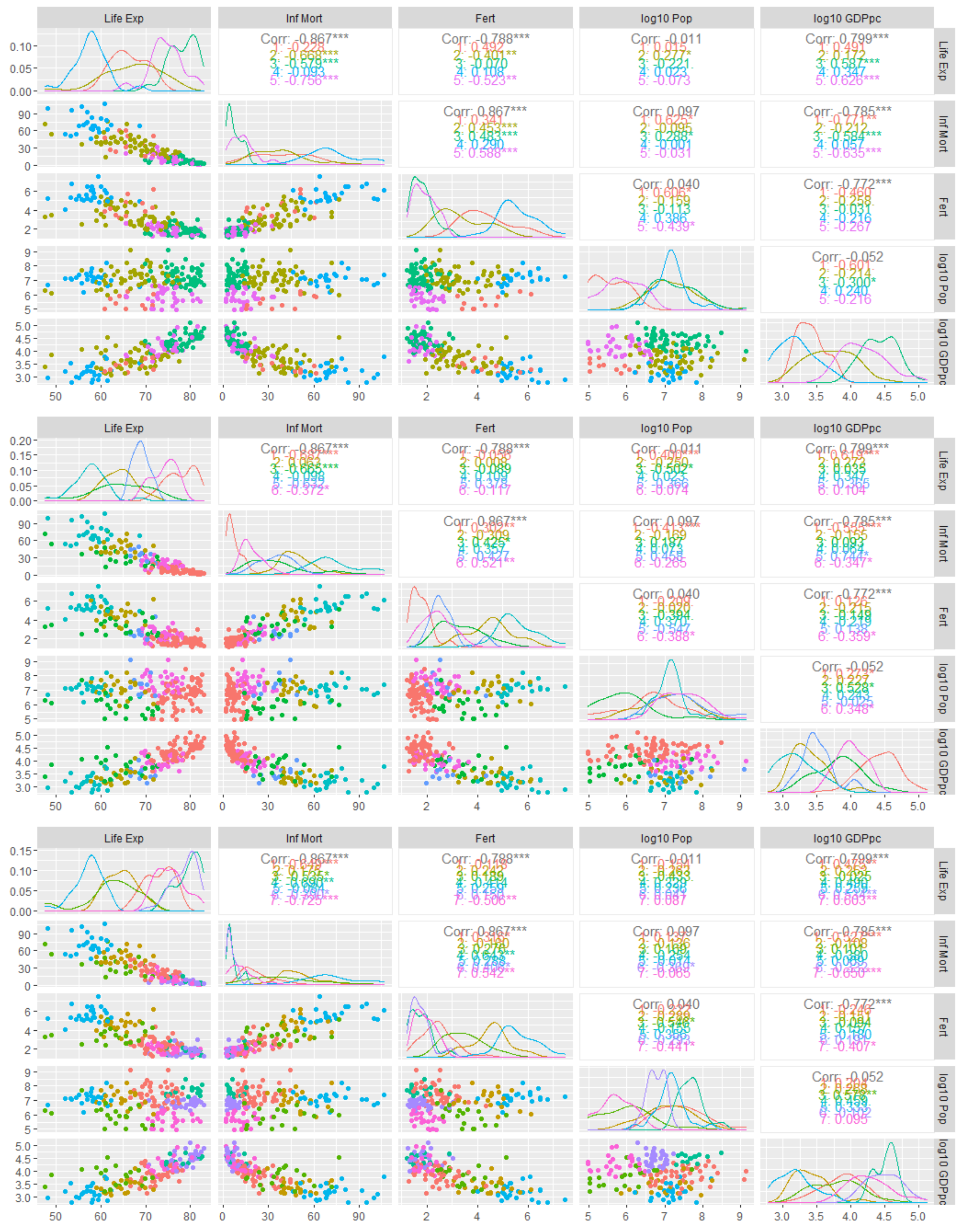


Figure 24. Realizations of the NJW algorithm on the Gapminder data using the default `specc()` settings, for $k = 5, 6, 7$.

4.3 Probability-Based Clustering

In contrast with the model-free approach density-based clustering and spectral clustering, **probabilistic-based clustering** attempt to optimize the fit between the observed data and some **mathematical model** of clustering, with the assumption that the data is generated *via* a number of underlying probability distributions.

In practice, we assume that clusters are represented by **parametric probability distributions**, and the objective is to **learn the parameters** for each of these distributions.

This assumption allows us to use probability theory to derive learning formulas for the parameters.[38] Expectation-Maximization clustering, the representative technique of probability-based clustering, will be revisited and given a lighter (and more visual) treatment in a subsequent section.

Mixture Models The main underlying assumptions of **mixture models** is that each observation is drawn (or generated) from one of several mechanisms (or components). In **model-based clustering**, we learn the parameters that provide the optimal fit to the data; in other words, we make a series of predictions about which components generated each of the observations.

This naturally leads to **clusters**, all observations generated by a given component belonging to the same cluster. Formally, we let

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} \in M_{n,p}(\mathbb{R}).$$

Assume that there are k mechanisms that generate data, and that each of them is determined by a vector of parameters $\boldsymbol{\theta}_\ell$, $1 \le \ell \le k$.

For $1 \le j \le n$, denote the probability of $\mathbf{x}_j$ being **generated by the ℓ–th mechanism**, $1 \le \ell \le k$, by

$$P(\mathbf{x}_j \mid \boldsymbol{\theta}_\ell).$$

The mixture vector $\pi = (\pi_1, \ldots, \pi_k)$ is a vector such that $\pi_\ell \in [0,1]$ for all $1 \le \ell \le k$ and $\pi_1 + \cdots + \pi_k = 1$.

If $z_j \sim \mathrm{C}(k;\pi)$, i.e., if $P(z_j = \ell) = \pi_\ell$, for $1 \le \ell \le k$, and if

$$P(\mathbf{x}_j \mid z_j = \ell) = P(\mathbf{x}_j \mid \boldsymbol{\theta}_\ell) \quad \forall j, \ell,$$

then the probability of observing $\mathbf{x}_j$ is

$$P(\mathbf{x}_j) = \sum_{\ell=1}^{k} \pi_\ell P(\mathbf{x}_j \mid \boldsymbol{\theta}_\ell) = \sum_{\ell=1}^{k} P(z_j = \ell) P(\mathbf{x}_j \mid z_j = \ell),$$

according to the **Law of Total Probability**.

In this set-up, we interpret z_j as the **cluster label** for $\mathbf{x}_j$.

[38]We borrow rather heavily from Deng and Han's *Probabilistic Models for Clustering* chapter in [2].

Alternatively, we could use

$$\mathbf{z}_j \in \{0,1\}^k, \quad \|\mathbf{z}_j\|_2 = 1$$

to denote the **cluster signature** of $\mathbf{x}_j$. The norm condition implies that exactly one of the components of $\mathbf{z}_j$ is 1; all others are 0. For instance, if there are $k = 5$ mechanisms (clusters) in the data and $\mathbf{x}_j \in C_4$, then $\mathbf{z}_j = (0,0,0,1,0)$.[39]

If we write

$$P(\mathbf{z}_j) = \pi_1^{z_{j,1}} \times \cdots \times \pi_k^{z_{j,k}} = \prod_{\ell=1}^{k} \pi_\ell^{z_{j,\ell}}$$

and

$$P(\mathbf{x}_j \mid \mathbf{z}_j) = P(\mathbf{x}_j \mid \boldsymbol{\theta}_1)^{z_{j,1}} \times \cdots \times P(\mathbf{x}_j \mid \boldsymbol{\theta}_k)^{z_{j,k}} = \prod_{\ell=1}^{k} P(\mathbf{x}_j \mid \boldsymbol{\theta}_\ell)^{z_{j,\ell}},$$

we recover the **mixture model**

$$P(\mathbf{x}_j) = \sum_{\ell=1}^{k} \pi_\ell P(\mathbf{x}_j \mid \boldsymbol{\theta}_\ell) = \sum_{\ell=1}^{k} P(\mathbf{z}_j \in C_\ell) P(\mathbf{x}_j \mid \mathbf{z}_j \in C_\ell).$$

Generative Process In practice, then, we can imagine that the dataset $\mathbf{X}$ is generated as follows. For $1 \le j \le n$:

1. draw a cluster signature $\mathbf{z}_j \sim \mathcal{G}_k(\pi) = \mathrm{Mult}_k(\pi)$, and
2. draw an observation $\mathbf{x}_j$ from the corresponding mechanism according to $P(\mathbf{x}_j \mid \mathbf{z}_j)$.

But we usually do not have access to this **generative process**; instead, we are given $\mathbf{X}$ and the clustering task is to determine how likely it is that component C_ℓ, $1 \le \ell \le k$, is **responsible** for observation $\mathbf{x}_j$, $1 \le j \le n$.

