Hierarchical Clustering

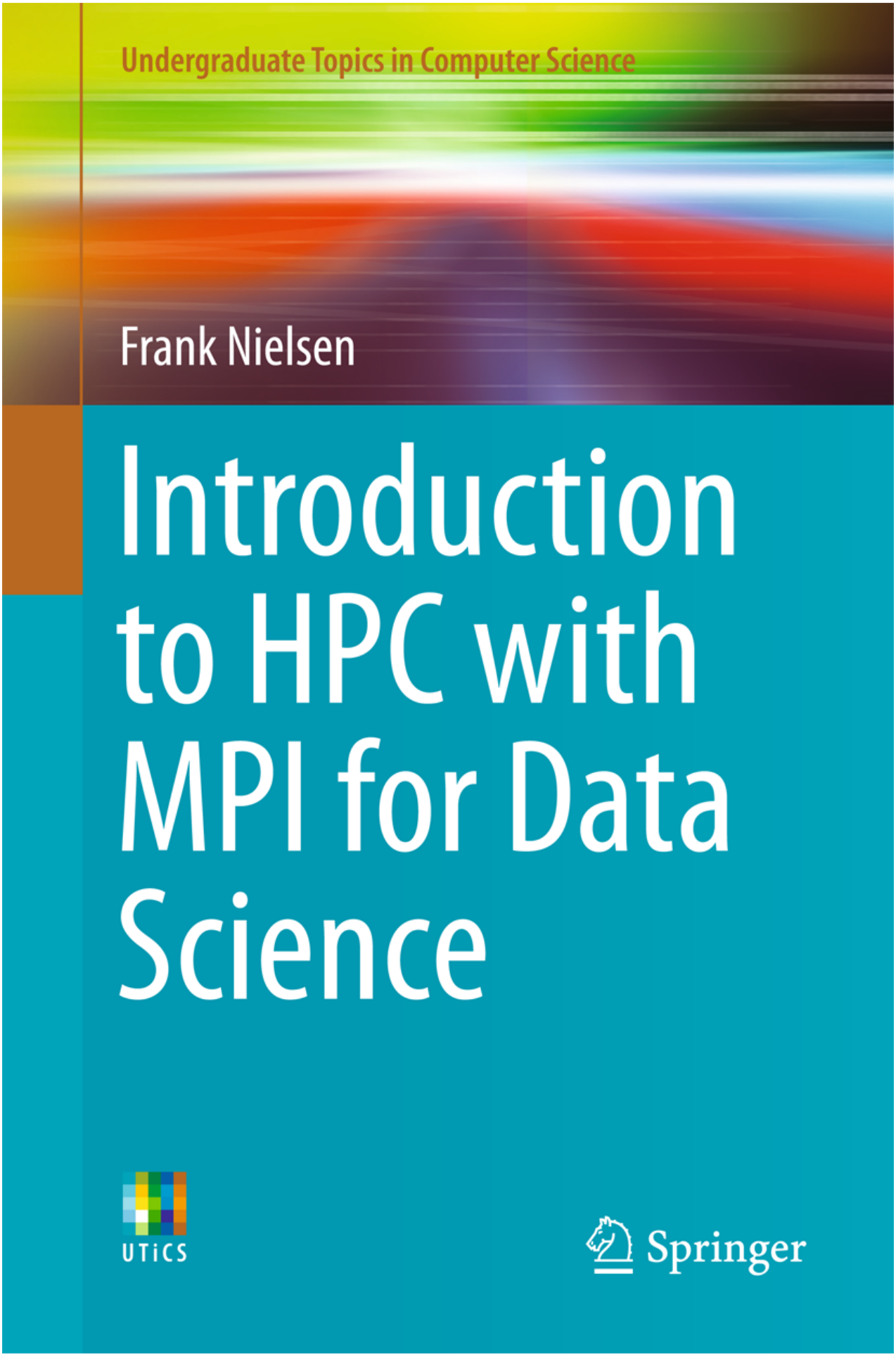
Chapter · February 2016

DOI: 10.1007/978-3-319-21903-5_8

1 author:

Frank Nielsen
Sony Computer Science Laboratories, Inc.
497 PUBLICATIONS **5,628** CITATIONS

SEE PROFILE


Undergraduate Topics in Computer Science
Frank Nielsen
Introduction to HPC with MPI for Data Science
UTiCS
Springer

8
Hierarchical clustering

A concise summary is provided at the end of this chapter, in §8.7.

8.1 Agglomerative versus divisive hierarchical clustering, and dendrogram representations

Hierarchical clustering is yet another technique for performing data exploratory analysis. It is an unsupervised technique. In the former clustering chapter, we have described at length a technique to partition a data-set $X = \{x_1, ..., x_n\}$ into a collection of groups called clusters $X = \uplus_{i=1}^{k} G_i$ by minimizing the k-means objective function (*i.e.*, the weighted sum of cluster intra-variances): In that case, we dealt with flat clustering that delivers a non-hierarchical partition structure of the data-set. To contrast with this flat clustering technique, we cover in this chapter another widely used clustering technique: Namely, *hierarchical clustering*.

Hierarchical clustering consists in building a binary merge tree, starting from the data elements stored at the leaves (interpreted as singleton sets) and proceed by merging two by two the "closest" sub-sets (stored at nodes) until we reach the root of the tree that contains all the elements of X. We denote by $\Delta(X_i, X_j)$ the distance between any two sub-sets of X, called the *linkage distance*. This technique is also called *agglomerative hierarchical clustering* since we start from the leaves storing singletons (the x_i's) and merge iteratively

subsets until we reach the root.

