Cluster Analysis and Applications

May 23, 2020

Contents

Chapter 2

Representatives

In applied research it is often necessary to represent a given set of data by a single datum which, in some sense, encompasses most of the features (properties) of the given set. The quantity commonly used is the well-known arithmetic mean of the data. For example, a student's grade point average can be expressed by arithmetic mean, but it wouldn't be appropriate to represent the average rate of economics growth during several years in such a way (see [31]).

In order to determine a best representative of a given set, first one has to decide how to measure the distance between points of the set. Of course, one could use some standard metric function, but various applications show (see e.g. [17]) that to measure the distance it is more useful to take a function which does not necessarily satisfy all the properties of a metric function (see [17, 37]).

Definition 2.1. A function $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, which satisfies[1]

(i) $d(x,y) = 0 \Leftrightarrow x = y$

(ii) $x \mapsto d(x,y)$ is continuous on $\mathbb{R}^n$ for every fixed $y \in \mathbb{R}^n$

(iii) $\lim\limits_{\|x\|\to\infty} d(x,y) = +\infty$ for every fixed $y \in \mathbb{R}^n$

will be called a ***distance-like function***.

It is readily seen that every ℓ_p metric, $p \geq 1$, is a distance-like function, but an important example is the well known *least squares* (LS) *distance-like function* $d_{LS}(x,y) = \|x-y\|^2$, where $\|\ \|$ is the usual 2-norm.[2]

[1] We use the following notation: $\mathbb{R}_+ = \{x \in \mathbb{R} : x \geq 0\}$ and $\mathbb{R}_{++} = \{x \in \mathbb{R} : x > 0\}$.

[2] If there is no risk of misapprehension, throughout the text we are going to use $\|\ \|$ to denote the Euclidean, i. e. the 2-norm $\|\ \|_2$.

In general, a distance-like function is neither symmetric nor does it satisfy the triangle inequality. But, as shown by the following lemma, given a finite set of data points[3] $\mathcal{A} = \{a^i = (a^i_1, \dots, a^i_n) : i = 1, \dots, m\} \subset \mathbb{R}^n$ with weights $w_1, \dots, w_m > 0$, there exists a point $c^\star \in \mathbb{R}^n$ such that the sum of its weighted d-distances to the points of $\mathcal{A}$ is minimal.

Lemma 2.2. *Let $\mathcal{A} = \{a^i : i = 1, \dots, m\} \subset \mathbb{R}^n$ be a set of data points with weights $w_1, \dots, w_m > 0$, let $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ be a distance-like function, and let $F\colon \mathbb{R}^n \to \mathbb{R}_+$ be the function given by*

$$F(x) = \sum_{i=1}^{m} w_i \, d(x, a^i) \,. \tag{2.1}$$

Then there exists a point $c^\star \in \mathbb{R}^n$ such that

$$F(c^\star) = \min_{x \in \mathbb{R}^n} F(x) \,. \tag{2.2}$$

Proof. Since $F(x) \geq 0$, $x \in \mathbb{R}^n$, there exists $F^\star := \inf\limits_{x \in \mathbb{R}^n} F(x)$. Let (c_k) be some sequence in $\mathbb{R}^n$ such that $\lim\limits_{k\to\infty} F(c_k) = F^\star$. Let us show that the sequence (c_k) in $\mathbb{R}^n$ is bounded. In order to do this, assume the contrary, i. e. that there exists a subsequence (c_{k_ℓ}) such that $\|c_{k_\ell}\| \to \infty$. Then, according to properties (ii) and (iii) from Definition 2.1, it follows that $\lim\limits_{\|c_{k_\ell}\|\to\infty} F(c_{k_\ell}) = +\infty$, and therefore the function F cannot attain its infimum. Finally, the sequence (c_k), being bounded, has a convergent subsequence (c_{k_j}), and let $c^\star$ be its limit point. Then $F(c^\star) = F(\lim\limits_{j\to\infty} c_{k_j}) = \lim\limits_{j\to\infty} F(c_{k_j}) = \lim\limits_{k\to\infty} F(c_k) = F^\star$, showing that (2.2) holds true. □

Remark 2.3. Note that for a global minimum point $c^\star \in \mathbb{R}^n$ and all $x \in \mathbb{R}^n$,

$$F(x) = \sum_{i=1}^{m} w_i \, d(x, a^i) \geq \sum_{i=1}^{m} w_i \, d(c^\star, a^i) = F(c^\star) \,, \tag{2.3}$$

holds true, and the equality holds if and only if $x = c^\star$, or some other point satisfying (2.4).

Lemma 2.2 enables the following definition:

[3] We are going to use upper indices for elements $a^i \in \mathbb{R}^n$, and the lower indices for the coordinates of elements in $\mathbb{R}^n$.