To do so, we need to compute the

$$\gamma(z_{j,\ell}) = P(\mathbf{z}_j \in C_\ell \mid \mathbf{x}_j), \quad \forall j, \ell.$$

This is difficult to do directly; we use **Bayes' Theorem** to provide an easier handle on the computations:

$$\begin{aligned}\gamma(z_{j,\ell}) = P(\mathbf{z}_j \in C_\ell \mid \mathbf{x}_j) &= \frac{P(\mathbf{z}_j \in C_\ell) P(\mathbf{x}_j \mid \mathbf{z}_j \in C_\ell)}{P(\mathbf{x}_j)} \\ &= \frac{P(\mathbf{z}_j \in C_\ell) P(\mathbf{x}_j \mid \mathbf{z}_j \in C_\ell)}{\sum_{\nu=1}^{k} P(\mathbf{z}_j \in C_\nu) P(\mathbf{x}_j \mid \mathbf{z}_j \in C_\nu)} \\ &= \frac{\pi_\ell P(\mathbf{x}_j \mid \boldsymbol{\theta}_\ell)}{\sum_{\nu=1}^{k} \pi_\nu P(\mathbf{x}_j \mid \boldsymbol{\theta}_\nu)}.\end{aligned}$$

The **clustering objective** is to infer $\{\pi_\ell\}_{\ell=1}^{k}$, $\{\boldsymbol{\theta}_\ell\}_{\ell=1}^{k}$ from $\mathbf{X}$ for a fixed k, to obtain the desired probabilities $\gamma(z_{j,\ell})$.

[39]This notation can be generalized to **fuzzy clusters** (see Section 4.5): the cluster signature of $\mathbf{x}_j$ is

$$\mathbf{z}_j \in [0,1]^k, \quad \|\mathbf{z}_j\|_2 = 1;$$

if $\mathbf{z}_j = (0, 0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0)$, say, then we would interpret $\mathbf{x}_j$ as belonging equally to clusters C_3 and C_4, or as having probability $1/2$ of belonging to either C_3 or C_4.

Denote

$$\mathbf{\Theta} = \{\pi_1, \ldots, \pi_k, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_\ell\}.$$

If we further assume that the $\mathbf{x}_j$ are **independently** drawn by the generative process, then

$$P(\mathbf{X} \mid \mathbf{\Theta}) = \prod_{j=1}^{n} \sum_{\ell=1}^{k} \pi_k P(\mathbf{x}_j \mid \boldsymbol{\theta}_\ell),$$

or

$$\text{LL}(\mathbf{\Theta}) = \ln P(\mathbf{X} \mid \mathbf{\Theta}) = \sum_{j=1}^{n} \ln\left(\sum_{\ell=1}^{k} \pi_k P(\mathbf{x}_j \mid \boldsymbol{\theta}_\ell)\right),$$

by construction.

The **maximum likelihood estimator** (MLE) of $\mathbf{\Theta}$ is

$$\mathbf{\Theta}_{\text{MLE}} = \arg\max_{\mathbf{\Theta}} \left\{\ln P(\mathbf{X} \mid \mathbf{\Theta})\right\};$$

if we have information about the **prior** $P(\mathbf{\Theta})$, then we may use the **maximum a posteriori estimator** (MAP) instead:

$$\mathbf{\Theta}_{\text{MAP}} = \arg\max_{\mathbf{\Theta}} \left\{\ln P(\mathbf{X} \mid \mathbf{\Theta}) + \ln P(\mathbf{\Theta})\right\}.$$

Whether we use MLE or MAP depend, in large part, on the form taken by the component distributions.

Gaussian Mixture Models A standard assumption is that all clusters are generated by Gaussian mechanisms, which is to say that $P(\mathbf{x}_j \mid \boldsymbol{\theta}_\ell)$ arises from a **multivariate Gaussian** distribution:

$$\begin{aligned}&\mathcal{N}(\mathbf{x}_j \mid \boldsymbol{\mu}_\ell, \mathbf{\Sigma}_\ell)\\ &\quad = \frac{1}{\sqrt{(2\pi)^p |\mathbf{\Sigma}_k|}} \exp\left(-\tfrac{1}{2}(\mathbf{x}_j - \boldsymbol{\mu}_\ell)^\top \mathbf{\Sigma}_\ell^{-1}(\mathbf{x}_j - \boldsymbol{\mu}_\ell)\right),\end{aligned}$$

where $\boldsymbol{\mu}_\ell \in \mathbb{R}^p$ and $\mathbf{\Sigma}_\ell$ is a symmetric positive semi-definite matrix. Thus, if there are k components, then

$$P(\mathbf{x}_j \mid \mathbf{\Theta}) = \sum_{\ell=1}^{k} \pi_\ell \mathcal{N}(\mathbf{x}_j \mid \boldsymbol{\mu}_\ell, \mathbf{\Sigma}_\ell)$$

and

$$\text{LL}(\mathbf{\Theta}) = \ln P(\mathbf{X} \mid \mathbf{\Theta}) = \sum_{j=1}^{n} \ln\left(\sum_{\ell=1}^{k} \pi_\ell \mathcal{N}(\mathbf{x}_j \mid \boldsymbol{\mu}_\ell, \mathbf{\Sigma}_\ell)\right).$$

It is straightforward to show that

$$\nabla\text{LL}(\boldsymbol{\mu}_\ell) = \mathbf{\Sigma}_\ell^{-1} \sum_{j=1}^{n} \gamma(z_{j,\ell})(\mathbf{x}_j - \boldsymbol{\mu}_\ell),$$

so that the MLE estimators for the mean vectors are

$$\hat{\boldsymbol{\mu}}_\ell = \frac{\sum_{j=1}^{n} \gamma(z_{j,\ell})\, \mathbf{x}_j}{\sum_{j=1}^{n} \gamma(z_{j,\ell})};$$

that this is a maximizer for $\text{LL}(\mathbf{\Theta})$ is due to positive semi-definiteness of $\mathbf{\Sigma}_\ell$.

Thus $\hat{\boldsymbol{\mu}}_\ell$ is a weighted mean of the observations of $\mathbf{X}$, with weights corresponding to the posterior probability $\gamma(z_{j,\ell})$ that the ℓ—th component was responsible for generating $\mathbf{x}_j$.

Simultaneously, we can show that

$$\nabla\text{LL}(\mathbf{\Sigma}_\ell) = \sum_{j=1}^{n} \frac{\pi_\ell}{P(\mathbf{x}_j \mid \mathbf{\Theta})} \cdot \frac{\partial \mathcal{N}(\mathbf{x}_j \mid \boldsymbol{\mu}_\ell, \mathbf{\Sigma}_\ell)}{\partial \mathbf{\Sigma}_\ell};$$

slightly more complicated manipulations show that the MLE estimators for the covariance matrices are also weighted averages:

$$\hat{\mathbf{\Sigma}}_\ell = \frac{\sum_{j=1}^{n} \gamma(z_{j,\ell})(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_\ell)(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_\ell)^\top}{\sum_{j=1}^{n} \gamma(z_{j,\ell})}.$$

Finally, to obtain the mixture probabilities π_ℓ, we must maximize $\text{LL}(\Theta)$ with respect to $\boldsymbol{\pi}$, subject to $\pi_\ell \in [0,1]$,

$$\pi_1 + \cdots + \pi_k = 1;$$

we can use **Lagrange multipliers** to show that the MLE estimates of the mixture probabilities are also an average:

$$\hat{\pi}_\ell = \frac{1}{n} \sum_{j=1}^{n} \gamma(z_{j,\ell}).$$

So we have nice expressions for the MLE estimates $\hat{\mathbf{\Theta}}$, but there is a problem: we need the clustering probabilities $\gamma(z_{j,\ell})$ in order to provide the MLE estimates, but these depend on the MLE estimates themselves.

Expectation-Maximization Algorithm While there is no **closed-form solution** allowing us to express the cluster signatures directly in terms of the observed data $\mathbf{X}$, there is a simple iterative solution based on the **Expectation-Maximization algorithm for Gaussian Mixture Models**:

Input: $\mathbf{X}$; Output: $\mathbf{\Theta}^*$ which maximizes $\text{LL}(\mathbf{\Theta})$.

1. Initialize $\mathbf{\Theta}^{[0]} = \left\{\boldsymbol{\mu}_\ell^{[0]}, \mathbf{\Sigma}_\ell^{[0]}, \pi_\ell^{[0]}\right\}_{\ell=1}^{k}$ and set $$\text{LL}^{[0]} = \text{LL}(\mathbf{\Theta}^{[0]});$$

For $i = 0$ to max_step, do:

2. **E**(xpectation)**-step**: compute the responsibilities $$\gamma(z_{j,\ell}^{[i]}) = \frac{\pi_\ell^{[i]} \mathcal{N}(\mathbf{x}_j \mid \boldsymbol{\mu}_\ell^{[i]}, \mathbf{\Sigma}_\ell^{[i]})}{\sum_{\nu=1}^{k} \pi_\nu^{[i]} \mathcal{N}(\mathbf{x}_j \mid \boldsymbol{\mu}_\nu^{[i]}, \mathbf{\Sigma}_\nu^{[i]})}, \quad \forall j, \ell;$$

3. **M**(aximization)**-step**: update the parameters $$\boldsymbol{\mu}_\ell^{[i+1]} = \frac{\sum_{j=1}^{n} \gamma(z_{j,\ell}^{[i]})\, \mathbf{x}_j}{\sum_{j=1}^{n} \gamma(z_{j,\ell}^{[i]})}, \quad \forall \ell;$$

$$\boldsymbol{\Sigma}_\ell^{[i+1]} = \frac{\sum_{j=1}^n \gamma(z_{j,\ell}^{[i]})(\mathbf{x}_j-\boldsymbol{\mu}_\ell^{[i]})(\mathbf{x}_j-\boldsymbol{\mu}_\ell^{[i]})^\top}{\sum_{j=1}^n \gamma(z_{j,\ell}^{[i]})}, \quad \forall \ell,$$

and

$$\pi_\ell^{[i+1]} = \frac{1}{n}\sum_{j=1}^n \gamma(z_{j,\ell}^{[i]}), \quad \forall \ell;$$

4. Set $\text{LL}^{[i+1]} = \text{LL}(\boldsymbol{\Theta}^{[i]})$ and check for convergence according to some **convergence criterion**

$$(\|\boldsymbol{\Theta}^{[i]} - \boldsymbol{\Theta}^{[i+1]}\| < \varepsilon, \quad \text{say}):$$

if satisfied, set $\boldsymbol{\Theta}^* = \boldsymbol{\Theta}^{[i+1]}$; otherwise, repeat steps 2 to 4.