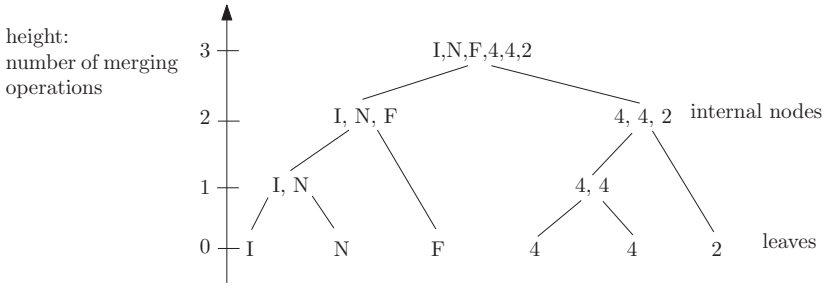


Figure 8.1 Drawing a dendrogram by embedding the nodes on the plane using a height function.

The graphical representation of this binary merge tree is called a *dendrogram*. This word stems from the greek *dendron* that means *tree* and *gramma* the means *draw*. For example, to draw a dendrogram, we can draw an internal node $s(X')$ containing a subset $X' \subseteq X$ at height $h(X') = |X'|$, where $|\cdot|$ denotes the cardinality of X', that is, its number of elements. We then draw edges between this node $s(X')$ and its two sibling nodes $s(X_1)$ and $s(X_2)$ with $X' = X_1 \cup X_2$ (and $X_1 \cap X_2 = \emptyset$). Figure 8.1 depicts conceptually the process of drawing a dendrogram. There exists several ways to visualize the hierarchical structures obtained by hierarchical clustering. For example, we may use special Venn diagrams using nested convex bodies, as depicted in Figure 8.2.

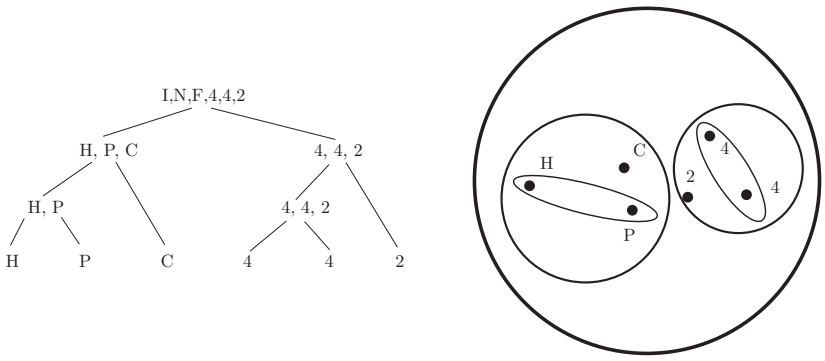


Figure 8.2 Several visualizations of a dendrogram: dendrogram (left) and equivalent Venn diagram (right) using nested ellipses (and disks).

Figure 8.3 shows such an example of a dendrogram that has been drawn from a agglomerative hierarchical clustering computed on a data-set provided

in the free multi-platform R language[1] (GNU General Public License). The (short) R code for producing this figure is the following:

```
d <- dist(as.matrix(mtcars))     # find distance matrix
hc <- hclust(d,method="average" )
plot(hc, xlab="x", ylab="height", main="Hierarchical␣
    clustering␣(average␣distance)", sub="(cars)")
```

We have chosen the *Euclidean distance* $D(x_i, x_j) = \|x_i - x_j\|$ as the basic distance between any two elements of X, and the minimum distance as the linkage distance for defining the *sub-set distance* $\Delta(X_i, X_j) = \min_{x \in X, y \in X_j} D(x, y)$. Here is an excerpt of that data-set that describes some features for the car data-set:

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

Notice that the visual drawing of hierarchical clusterings, dendrograms, conveys rich information for both qualitative and quantitative evaluations of various hierarchical clustering techniques that we shall present below.

To contrast with agglomerative hierarchical clustering, we also have *divisive hierarchical clustering* that starts from the root containing all the data-set X, and splits this root node into two children nodes containing respectively X_1 and X_2 (so that $X = X_1 \cup X_2$ and $X_1 \cap X_2 = \emptyset$), and so on recursively until we reach leaves that store in singletons the data elements. In the remainder, we concentrate on agglomerative hierarchical clustering (AHC) that is mostly used in applications.

[1] Download and install R from the following URL: `http://www.r-project.org/`

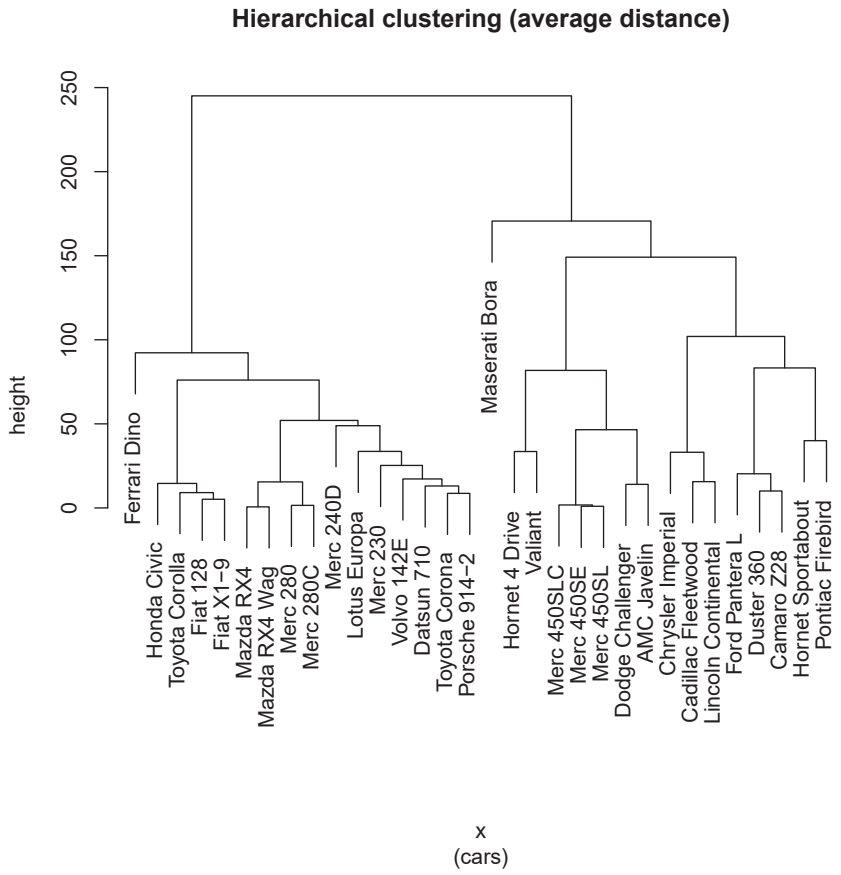


Figure 8.3 Example of a dendrogram for a car data-set: The data elements are stored at the leaves of the binary merge tree.

8.2 Strategies to define a good linkage distance

Let $D(x_i, x_j)$ denote the elementary distance between any two elements of X (for example, the Euclidean distance). In order to select at each stage of the hierarchical clustering the closest pair of sub-sets, we need to define a sub-set distance $\Delta(X_i, X_j)$ between any two sub-sets of elements. Of course, when both sub-sets are singletons $X_i = \{x_i\}$ and $X_j = \{x_j\}$, we should have $\Delta(X_i, X_j) = D(x_i, x_j)$. We present below three such common *linkage functions*:

1. *Single Linkage* (SL):

$$\Delta(X_i, X_j) = \min_{x_i \in X_i, x_j \in X_j} D(x_i, x_j)$$

2. *Complete Linkage* (CL) (or diameter):

$$\Delta(X_i, X_j) = \max_{x_i \in X_i, x_j \in X_j} D(x_i, x_j)$$

3. *Group Average Linkage* (GAL):

$$\Delta(X_i, X_j) = \frac{1}{|X_i||X_j|} \sum_{x_i \in X_i} \sum_{x_j \in X_j} D(x_i, x_j)$$

Figure 8.4 visualizes pictorially those three different linkage functions.

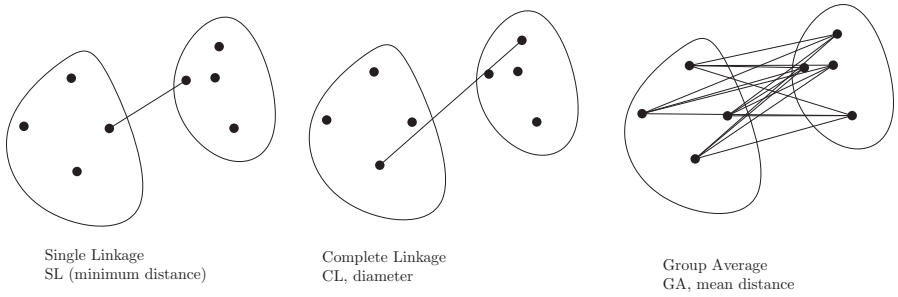


Figure 8.4 Illustrating the common linkage functions defining distances between sub-sets: single linkage, complete linkage and group average linkage.

There exist many other sub-set distances Δ that are commonly called linkage distances because they literally allow one to *link sub-trees* representing the sub-sets in the dendrogram representation.

8.2.1 A generic algorithm for agglomerative hierarchical clustering

We summarize below the principle of the generic agglomerative hierarchical clustering (AHC) for a prescribed linkage distance $\Delta(\cdot, \cdot)$ (user-defined and relying on yet another used-defined element distance):

Algorithm **AHC**

- Initialize for each data element $x_i \in X$ its cluster singleton $G_i = \{x_i\}$ in a list
- While there remains two elements in the list, do:
  - Choose G_i and G_j so that $\Delta(G_i, G_j)$ is minimized among all pairs,
  - Merge $G_{i,j} = G_i \cup G_j$, and
    - add $G_{i,j}$ to the list, and
    - remove G_i and G_j from the list.
- Return the remaining group in the list ($G_{\text{root}} = X$) as the dendrogram root.

Since we start from $n = |X|$ leaves to finish with a root containing the full set X, we perform exactly $n - 1$ merge operations. A straightforward implementation of this AHC algorithm yields a cubic time complexity, in $O(n^3)$. Depending on the linkage distance, we can optimize this naive algorithm and obtain far better time complexities.

Observation 4

Notice that in general the dendrogram may not be unique for a linkage distance function: Indeed, there can be *several* "closest" pairs of subsets, but we choose only one pair at each iteration and reiterate (thus breaking the symmetry, say, by introducing a lexicographic order on the pairs). In other words, if we had applied a permutation σ on the elements of X, and re-run the AHC algorithm, we could have obtained another dendrogram in output. For numerical data, we can slightly perturbate the initial data-set by adding some small random noise drawn uniformly in $(0, \epsilon)$ to bypass this problem. However, for *categorical data*, the problem still remains and therefore careful attention should be given to handle this problem.