There are two main limitations to using EM for GMM:

- EM is **costlier** (has a longer run-time) than k—means, and depending on the initialization, the algorithm may converge to a **local critical point** which is not necessarily the global maximizer;
- as the algorithm iterates, two (or more) GMM clusters can **collapse** into a single GMM cluster.

The EM algorithm can be sped-up by **first running** k**—means** and using the mean vector, covariance matrix, and proportion of observations of observations in the k—means cluster C_ℓ for the initialization of $\boldsymbol{\mu}_\ell^{[0]}$, $\boldsymbol{\Sigma}_\ell^{[0]}$, and π_ℓ for $1 \leq \ell \leq k$.

The collapsing of clusters can be mitigated by monitoring $\|\boldsymbol{\Sigma}_\ell^i\|_2$ and randomly resetting $\boldsymbol{\mu}_\ell^{[i]}$, $\boldsymbol{\Sigma}_\ell^{[i]}$ when some threshold is reached.

Special Cases and Variants In a GMM with k components, if $\boldsymbol{\Sigma}_\ell = \boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n$ for $1 \leq \ell \leq \ell$, then

$$P(\mathbf{x}_j \mid \boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p}\sigma} \cdot \exp\left(-\frac{1}{2\sigma^2}\|(\mathbf{x}-\boldsymbol{\mu}_\ell)\|_2^2\right);$$

the EM algorithm applied to this special case leads to

$$\textbf{E-step:} \quad \gamma(z_{j,\ell}^{[i]}) = \frac{\pi_\ell^{[i]} \exp\left(-\|\mathbf{x}_j - \boldsymbol{\mu}_\ell^{[i]}\|_2^2/2\sigma^2\right)}{\sum_{\nu=1}^k \pi_\nu^{[i]} \exp\left(-\|\mathbf{x}_j - \boldsymbol{\mu}_\nu^{[i]}\|_2^2/2\sigma^2\right)}$$

$$\textbf{M-step:} \quad \boldsymbol{\mu}_\ell^{[i+1]} = \frac{\sum_{j=1}^n \gamma(z_{j,\ell}^{[i]})\,\mathbf{x}_j}{\sum_{j=1}^n \gamma(z_{j,\ell}^{[i]})}$$

$$\pi_\ell^{[i+1]} = \frac{1}{n}\sum_{j=1}^n \gamma(z_{j,\ell}^{[i]}).$$

When $\sigma \to 0$, we can show that

$$\gamma(z_{j,\ell}) \to \begin{cases} 1 & \text{if } \ell = \arg\min_\nu \left\{\|\mathbf{x}_j - \boldsymbol{\mu}_\nu\|_2^2\right\} \\ 0 & \text{otherwise} \end{cases}$$

which is simply the formulation for k—means.

Note that the components do not need to be multivariate Gaussians; there is a **general EM algorithm** that takes advantage of the concavity of the ln function [2].

If the dataset of observations is **binary**, as may occur in image datasets (each pixel taking on the values 0 or 1, depending as to whether the pixel is white or black, say), we can modify GMM so that $P(\mathbf{x}_j \mid \boldsymbol{\mu}_\ell)$ arises from a **multivariate Bernoulli** distribution:

$$\mathscr{B}(\mathbf{x}_j \mid \boldsymbol{\mu}_\ell) = \prod_{\nu=1}^p \mu_{\ell,\nu}^{x_{j,\nu}}(1-\mu_{\ell,\nu})^{1-x_{j,i}},$$

where $\boldsymbol{\mu}_\ell \in [0,1]^p$. Thus, if there are k components, then

$$P(\mathbf{x}_j \mid \boldsymbol{\Theta}) = \sum_{\ell=1}^k \pi_\ell \mathscr{B}(\mathbf{x}_j \mid \boldsymbol{\mu}_\ell)$$

and

$$\text{LL}(\boldsymbol{\Theta}) = \ln P(\mathbf{X} \mid \boldsymbol{\Theta}) = \sum_{j=1}^n \ln\left(\sum_{\ell=1}^k \pi_\ell \prod_{\nu=1}^p \mu_{\ell,\nu}^{x_{j,\nu}}(1-\mu_{\ell,\nu})^{1-x_{j,i}}\right).$$

We can find $\boldsymbol{\Theta}^*$ that maximizes $\text{LL}(\boldsymbol{\Theta})$ by using the EM algorithm for the Bernoulli Mixture Models: the EM algorithm applied to this special case leads to

$$\textbf{E-step:} \quad \gamma(z_{j,\ell}^{[i]}) = \pi_\ell^{[i]} \prod_{\nu=1}^p \left(\mu_{\ell,\nu}^{[i]}\right)^{x_{j,\nu}} (1-\mu_{\ell,\nu}^{[i]})^{1-x_{j,i}}$$

$$\textbf{M-step:} \quad \boldsymbol{\mu}_\ell^{[i+1]} = \frac{\sum_{j=1}^n \gamma(z_{j,\ell}^{[i]})\,\mathbf{x}_j}{\sum_{j=1}^n \gamma(z_{j,\ell}^{[i]})}$$

$$\pi_\ell^{[i+1]} = \frac{1}{n}\sum_{j=1}^n \gamma(z_{j,\ell}^{[i]}),$$

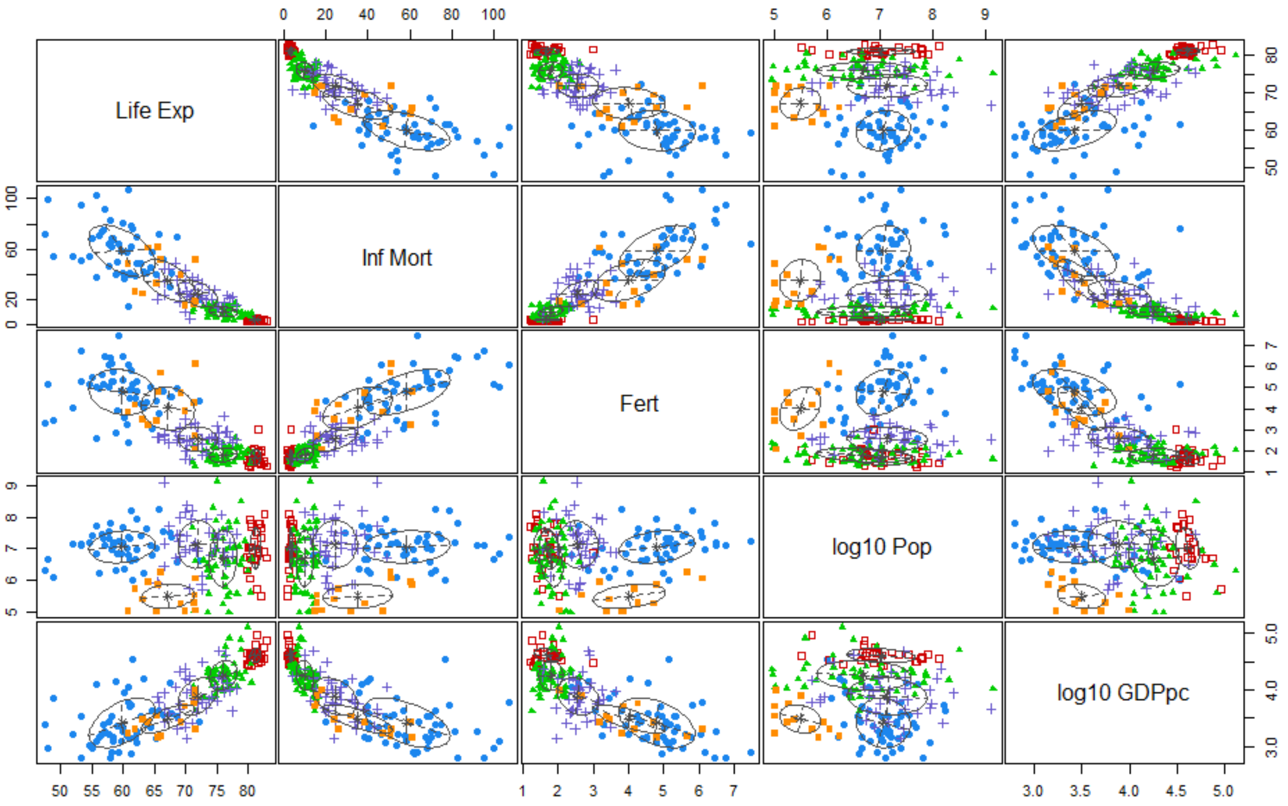
with initialization $\pi_\ell^{[0]} = \frac{1}{k}$ and

$$\boldsymbol{\mu}_\ell \sim \prod_{\nu=1}^p \mathscr{U}(0.25, 0.75)$$

for $1 \leq \ell \leq k$.

Other variants include **Generalized EM**, **Variational EM**, and **Stochastic EM** [2]. Note that the essence of EM methods remains the same for all algorithms: we attempt to "guess" the value of the "hidden" cluster variable $z_{j,\ell}$ in the **E-step**, and we update the model parameters in the **M-step**, based on the approximated responsibilities found in the E—step.

Interestingly, EM can detect overlapping clusters (unlike k—means, see Figure 25). But most variants share the same limitations: convergence to a global maximizer is not guaranteed; it may be quite slow even when it does converge, and the correct number of components is assumed to be known prior to analysis.



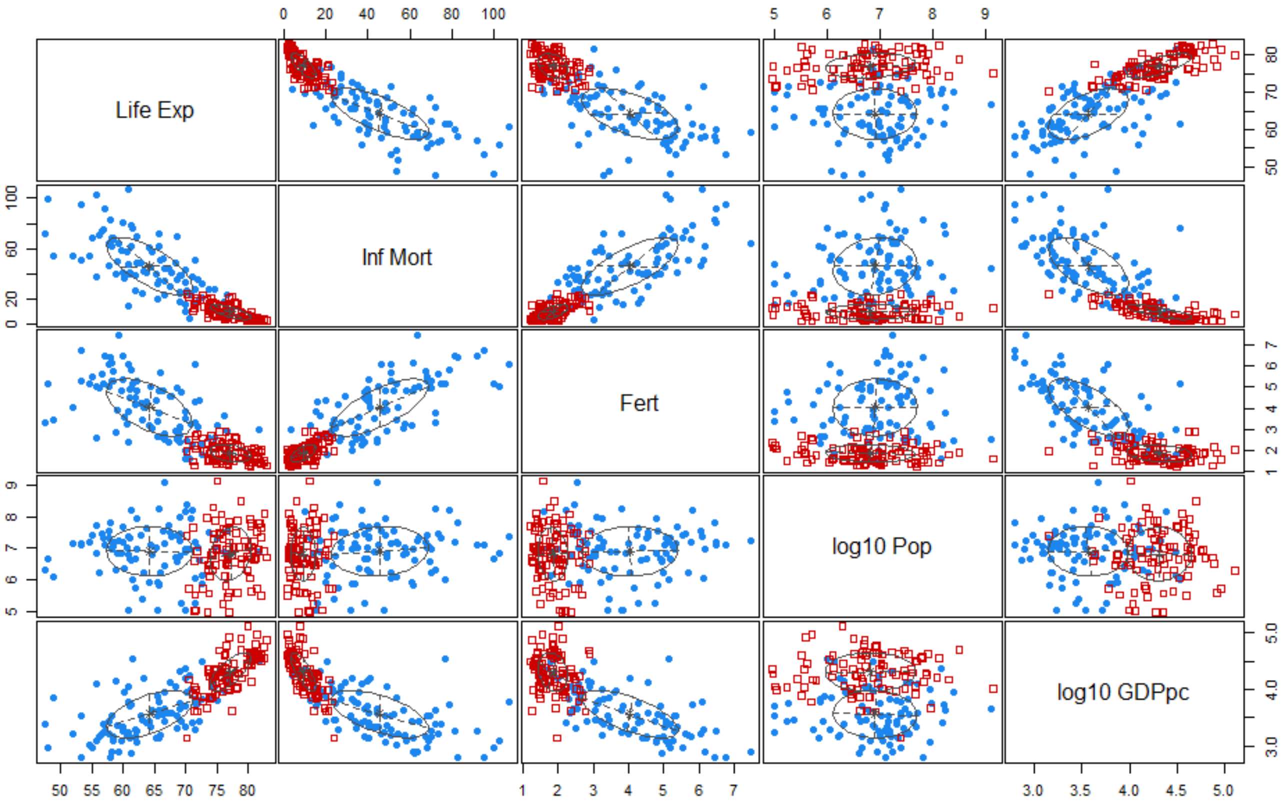


Figure 25. EM results on the Gapminder data using the default `mclust()` settings (no parameters are specified); on the raw data, EM finds 5 clusters (top); on the scaled data, it finds 2 clusters (bottom). This implementation determines the optimal number of clusters using BIC.

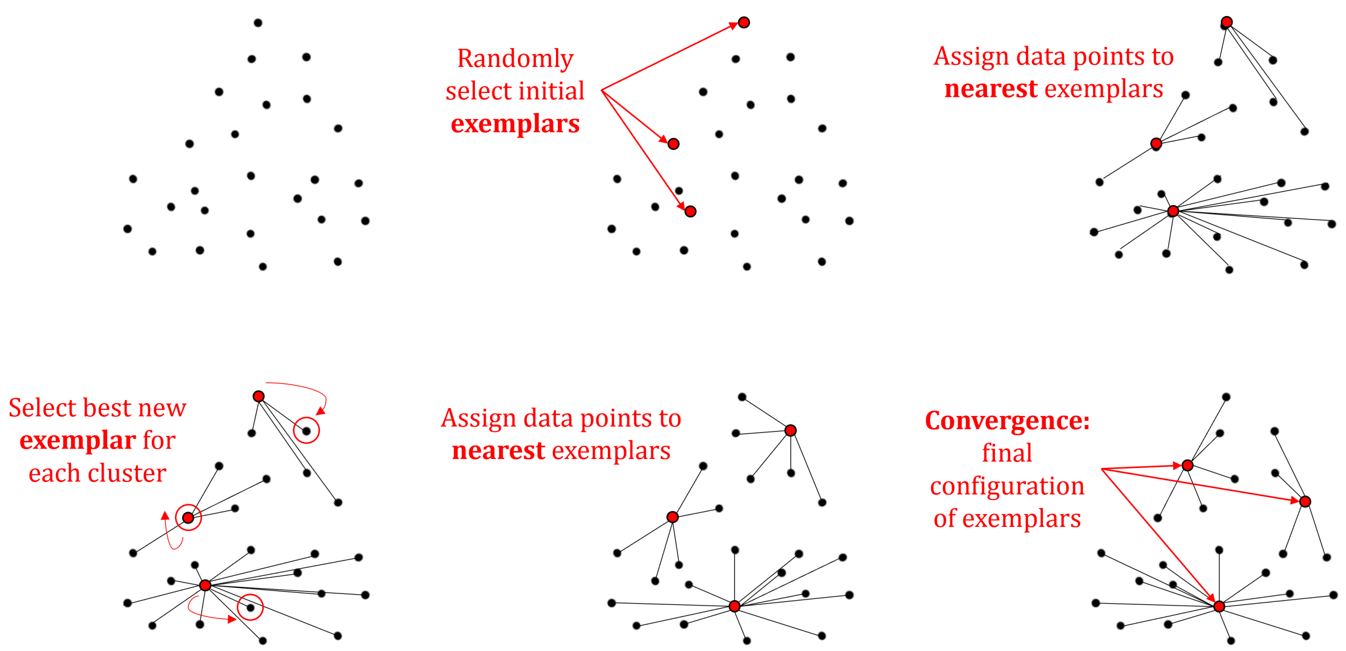

Randomly select initial exemplars
Assign data points to nearest exemplars
Select best new exemplar for each cluster
Assign data points to nearest exemplars
Convergence: final configuration of exemplars

Figure 26. Illustration of 3—mediods on an artificial dataset; modified from [16].

4.4 Affinity Propagation

Affinity propagation (AP) is a fairly recent arrival on the clustering stage [16,17]; it takes a somewhat novel perspective on clustering although, as might be expected, there are still similarities to other clustering methods, in particular, DBSCAN and k—means.

AP takes the k—**medoids** algorithm as a jumping off point. Unlike k—means or EM, this algorithm does not operate on statistical principles; rather, it selects existing observations to act as the **exemplar** for a particular cluster (rather than a mean vector, as in k—means; see Figure 26).

The k—mediods algorithm refines the selection of these exemplars so that in the final (stable) configuration, the observations assigned to an exemplar are quite similar to it, relative to other exemplars.

As the name suggests, the number of clusters k must be selected prior to running the algorithm; as is the case with k—means, k—medoids is non-deterministic and is sensitive to the initial choice of exemplars and similarity metric.

The AP algorithm attempts to overcome the issues arising with k—medoids, using Bayesian network theory (in particular, belief propagation networks and factor graphs), and treats observations as a connected graph. In this approach, each graph vertex can:

- **communicate** with any other vertex, and
- act as a **possible** exemplar for other observations.