The standard optimized AHC algorithm is called SLINK [79] (1973), and has a quadratic complexity, in $O(n^2)$ time. Single-linkage AHC yields a "*chaining phenomenon*" in dendrograms as depicted in Figure 8.5. The AHC algorithm with complete linkage (also called diameter linkage) is called CLINK [23] (1977), and can be computed in $O(n^2 \log n)$ time. One disadvantage of complete linkage is that it is very sensitive to *outliers* (that is, artifact data that should have been removed beforehand when possible — the cleaning stage of data-sets). At first glance, the group average AHC is more computationally costly to compute but can also be optimized as well to get a sub-cubic time complexity. Usually, we recommend in applications the group

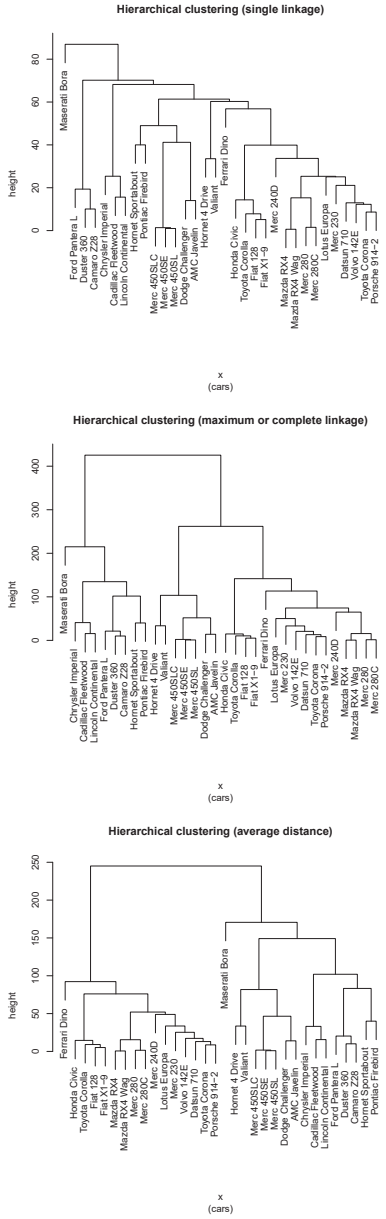


Figure 8.5 Comparisons of dendrograms obtained from agglomerative hierarchical clustering for three commonly used linkage functions: single linkage (top), complete linkage (middle) and group average linkage (bottom).

average AHC algorithm that does not produce chaining phenomena and is more robust to noisy input.

8.2.2 Choosing the appropriate elementary distance between elements

The *base distance function* $D(\cdot,\cdot)$ plays a crucial role on the shape of dendrograms. This distance function is a *dissimilarity measure* that evaluates how different element x_i is from element x_j (for any pair of elements). Although we often use the Euclidean distance, we can also choose other *metric distances*[2] like the *city block distance* (called the *Manhattan distance* or the L_1-norm induced distance[3]):

$$\boxed{D_1(p,q) = \sum_{j=1}^{d} |p^j - q^j|}$$

Recall that we use the super-script notation $x = (x^1, ..., x^j, ..., x^d)$ for an attribute vector x with d components: the x^j's are the coordinates of a d-dimensional vector x.

We can also use the *Minkowski distances* that generalize both the Euclidean distance (for $m = 2$) and the Manhattan distance (for $m = 1$):

$$\boxed{D_m(p,q) = \left(\sum_{j=1}^{d} |p^j - q^j|^m\right)^{\frac{1}{m}}} = \|p-q\|_m, m \geq 1$$

When the data coordinates have different scale factors, or are correlated, we better use the *Mahalanobis distance*[4]:

$$\boxed{D_\Sigma(p,q) = \sqrt{(p-q)^\top \Sigma^{-1}(p-q)} = D_2(L^\top p, L^\top q)}$$

with the *precision matrix* (inverse of the *covariance matrix*) $\Sigma^{-1} = L^\top L$ being factorized by the *Cholesky matrix* (matrix L is a *lower triangular matrix*). That is, the Mahalanobis distance $D_\Sigma(p,q)$ amounts to compute a traditional Euclidean distance $D_2(L^\top p, L^\top q)$ after an *affine change of variable*: $x \leftarrow L^\top x$. Matrix Σ is called the covariance matrix, and its inverse matrix Σ^{-1} is called

[2] satisfying the symmetry ($D(p,q) = D(q,p)$, the law of indiscernability ($D(p,q) = 0$ if and only if $p = q$), and the triangular inequality (for all triples $D(p,q) \leq D(p,r) + D(q,r)$). See Section 8.5 that introduces ultra-metrics.

[3] A norm $\|.\|$ induces a distance $D(p,q) = \|p-q\|$

[4] A metric distance that is symmetric and satisfies the triangle inequality.

the precision matrix. We can estimate the covariance matrix from a data-set sample $x_1, ..., x_n$ by computing:

$$\Sigma = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top,$$

with $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ the empirical mean, also called the *sample mean*.

For categorical data (that is non-numerical), we often use an *agreement distance* like the *Hamming distance*:

$$\boxed{D_H(p,q) = \sum_{j=1}^{d} 1_{[p^j \neq q^j]}}$$

where $1_{[a\neq b]} = 1$ if and only if $a \neq b$, and zero otherwise. That is, the Hamming distance counts the number of times corresponding attributes are different from each other. The Hamming distance is a metric distance.

Often, we can link a *similarity* measure to a *dissimilarity* measure, and vice-versa. For example, considering the Hamming distance on d-dimensional binary vectors, we can define the corresponding similarity measure by $S_H(p,q) = \frac{d-D_H(p,q)}{d}$ (with $0 \leq S_H(p,q) \leq 1$, and maximal similarity when $p = q$).

There exist many other distance functions that have been used in a broad panel of applications. Let us cite the *Jaccard distance* $D_J(A,B) = \frac{|A\cap B|}{A\cup B}$ defined on sets, the *edit distance* for finding distance between combinatorial structures (like texts or DNA sequences), the *cosine distance* $D_{\cos}(p,q) = 1 - \frac{p^\top q}{\|p\|\|q\|}$ (very useful when analyzing a corpus of texts with documents represented by a frequency histogram of word occurencies), etc.

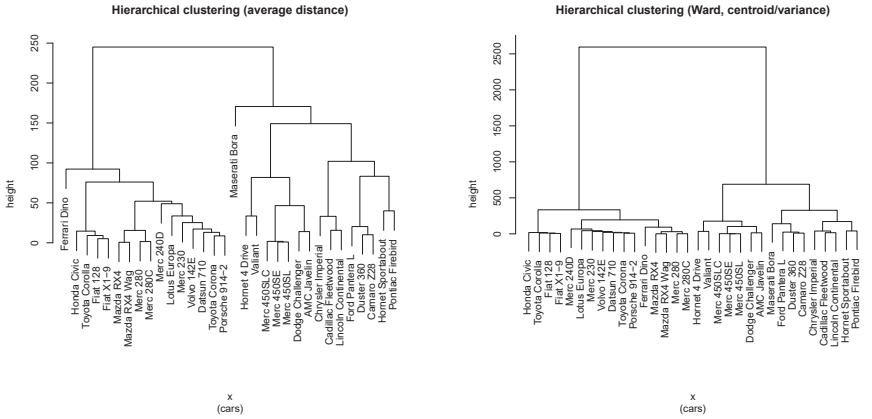
8.3 Ward merging criterion and centroids

One can also take a sub-set distance Δ according to the centroids of the sub-sets. This criterion allows us to implement a variance minimization process. This yields the *Ward linkage* function: To merge X_i ($n_i = |X_i|$) with X_j ($n_j = |X_j|$), we consider the following Ward criterion:

$$\boxed{\Delta(X_i, X_j) = \frac{n_i n_j}{n_i + n_j}\|c(X_i) - c(X_j)\|^2}$$

where $c(X')$ denotes the centroid of subset $X' \subseteq X$: $c(X') = \frac{1}{|X'|}\sum_{x\in X'} x$ (we may consider weighted points too). Observe that the distance between two elements induced from the sub-set distance Δ is merely half of the squared

Euclidean distance: $\Delta(\{x_i\},\{x_j\}) = D(x_i,x_j) = \frac{1}{2}\|x_i - x_j\|^2$. Figure 8.6 illustrates visually the difference between the dendrograms obtained from the group average AHC and from the Ward AHC (of minimal variance).



(a) (b)

Figure 8.6 Comparing dendrograms obtained for (a) the group average linkage, and (b) Ward linkage.

Notice that we can always define the similarity $S(X_i,X_j)$ between two sub-sets by defining $S(X_i,X_j) = -\Delta(X_i,X_j)$. The merging steps of a path sequence of length l in the tree dendrogram are said monotonous when we have the property that $S_1 \geq S_2 \geq ... \geq S_l$. A hierarchial clustering is said *non-monotonous* when there exists at least one *inversion*, say $S_i < S_{i+1}$, on a path from the leaves to the root of the dendrogram. The Ward AHC is not monotonous because there can exist inversions. However, the single linkage, complete linkage and group average linkage are all guaranteed to be monotonous.

When we draw the nodes of the merge tree (*i.e.*, nodes of the dendrogram) using a height function defined as the similarity, an inversion in a dendrogram can be noticed graphically by the fact that one horizontal height line can be lower than another horizontal height line for a former merging operation. Indeed, this contradicts the fact the nodes on a path from a leaf to the root should be y-monotonous. Figure 8.7 illustrates an inversion in a dendrogram.

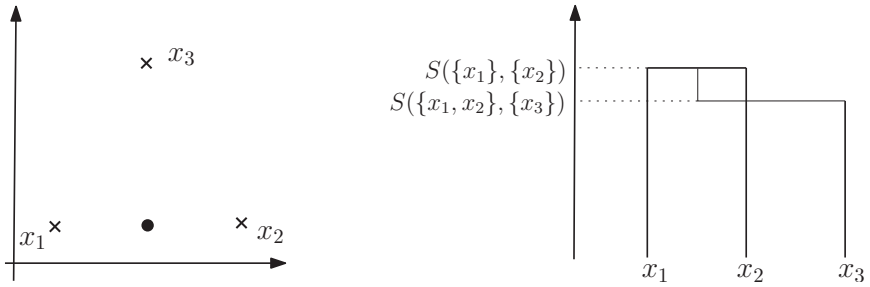


Figure 8.7 Example of an inversion phenomenon in a dendrogram obtained when using Ward's criterion for hierachical clustering on a toy data-set of a triple of elements.

8.4 Retrieving flat partitions from dendrograms

From a dendrogram, we can extract many different flat partitions. Figure 8.8 illustrates this concept by displaying two constant-height cuts that induce respective partitions of the data sets. Note that the cutting path on the dendrogram *does not need* to be at constant height in general (see exercise 8.8).