The selection of exemplars is determined by exchanging real-valued **messages** between points. Eventually, sets of exemplars and data points associated with each exemplar are generated from this iterative process, forming clusters.

Messages are updated on the basis of fairly simple formulae. As in all clustering contexts, a similarity function s must first be selected prior to clustering: for distinct pairs (i,k), $s(i,k)$ represents the **suitability of k as an exemplar of i**.

Each observation k is further assigned a **preference** $s(k,k)$ that it be chosen as an exemplar. The preference can be constant, to indicate no particular initial preference.

Two types of messages get sent:

- the **availability** $a(i,k)$ sent from k to i, which reports on the suitability of k to be an exemplar of i;
- the **responsiblity** $r(i,k)$ sent from i to k, which reports on the suitability of i to be represented by k.

The availabilities are initialized to 0, the responsibilities to

$$r(i,k) \leftarrow s(i,k) - \max_{k'\neq k}\{a(i,k') + s(i,k')\}.$$

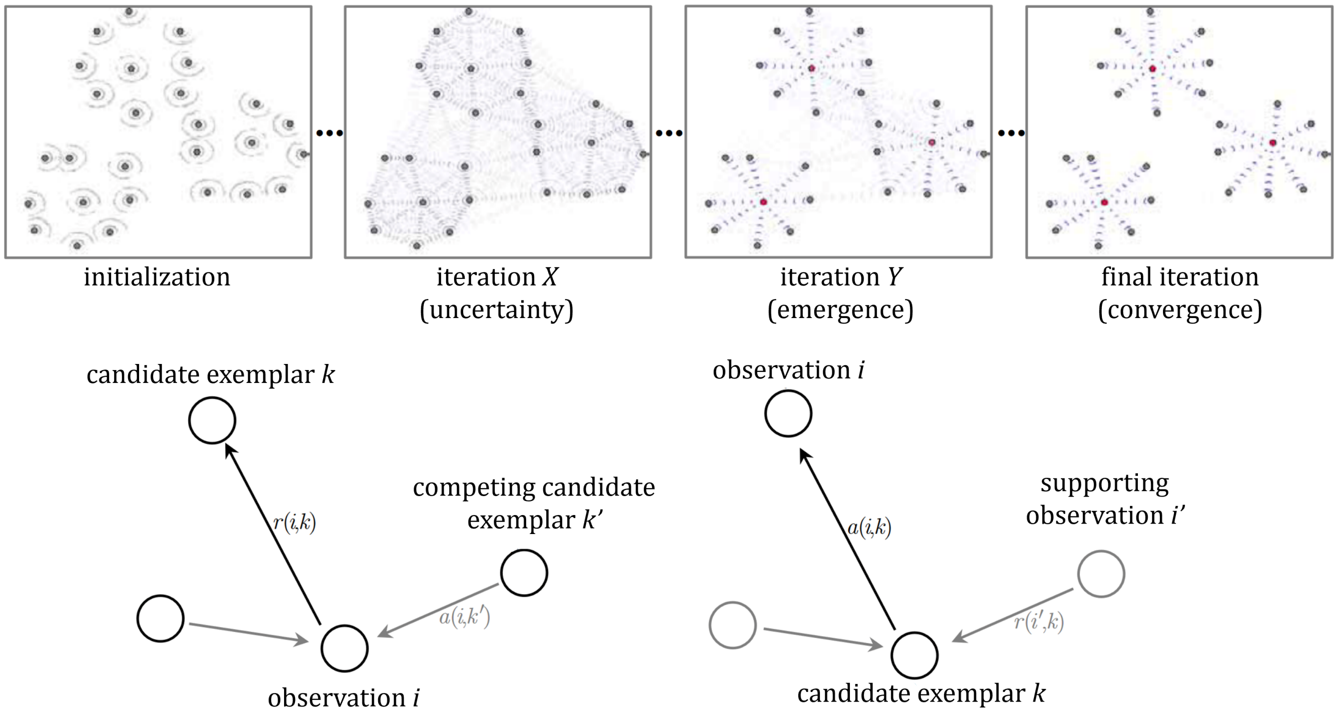
This calculation allows eligible exemplars of an observation to "**compete**" for each observations, in a sense, so they can become that observation's exemplar.[40]

Subsequently, the focus switches back and forth between the exemplar and the observation perspective, with observations looking for available exemplars:

$$a(i,k) \leftarrow \begin{cases} \min\left\{0, r(k,k) + \sum_{i'\notin\{i,k\}} \max\{0, r(i,k)\}\right\} & i \neq k \\ \sum_{i'\neq k} \max\{0, r(i',k)\} & i = k \end{cases}$$

The case $i = k$ is intended to reflect **current evidence** that point k is an exemplar.

[40] As candidate exemplars are themselves observations, we can also compute **self-responsibility** as $r(k,k) \leftarrow s(k,k) - \max_{k\neq k'}\{s(k,k')\}$.



initialization
iteration X (uncertainty)
iteration Y (emergence)
final iteration (convergence)
candidate exemplar k
$r(i,k)$
competing candidate exemplar k'
$a(i,k')$
observation i
observation i
$a(i,k)$
supporting observation i'
$r(i',k)$
candidate exemplar k

Figure 27. Illustration of affinity propagation on an artificial dataset (top); illustration of availability and responsibility (bottom); modified from [16].

The responsibilities and availabilities are updated, reflecting the **current affinity** that one observation has for choosing another observation as its exemplar (hence the name), until the quantities converge to $r(i,k)$ and $a(i,k)$, respectively, for all pairs of observations (i,k).

This leads to the **cluster assignment** $\{c_1,\ldots,c_n\}$, where

$$c_i = \underset{k}{\operatorname{argmax}}\{a(i,k)+r(i,k)\}, \quad 1\leq i\leq n;$$

if i is an observation with associated exemplar k, then $c_i = c_k = k$.

The fact that any observation can become an exemplar when the quantities are updated, and thus that the number of clusters is not an algorithm parameter, is an important distinction between AP and k—medoids (and other segementation clustering approaches).

Setting Algorithm Parameters AP has two parameters which impact its clustering behaviour: the **input preference** (which influences the eventual number of clusters) and the **dampening parameter**.

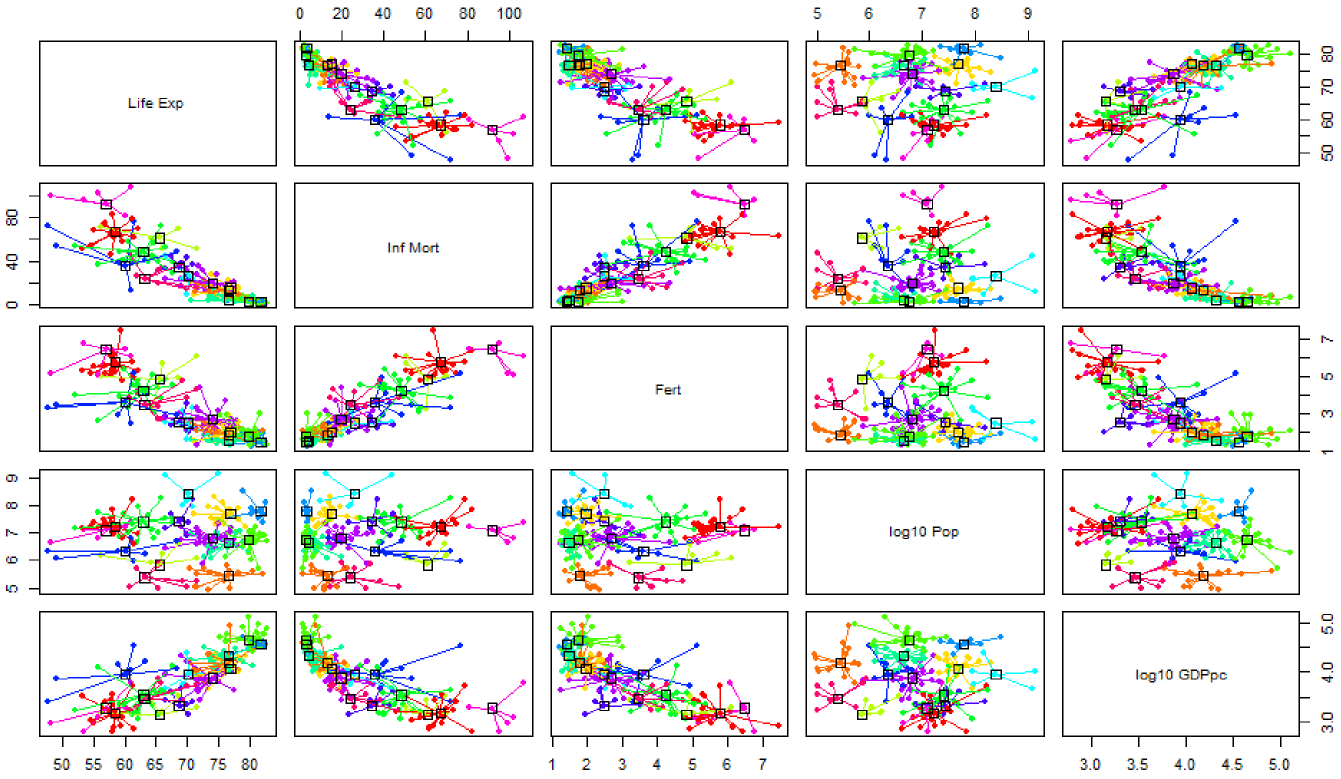
The input preference determines the suitability of each observation to act as an exemplar; this is often set as the **median similarity** in the data, but it can be tweaked. In principle, certain observations could be assigned preference values in a different manner, perhaps relating to domain knowledge (or previous results).

The **dampening parameter** is slightly more technical. Because affinity propagation creates a directed graph to generate clusters, it can become vulnerable to **graph loops**, which could result in algorithmic oscillations (the algorithm may not converge to a particular solution). The dampening factor acts to control this oscillation problem.

Comparison with Other Algorithms Performance of clustering algorithms can be considered both in general (e.g., based on best/worst cases of an implemented algorithm) or in the context of applications in particular domains. One major AP drawback is the cost of calculation of the similarity matrix, which is $O(n^2)$. Once the similarity matrix has been calculated, the number of scalar computations scales linearly in the number of similarities or quadratically in the number of observations if all possible pairwise similarities are used [16]. In other words, AP is slow on larger datasets.

Arguably, one of the major advantages of AP (other than not having to specify the number of clusters up front) is its ability to use any similarity measure. As a result, we do not need to alter the dataset to 'fit' with a distance/similarity framework (e.g., by changing categorical variables into numeric variables in some way, or ignoring categorical variables altogether).

Example We once again re-visit the 2011 (scaled) Gapminder dataset. We use the AP implementation found in the R package `apcluster`, with similarity $s(i,k) = -\|\mathbf{x}_i - \mathbf{x}_k\|^2$. We start by setting the input preference as the median similarity and obtain 13 clusters; if instead we use the minimum similarity, we obtain 4 clusters (exemplars: Guinea, Guyana, Croatia, Morocco). The results are displayed in Figure 28.



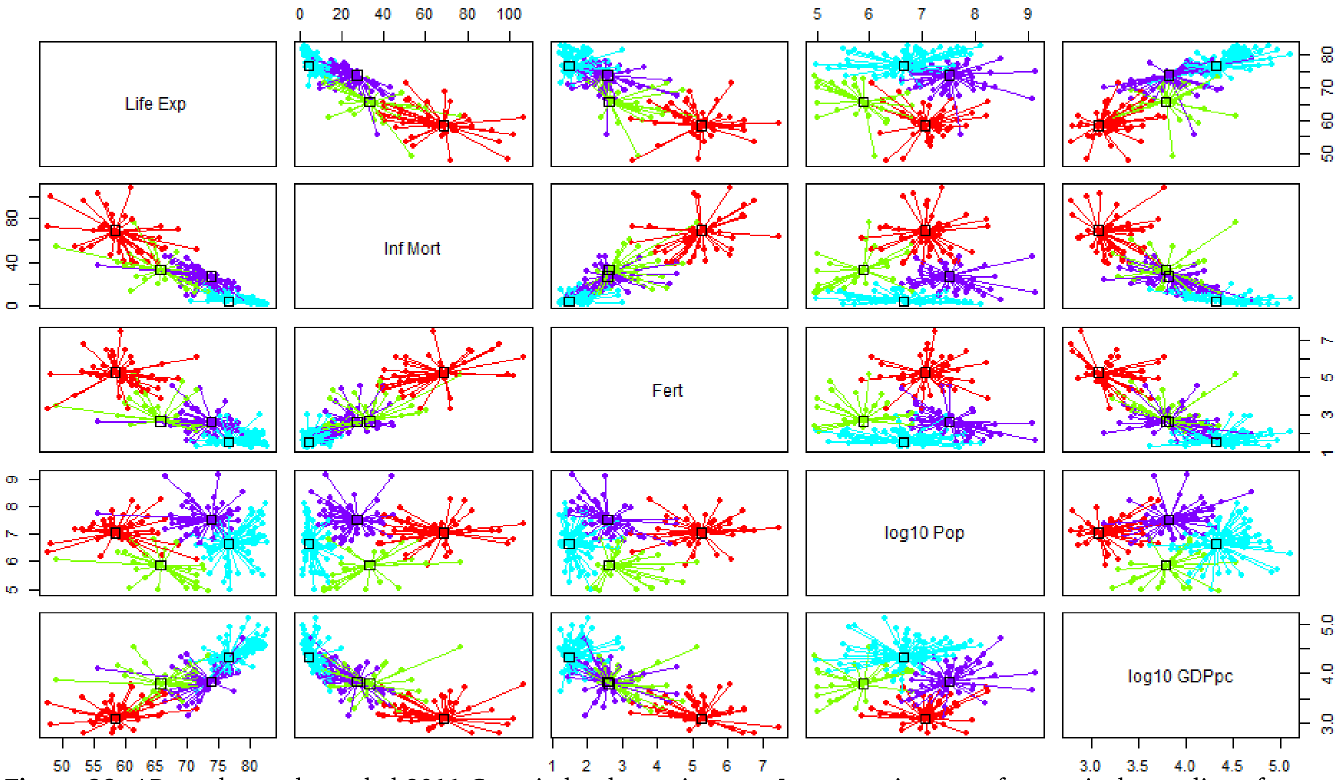


Figure 28. AP results on the scaled 2011 Gapminder data using `apcluster`; input preference is the median of similiarities in the data, yielding 13 clusters (top) or the minimum of input similarities, yielding 4 clusters.

4.5 Fuzzy Clustering

Fuzzy clustering (FC) is also called "soft" clustering (in opposition to "hard" clustering). Rather than assigning each observation to a cluster, they are assigned a **cluster signature**, a set of values that indicate their relative membership in each of the clusters.

The signature vector is often interpreted as a **probability vector**: observation $\mathbf{x}_i$ belongs to cluster ℓ with probability $p_{i,\ell} \geq 0$, with

$$p_{i,1} + \cdots + p_{i,c} = 1, \quad \text{for all } 1 \leq i \leq n.$$

Fuzzy c—Means: The Typical Approach The most prevalent algorithm for carrying out FC is called fuzzy c—means (FCM). It is a variant of k—means with two modifications:

- the presence of a new parameter $m > 1$, called the **fuzzyfier**, which determines the degree of "fuzziness" of the clusters, and
- cluster membership is output as a **weight vector**, with weights in $[0,1]$ adding to 1.

As in k—means, c observations are selected randomly as the initial cluster centroids, as are the membership weights of each observation.

The membership weights of each observations, relative to the current centroid, are re-calculated based on how "close" the point is to the given centroid in comparison to the distance to all of the other centroids.[41]

Effectively, we are looking for clusters that minimize the objective function

$$\sum_{\ell=1}^{c} \sum_{\mathbf{x}_i \in C_\ell} u_{i,\ell}^m \text{variation}(\mathbf{x}_i, \boldsymbol{\mu}_\ell),$$

where the **degree $u_{i,\ell}^m$ to which observation $\mathbf{x}_i$ belongs to cluster C_ℓ** is

$$u_{i,\ell}^m = \frac{1}{\sum_{j=1}^{c} \left(\frac{\text{variation}(\mathbf{x}_i, \boldsymbol{\mu}_\ell)}{\text{variation}(\mathbf{x}_i, \boldsymbol{\mu}_j)} \right)^{2/(m-1)}}.$$

The value of m effectively determines the width of **fuzziness bands** around clusters, where clusters may overlap with other clusters.

Within these bands, if there are overlaps, points will have weights between 0 and 1. Outside of these bands, points will have a membership of 1 for a particular cluster (that it is close to) and a membership of 0 for other bands.

As with k—means, the algorithm is generally run until the change in membership values, or in this case the weights, falls below a particular threshold.

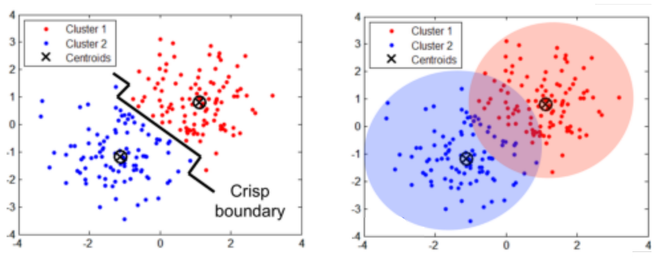
[41]The centroid of the ℓth cluster is the weighted average of ALL observations by the degree to which they belong to cluster ℓ.

In practice, we typically use $m = 2$ and

$$\text{variation}(\mathbf{x}_i, \boldsymbol{\mu}_\ell) = \|\mathbf{x}_i - \boldsymbol{\mu}_\ell\|^2.$$

As $m \to 1$, FCM converges to k—means.

Comparison Between Fuzzy c—Means and k—Means To gain an appreciation for how FCM works, it can be useful to compare its results to those provided by k—means. The image below shows the same dataset clustered by k—means (left) and fuzzy c—means (right) [6].



On the right, we can see observations that "belong" to the 2 clusters. FCM is useful in this context because it would seem almost arbitrary for some of the points to be assigned to one or the other cluster (which is what k—means does).