8.5 Ultra-metric distances and phylogenetic trees

A distance function $D(\cdot,\cdot)$ is called *a metric* if it satisfies the following three axioms:

Law of indiscernability. $D(x,y) \geq 0$ with equality iff. $x = y$,

Symmetry. $D(x,y) = D(y,x)$

Triangular inequality. $D(x,y) \leq D(x,z) + D(z,y)$,

The Euclidean distance and the Hamming distance are two examples of metric distances. Beware that the squared Euclidean distance is not a metric although it is symmetric and satisfies the law of indiscernability. Indeed, the triangular inequality is not anymore satisfied when we take the square of the Euclidean distance (however, recall that the squared Euclidean distance is used to define the potential function of the k-means in flat clustering in order to get centroids and minimizes of cluster variances). The law of indiscernability can further be split into two sub-axioms: The law of non-negativity $D(p,q) \geq 0$, and the law of reflexivity: $D(p,q) = 0 \Leftrightarrow p = q$.

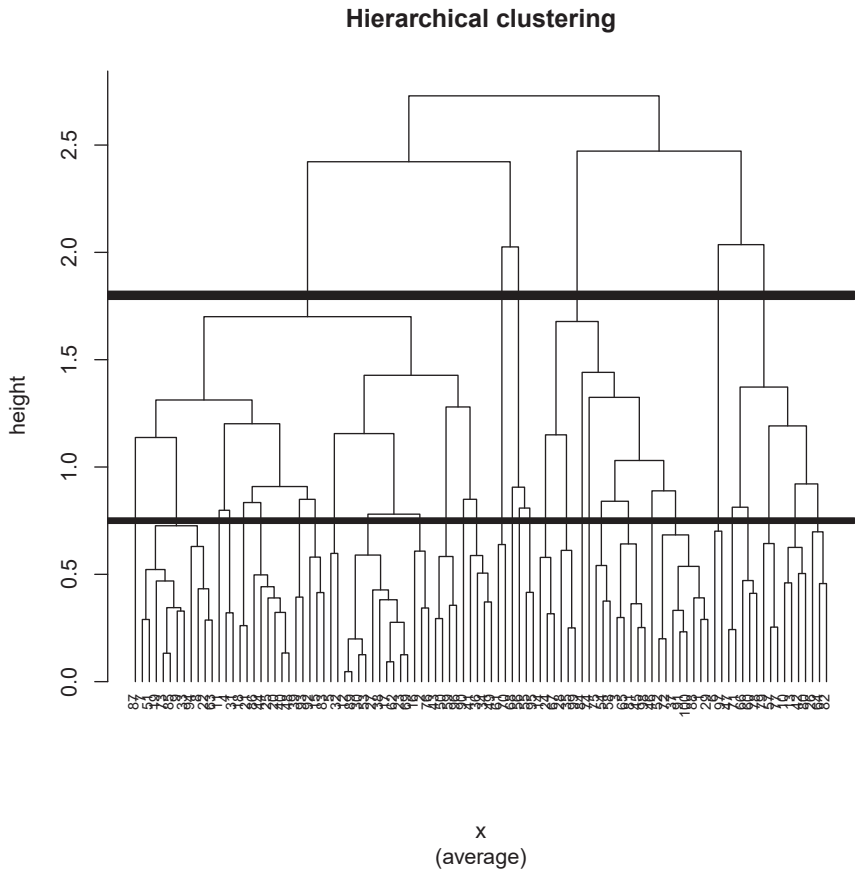


Figure 8.8 Retrieving flat partitions from a dendrogram: We choose the height for cutting the dendrogram. At a given height, we obtain a flat clustering (that is a partition of the full data-set). The cut path does not need to be at a constant height. Thus a dendrogram allows one to obtain many flat partitions. Here, we show two different cuts at constant height, for $h = 0,75$ and $h = 1,8$.

Hierarchical clustering is tightly linked to a class of distances called the class of *ultra-metrics*. A distance is said ultra-metric if it is a metric and further ensures that:

$$D(x, y) \leq \max_z (D(x, z), D(z, y)).$$

Let us now explain the link between ultra-metrics and hierarchical clustering: In evolution theory, species evolve with time, and the distance between species is represented by a so-called *phylogenetic tree*. Let us write for short

$D_{i,j} = D(x_i, x_j)$. A tree is said *additive* if and only if we can attach to each edge a weight so that for each pair of leaves, the distance between them is equal to the sum of the distances of the edges linking them. A tree is said ultra-metric when the distance between two leaves, say i and j, and their *common ancestor*, say k, is equal: $D_{i,k} = D_{j,k}$. We can draw an ultra-metric tree by choosing the height distance $\frac{1}{2}D_{i,j}$ for visualizing a dendrogram. This distance can be interpreted as a clock time among all the elements of X (for species, it represents the biological time).

The group average AHC guarantees to produce an ultra-metric tree. We shall call this hierarchical clustering that embeds the nodes of the tree with its height the *Unweighted Pair Group Method using arithmetic Averages* algorithm (or UPGMA, for short). We write in pseudo-code this algorithm below:

Algorithm **UPGMA** :

- For all i, initialize x_i to its cluster $C_i = \{x_i\}$, and set this node leaf to height 0.
- While there remains at least two clusters:
  - Find the closest pair of clusters C_i and C_j that minimizes the group average distance $\Delta_{i,j}$,
  - Define a new cluster $C_k = C_i \cup C_j$ and compute the distance $\Delta_{k,l}$ for all l,
  - Add a node k to the children C_i and C_j, and set the height of that node to $\frac{1}{2}\Delta(C_i, C_j)$,
  - Remove both C_i and C_j from the cluster list, and reiterate until we get two remaining clusters.
- For the last two clusters C_i and C_j, set the root node at height $\frac{1}{2}\Delta(C_i, C_j)$.

Theorem 9

When the matrix distance $M = [D_{i,j}]_{i,j}$ with $D_{i,j} = D(x_i, x_j)$ of a data-set X satisfies the ultra-metric property, then there exists a unique ultra-metric tree that can be built with the UPGMA algorithm.