Other Fuzzy Clustering Options Although FCM is the most popular fuzzy clustering algorithm, it is not a particularly **nuanced** algorithm. Like k—means, the resulting clusters are essentially blob-shaped.

Sophisticated results can be gained by using more complex algorithms. The **Gustafson–Kessel** (GK) clustering algorithm [19] is an early extension of FCM which replaces the simple distance measure used in FCM with a (covariance) matrix. This brings FCM more in-line with EM clustering, which also provides fuzzy results, and can be carried out with a variety of statistical models, resulting in a more mature clustering results, albeit at the cost of heavier processing. **FANNY** [27] is another fuzzy approach; it is less sensitive to outliers than FCM is.

Fuzzy Clustering Validation As with hard clustering, it is important to **validate** fuzzy clusters. A number of validation strategies have been developed; the **Xie-Beni index** is a popular choice. It can be calculated for non-fuzzy clusters as well as for fuzzy clusters. However, it takes into accounts the weights of the points for each clustering by weighting the clustering separation and compactness measures using the membership matrix (i.e., the matrix that contains the weights for each observation with respect to each cluster). Other metrics include the **Tang index** and the **Kwon index** [29, 45, 56].

Example We show some results of FANNY (with $c = 2, 3, 4$, and 6 clusters) and FCM (with $c = 4$ clusters) on the (scaled) 2011 Gapminder dataset in Figure 29.

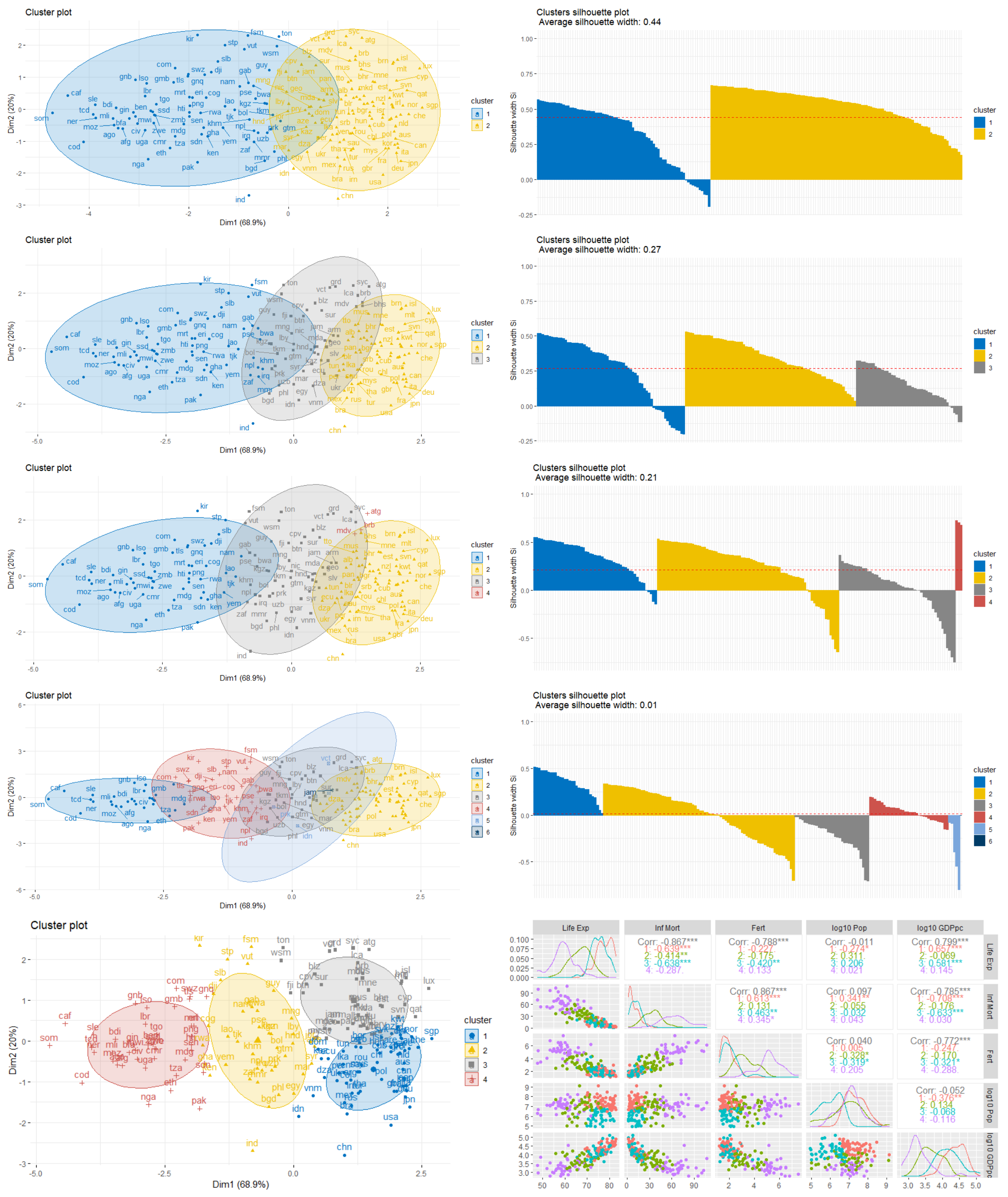


Figure 29. FANNY (for $c = 2, 3, 4, 6$ clusters, first 4 rows) and FCM results (bottom row) on the scaled 2011 Gapminder dataset. The scatterplots in the column on the left are projected on the first two principal components of the data. Note the cluster overlap. The silhouette plots suggest that there are probably 2 or 3 fuzzy clusters in the data.

4.6 Cluster Ensembles

We have seen that the choice of clustering method and algorithm parameters may have an impact on the nature and number of clusters in the data; quite often, the resulting clusters are **volatile**. This is aligned with the idea that the ability to accurately assess the quality of a clustering outcome remains **elusive**, for the most part.

The goal of **ensemble clustering** is to combine the results of multiple clustering runs to create a more **robust** outcome.

Most ensemble models use the following two steps to generate an outcome:

1. **generate** different clustering schemes, using different models, parameters, or data selection mechanisms (the **ensemble components**), and
2. **combine** the different results into a single outcome.

Selecting Different Ensemble Components The ensemble components are either **model-based** or **data selection-based**.

In **model-based ensembles**, the different components of the ensemble reflect different models, such as the use of

- different clustering **approaches**;
- different **parameter settings** for a given approach;
- different **randomizations** (for stochastic algorithms),
- or some **combination** of these.

For instance, an ensemble's components could be built from:

1. 5 runs of k—means for each of $k = 2, \dots, 10$, for each of the Euclidean and Manhattan similarities (90 components);
2. the hierarchical clustering outcome for each of the complete, single, average, centroid, and Ward linkage, for each of the Euclidean and Manhattan distances, for each of $k = 2, \dots, 10$ clusters (90 components);
3. the DBSCAN outcome for each of 5 values of ε^*, for each of $minPts = 2, \dots, 10$, for each of the Euclidean and Manhattan distances (90 components), and
4. the spectral clustering outcome for each of 3 threshold values τ, for each of the 3 types of Laplacians, for $k = 2, 4, 6, 8, 10$, for each of the Euclidean and Manhattan distances (90 components),

for a total of $4 \times 90 = 360$ components. Note that we could also pick algorithms, settings, and similarity measures randomly, from a list of reasonable options.

In **data selection-based ensembles**, we might select a specific clustering approach, combined with a set of parameters, and a given randomization (if the approach is stochastic) and instead build the different components of the model by running the algorithm on different subsets of the data, either *via*:

- selecting **subsets of observations** using random or other probabilistic sampling scheme;
- selecting **subsets of variables**, again using probabilistic sampling, or
- some **combination** of both.

For instance, an ensemble's components could be built using affinity propagation with Euclidean distance and a specific combination of input preference and dampening parameter, and 360 subsets of the data, obtained as follows:

1. for each component, draw a % of observations to sample and a # of variables to select from the data;
2. randomly select a subset with these properties;
3. run affinity propagation on the subset to obtain a clustering outcome.

We could also combine model-based and data selection-based approaches to create the components.

Combining Different Ensemble Components However the components are obtained, we need to find a way to combine them to obtain a **robust clustering consensus**. There are three basic methods to do this:

- **general affiliation**;
- **hypergraph partitioning**, and
- **meta-clustering**.

In the **general affiliation** approach, we consider each pair of observations and determine how frequently they are found in the same clusters in each of the ensemble components. The corresponding proportions create a **similarity matrix**, which can then be used to cluster the data using some graph-based method, such as DBSCAN.

In the **hypergraph partitioning** approach, each observation in the data is represented by a **hypergraph vertex**. A cluster in any of the ensemble components is represented as a **hypergraph hyperedge**, a generalization of the notion of edge which connects (potentially) more than two vertices in the form of a **complete clique**. This hypergraph is then partitioned using graph clustering methods.[42]

The **meta-clustering** approach is also a graph-based approach, except that vertices are associated with each cluster in the ensemble components; each vertex therefore represents a set of **data objects**. A graph partitioning algorithm is then applied to this graph.[43] **Balancing constraints** may be added to the meta-clustering phase to ensure that the resulting clusters are balanced.