Phylogenetic trees are often used when modeling the evolution of species: We associate to the vertical axis the chronological time of evolution, as depicted in Figure 8.9. The UPGMA allows to build such an ultra-metric tree. However let us emphasize that data-sets are often noisy and therefore the matrix distance is often not ultra-metric since corrupted. Another drawback is that we need to

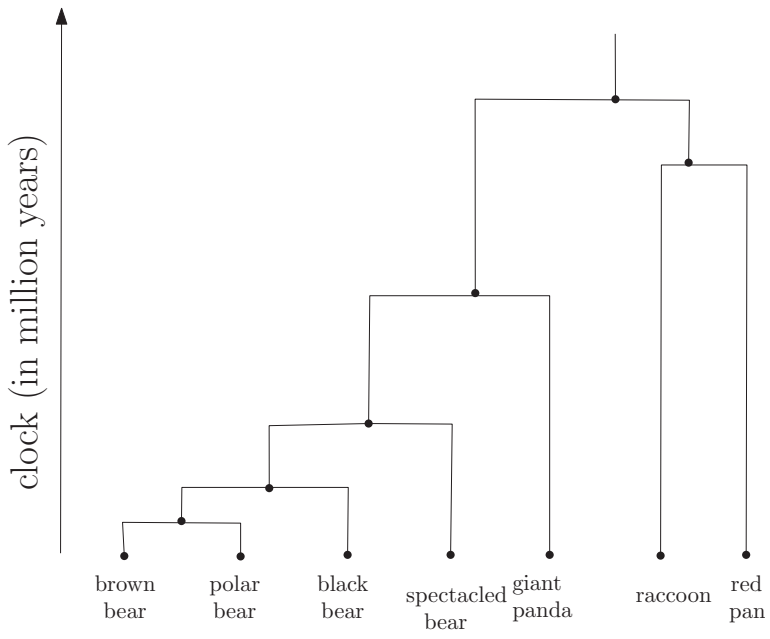


Figure 8.9 Dendrograms and phylogenetic trees for visualizing the evolution of species.

consider the matrix of pairwise distances that requires a quadratic memory space, and can therefore only be limited to reasonable size data-sets (but not big data as is!).

8.6 Notes and references

There exist many hierarchical clustering algorithms. Let us cite *SLINK* [79] (Single Linkage, 1973), *CLINK* [23] (Complete Linkage, 1977), and a general survey [68] providing a high-level abstraction of hierarchical clustering. Although that flat clustering minimizing the k-means objective function is NP-hard (even in the plane), it has been recently proved (2012) that we can extract from a single linkage hierarchical clustering the optimal k-means clustering provided that some stability criterion is satisfied, see [6] (the extraction of the flat partition is performed using dynamic programming to find the best non-constant height dendrogram cut). The hierarchical clustering that minimizes Ward's variance criterion and its related criteria have been thoroughly investigated in [89, 69]. Various hierarchical clustering algorithms

(including SLINK, CLINK and Ward) can be unified in the generic Lance-Williams framework, see [57] and exercise 8.8. Uniqueness and monotonic properties of hierarchical clustering have been studied in [67]. Although that hierarchical clustering algorithms are *a priori* harder to parallelize compare to flat clustering techniques (like k-means), let us mention this work [74] that reports an efficient parallel algorithm. We refer to [70] for an explanation of the divisive hierarchical clustering technique that maximizes the notion of modularity. Distances is at the core of many algorithms: We recommend the encyclopedia of distances [24] for a compact review of main distances.

8.7 Summary

Agglomerative hierarchical clustering differs from partition-based clustering since it builds a binary merge tree starting from leaves that contain data elements to the root that contains the full data-set. The graphical representation of that tree that embeds the nodes on the plane is called a dendrogram. To implement a hierarchical clustering algorithm, one has to choose a linkage function (single linkage, average linkage, complete linkage, Ward linkage, etc.) that defines the distance between any two sub-sets (and rely on the base distance between elements). A hierarchical clustering is monotonous if and only if the similarity decreases along the path from any leaf to the root, otherwise there exists at least one inversion. The single, complete, and average linkage criteria guarantee the monotonic property, but not the often used Ward's criterion. From a dendrogram, one can extract many data-set partitions that correspond to flat clustering output. Phylogenetic trees used to model the evolution of species are ultra-metric trees. Hierarchical clustering using the average linkage guarantees to build an ultra-metric tree when the base distance between any two elements is ultra-metric.

8.8 Exercises

Exercise 1: *Checking the ultra-metric property of a distance matrix*

Let M denote a square matrix of dimension $n \times n$ that stores at index (i, j) the distance $D(x_i, x_j)$ between element x_i and element x_j.

– Design an algorithm that checks whether the distance matrix satisfies the ultra-metric property or not,

– What is the time complexity of your algorithm?

Exercise 2: *Euclidean metric distance and Hamming metric distance*

– Prove that the Euclidean distance is a metric, but not the squared Euclidean distance.

– Prove that the Hamming distance satisfies the axioms of a metric.

– Prove that the distance $D(p,q) = \left(\sum_{j=1}^{d} |p^j - q^j|^m\right)^{\frac{1}{m}}$ for $0 < m < 1$ is not a metric (when $m \geq 1$, recall that it is the m-norm induced Minkowski metric distance).

Exercise 3: *Combining flat clustering with hierarchical clustering*

Let $X = \{x_1, ..., x_n\}$ be n data elements, each datum has d attributes.

– Give an algorithm that clusters hierarchically the data, and retrieve a partition of at most l elements (for large l, it produces an over-clustering), and use after a k-means algorithm on the centroids of these groups. What kind of applications can you think of that strategy?