Cluster ensembles are implemented in `R` *via* the packages `diceR` and `clue`. More information is available in [1, 2, 49].

[42]One major challenge with hypergraph partitioning is that a hyperedge can be "broken" by a partitioning in many different ways, not all of which are qualitatively equivalent. Most hypergraph partitioning algorithms use a constant penalty for breaking a hyperedge.

[43]The distribution of the membership of different instances to the meta-partitions can be used to determine its meta-cluster membership, or soft assignment probability.

References

[1] C. C. Aggarwal. *Data Mining: The Textbook*. Springer, Cham, 2015.

[2] C. C. Aggarwal and C. K. Reddy, editors. *Data Clustering: Algorithms and Applications*. CRC Press, 2014.

[3] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr.*, 12(5):613, 2009.

[4] R. C. Amorim. Feature relevance in ward's hierarchical clustering using the lp norm. *J. Classif.*, 32(1):46–62, apr 2015.

[5] F. R. Bach and M. I. Jordan. Learning spectral clustering, with application to speech separation. *J. Mach. Learn. Res.*, 7:1963–2001, Dec. 2006.

[6] A. Bagnaro, F. Baltar, and G. Brownstein. Reducing the arbitrary: fuzzy detection of microbial ecotones and ecosystems – focus on the pelagic environment. *Environmental Microbiome*, 15(16), 2020.

[7] Y. Bengio, P. Vincent, and J.-F. Paiement. Learning eigenfunctions of similarity: Linking spectral clustering and kernel PCA. Technical Report 1232, Département d'informatique et recherche opérationnelle, Université de Montréal, 2003.

[8] P. Boily. Non-Technical Aspects of Consulting. *Introduction to Quantitative Consulting*, 2021.

[9] P. Boily. Statistical Methods for Supervised Learning. *Data Science Report Series*, 2021.

[10] P. Boily and J. Schellinck. Machine Learning 101. *Data Science Report Series*, 2021.

[11] P. Boily and J. Schellinck. The Fundamentals of Data Insight. *Data Science Report Series*, 2021.

[12] Y. Cissokho, S. Fadel, R. Millson, R. Pourhasan, and P. Boily. Anomaly Detection and Outlier Analysis. *Data Science Report Series*, 2020.

[13] J. Cranshaw, R. Schwartz, J. I. Hong, and N. M. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In J. G. Breslin, N. B. Ellison, J. G. Shanahan, and Z. Tufekci, editors, *ICWSM*. The AAAI Press, 2012.

[14] B. Desgraupes. *clusterCrit: Clustering Indices*, 2018. R package version 1.2.8.

[15] J. d'Huy. Scientists trace society's myths to primordial origins. *Scientific American (Online)*, Sep 2016.

[16] D. Dueck. Affinity propagation: Clustering data by passing messages. *PhD thesis*, 01 2009.

[17] B. Frey and D. Dueck. Clustering by passing messages between data points. *Science (New York, N.Y.)*, 315:972–6, 03 2007.

[18] M. Grootendorst. 9 Distance Measures in Data Science. *Towards Data Science*, Feb 2021.

[19] D. Gustafson and W. C. Kessel. Fuzzy clustering with a fuzzy covariance matrix. *1978 IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes*, pages 761–766, 1978.

[20] U. Habib, K. Hayat, and G. Zucker. Complex building's energy system operation patterns analysis using bag of words representation with hierarchical clustering. *Complex Adapt. Syst. Model.*, 4:8, 2016.

[21] M. Harmouch. 17 types of similarity and dissimilarity measures used in data science. *Towards Data Science*, Mar 2021.

[22] N. Harris. Visualizing DBSCAN Clustering.

[23] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed.* Springer, 2008.

[24] S. Hwang and J.-C. Thill. Delineating urban housing submarkets with fuzzy clustering. *Environment and Planning B: Planning and Design*, 36:865–882, 09 2009.

[25] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer, 2014.

[26] A. Jawad, K. Kersting, and N. Andrienko. Where traffic meets dna: Mobility mining using biological sequence analysis revisited. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, page 357–360, New York, NY, USA, 2011. Association for Computing Machinery.

[27] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley, 1990.

[28] H. T. Kung and D. Vlah. A spectral clustering approach to validating sensors via their peers in distributed sensor networks. *Int. J. Sen. Netw.*, 8(3/4):202–208, Oct. 2010.

[29] S. H. Kwon, J. Kim, and S. H. Son. Improved cluster validity index for fuzzy clustering. *Electronics Letters*, 57(21):792–794, 2021.

[30] T. Lange, M. Braun, V. Roth, and J. Buhmann. Stability-based model selection. *Advances in Neural Information Processing Systems (NIPS 2002): 2002*, 06 2003.

[31] O. Leduc, A. Macfie, A. Maheshwari, M. Pelletier, and P. Boily. Feature Selection and Dimension Reduction. *Data Science Report Series*, 2020.

[32] J. M. Lewis, M. Ackerman, and V. R. de Sa. Human cluster evaluation and formal quality measures: A comparative study. *Cognitive Science*, 34, 2012.

[33] Y. Murat and Z. Cakici. An integration of different computing approaches in traffic safety analysis. *Transportation Research Procedia*, 22:265–274, 12 2017.

[34] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, page 849–856, Cambridge, MA, USA, 2001. MIT Press.

[35] J. Ning, L. Zhang, D. Zhang, and C. Wu. Interactive image segmentation by maximal similarity based region merging. *Pattern Recognition*, 43:445–456, 02 2010.

[36] M. Orlowska, K. Pytlakowska, A. Mrozek-Wilczkiewicz, R. Musiol, M. Waksmundzka-Hajnos, M. Sajewicz, and T. Kowalska. A comparison of antioxidant, antibacterial, and anticancer activity of the selected thyme species by means of hierarchical clustering and principal component analysis. *Acta Chromatographica Acta Chromatographica*, 28(2):207 – 221, 2016.

[37] V. U. Panchami and N. Radhika. A novel approach for predicting the length of hospital stay with dbscan and supervised classification algorithms. In *ICADIWT*, pages 207–212. IEEE, 2014.

[38] V. U. Panchami and N. Radhika. A novel approach for predicting the length of hospital stay with dbscan and supervised classification algorithms. In *ICADIWT*, pages 207–212. IEEE, 2014.

[39] C. Plant, S. J. Teipel, A. Oswald, C. Böhm, T. Meindl, J. M. Miranda, A. L. W. Bokde, H. Hampel, and M. Ewers. Automated detection of brain atrophy patterns based on mri for the prediction of alzheimer's disease. *NeuroImage*, 50(1):162–174, 2010.

[40] H. Rosling. *The Health and Wealth of Nations* . Gapminder Foundation, 2012.

[41] G. Schoier and G. Borruso. Individual movements and geographical data mining. clustering algorithms for highlighting hotspots in personal navigation routes. In B. Murgante, O. Gervasi, A. Iglesias, D. Taniar, and B. O. Apduhan, editors, *Computational Science and Its Applications - ICCSA 2011*, pages 454–465, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[42] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu. Dbscan revisited, revisited: Why and how you should (still) use dbscan. *ACM Trans. Database Syst.*, 42(3), July 2017.

[43] J.-B. Sheu. An emergency logistics distribution approach for quick response to urgent relief demand in disasters. *Transportation Research Part E: Logistics and Transportation Review*, 43:687–709, 11 2007.

[44] W. Sun, Y. He, and H. Chang. Regional characteristics of co2 emissions from china's power generation: Affinity propagation and refined laspeyres decomposition. *International Journal of Global Warming*, 11:38, 01 2017.

[45] Y. Tang, F. Sun, and Z. Sun. Improved validation index for fuzzy clustering. In *Proceedings of the 2005, American Control Conference, 2005.*, pages 1120–1125. IEEE, 2005.

[46] O. Tominaga, F. Ito, T. Hanai, H. Honda, and T. Kobayashi. Modeling of consumers' preferences for regular coffee samples and its application to product design. *Food Science and Technology Research - FOOD SCI TECHNOL RES*, 8:281–285, 08 2002.

[47] F. Tung, A. Wong, and D. A. Clausi. Enabling scalable spectral clustering for image segmentation. *Pattern Recognition*, 43(12):4069–4076, 2010.

[48] L. Vendramin, R. J. G. B. Campello, and E. R. Hruschka. Relative clustering validity criteria: A comparative overview. *Stat. Anal. Data Min.*, 3:209–235, 2010.

[49] N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.*, 11:2837–2854, dec 2010.

[50] U. von Luxburg. A tutorial on spectral clustering. *Stat. Comput.*, 17(4):395–416, 2007.

[51] U. von Luxburg. Clustering stability: An overview. *Foundations and Trends in Machine Learning*, 2(3):235–274, 2010.

[52] L. Wang, X. Zhou, Y. Xing, M. Yang, and C. Zhang. Clustering ecg heartbeat using improved semi-supervised affinity propagation. *IET Software*, 11, 06 2017.

[53] Wikipedia. Cluster Analysis Algorithms .

[54] Ward's Method .

[55] D. Wolpert and W. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1997.

[56] X. Xie and G. Beni. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):841–847, 1991.

[57] R. Yedida. Evaluating Clusters , 2019.

[58] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2005.