– What is the complexity of your algorithm? Explain its advantages compare to only hierarchical clustering or to only partition-based clustering?

Exercise 4: *Hierarchical clustering of Lance and Williams [57]*

– State the hierarchical clustering algorithm using the following shortcut notations $D_{ij} = \Delta(C_i, C_j)$ and $D_{(ij)k} = \Delta(C_i \cup C_j, C_k)$ for disjoint groups C_i, C_j and C_k.

– A hierarchical clustering belongs to the Lance-Williams family if and only if it can be written canonically as:

$$D_{(ij)k} = \alpha_i D_{ik} + \alpha_j D_{jk} + \beta D_{ij} + \gamma |D_{ik} - D_{jk}|,$$

with $\alpha_i, \alpha_j, \beta$, and γ parameters depending on the size of clusters. Prove that Ward minimum variance criterion ($D(x_i, x_j) = \|x_i - x_j\|^2$) for disjoint groups C_i, C_j and C_k yields the following formula:

$$D(C_i \cup C_j, C_k) = \frac{n_i + n_k}{n_i + n_j + n_k} D(C_i, C_k) + \frac{n_j + n_k}{n_i + n_j + n_k} D(C_j, C_k) - \frac{n_k}{n_i + n_j + n_k} D(C_i, C_j).$$

– Deduce that Ward's algorithm is a particular case Lance-Williams's generic hierarchical clustering with the following parameterization:

$$\alpha_l = \frac{n_l + n_k}{n_i + n_j + n_k}, \qquad \beta = \frac{-n_k}{n_i + n_j + n_k}, \qquad \gamma = 0.$$

– Prove that Lance-Williams' algorithm unify single linkage, complete linkage and group average linkage.

Exercise 5: *Centroid-based hierarchical clustering for an arbitrary convex distance function*

For a convex distance $D(\cdot, \cdot)$, let us define the centroid of X as the unique minimizer of $\min_c \sum_{x \in X} D(x, c)$. Prove that the inversion phenomenon that can happen for Ward criterion does not happen for the Euclidean distance nor for the Manhattan distance (two examples of convex distances).

Exercise 6: * *Retrieving the best k-means flat partition from a hierarchical clustering [6]*

Given a dendrogram, one can extract many different partitions:

– How many distinct partitions can be retrieved from a dendrogram?

– For a sub-set X', let us denote by $c(X')$ the centroid of X' and by $v(X')$ its variance: $v(X') = \frac{1}{|X'|} \sum_{x \in X'} x^\top x - (c(X')^\top c(X'))^2$. Give a dynamic programming code for retrieving the best k-means flat clustering from a dendrogram. What is the time complexity of your algorithm?

Exercise 7: * *Cosine distances between documents and spherical k-means*

Let p and q be two vectors of d attributes, and consider the cosine distance: $D(p, q) = \cos(\theta_{p,q}) = 1 - \frac{p^\top q}{\|p\| \|q\|}$. The cosine distance is an angular distance that does not account for the magnitude of vectors. For a collection of text

documents, we model a text t by its word frequency/counting vector $f(t)$ (given a word dictionary).

– Prove that the cosine distance is a metric,

– Design an agglomerative hierarchical clustering that allows one to cluster text documents,

– Generalize the k-means flat clustering to a partition-based clustering algorithm relying on the cosine distance. We shall consider attribute vector as a point set lying on the unit sphere, and prove that the spherical centroid is the Euclidean centroid projected back to the unit sphere (when all points are enclosed into the same hemisphere). How to define the spherical centroid of two antipodal points on the unit sphere centered at the origin?

Exercise 8: ** Hierarchial clustering for Bregman divergences [83]*

Bregman divergences are non-metric distances that are defined according a strictly convex and differentiable convex generator function $F(x)$ by:

$$D_F(x,y) = F(x) - F(y) - (x-y)^\top \nabla F(y),$$

where $\nabla F(y) = (\frac{\mathrm{d}}{\mathrm{d}y^1}F(y), ..., \frac{\mathrm{d}}{\mathrm{d}y^d}F(y))$ denotes the gradient vector.

– Prove that for $F(x) = x^\top x$, the Bregman divergence amounts to the squared Euclidean distance.

– Prove that Bregman divergences can never be a metric, and that the squared Mahalanobis distance is a symmetric Bregman divergence.

– Generalize Ward's criterion for Bregman divergences as follows:

$$\Delta(X_i, X_j) = |X_i| \times D_F(c(X_i), c(X_i \cup X_j)) + |X_j| \times D_F(c(X_j), c(X_i \cup X_j)),$$

where $c(X_l)$ is the center of mass of X_l. Check that for the Bregman generator $F(x) = \frac{1}{2}x^\top x$, we get the usual Ward's criterion.

– Report a Bregman hierarchical clustering algorithm. Can inversion phenomena happen?

Exercise 9: *** Single linkage hierarchical clustering and minimum spanning tree [36]*

Give a naive implementation of the single linkage hierarchical clustering. What is the time complexity of your naive algorithm? Given a planar point set $X =$

$\{x_1, ..., x_n\}$, the Euclidean Minimum Spanning Tree (MST) is a tree with nodes anchored at all points of X so that the sum of all tree edge lengths is minimized. Prove that the MST is a subgraph of the Delaunay triangulation (the dual structure of the Voronoi diagram). Prove that the edge information contained in the Euclidean minimum spanning tree allows one to easily deduce the structure of the single linkage dendrogram. As a byproduct, report a quadratic time algorithm for the single linkage hierarchical clustering.