Definition 2.4. Let $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ be a distance-like function. A best representative of the set $\mathcal{A}$ with weights $w_1, \dots, w_m > 0$, with respect to the distance-like function d, is any point

$$c^\star \in \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \sum_{i=1}^{m} w_i \, d(x, a^i)\,. \tag{2.4}$$

The notation (2.4) suggests that best representative might not be unique, i.e. there might exist more best representatives of the set $\mathcal{A}$.

In this chapter we shall consider the two most commonly used representatives of a data set — *arithmetic mean* and *median*.

2.1 Representative of data sets with one feature

A set of data without weights, with a single feature is usually interpreted as a finite subset $\mathcal{A} = \{a^1, \dots, a^m\}$ of real numbers $a^i \in \mathbb{R}$, $i = 1, \dots, m$.

The two most frequently used distance-like functions on $\mathbb{R}$ are the *LS distance-like function* and ℓ_1 metric, also known as the *Manhattan* or *taxicab metric function* (see e.g. [6, 7, 16, 27])

$$d_{LS}(x, y) = (x - y)^2, \qquad \text{[LS distance-like function]}$$
$$d_1(x, y) = |x - y|\,. \qquad \text{[}\ell_1\text{ metric function]}$$

Exercise 2.5. Check whether

$$d_1(x, y) = d_2(x, y) = d_\infty(x, y) = d_p(x, y), \quad p \geq 1, \quad x, y \in \mathbb{R},$$

holds, where d_p is the p-metric on $\mathbb{R}$ (see e.g. [38]).

Exercise 2.6. Show that the function d_{LS} is not a metric function on $\mathbb{R}$, but the function d_1 is a metric on $\mathbb{R}$.

2.1.1 Best LS-representative

In case of the LS distance-like function, the function (2.1) becomes

$$F_{LS}(x) := \sum_{i=1}^{m} (x - a^i)^2\,, \tag{2.5}$$

and because it is a convex function and $F'_{LS}(c^\star_{LS}) = 0$ and $F''_{LS}(x) = 2m > 0$ for all $x \in \mathbb{R}$, it attains its global minimum at the unique point

$$c^\star_{LS} = \operatorname*{arg\,min}_{x \in \mathbb{R}} \sum_{i=1}^{m} d_{LS}(x, a^i) = \tfrac{1}{m} \sum_{i=1}^{m} a^i\,. \tag{2.6}$$

Hence, the best LS-representative of the set $\mathcal{A} \subset \mathbb{R}$ is the ordinary *arithmetic mean*[4], and it has the property (cf. Remark 2.3) that the sum of squared deviations to the given data is minimal:

$$\sum_{i=1}^{m}(x - a^i)^2 \geq \sum_{i=1}^{m}(c_{LS}^\star - a^i)^2, \tag{2.7}$$

where the equality holds for $x = c_{LS}^\star$.

As a measure of dispersion of a data set $\mathcal{A}$ around the arithmetic mean $c_{LS}^\star$, in statistics literature [**?**] one uses the *variance of data* (average squared deviation)

$$s^2 = \tfrac{1}{m-1}\sum_{i=1}^{m}(c_{LS}^\star - a^i)^2. \tag{2.8}$$

The number s is called the *standard deviation*.

Example 2.7. Given the set $\mathcal{A} = \{2, 1.5, 2, 2.5, 5\}$, its arithmetic mean is $c_{LS}^\star = 2.6$.

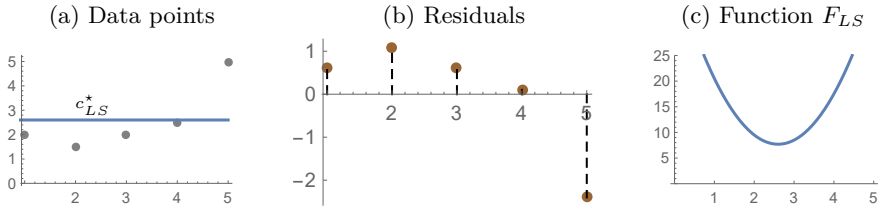


Figure 2.1: Arithmetic mean of the set $\mathcal{A} = \{2, 1.5, 2, 2.5, 5\}$

Figure 2.1a shows the data and the arithmetic mean $c_{LS}^\star$, Figure 2.1b shows the so-called *residuals* (the numbers $c_{LS}^\star - a^i$), and Figure 2.1c depicts the graph of the function F_{LS}. Note that the graph is a parabola and $F_{LS}(c_{LS}^\star) = 7.7$. What is the variance and what is the standard deviation of this set?

What would happen if there were an *outlier* (strongly jutting datum) among the data? How would it effect the best LS-representative (arithmetic mean) of the set $\mathcal{A}$? What would be the result if 5 were changed to 10?

[4]The problem of finding the best LS-representative of a data set occurs in the literature as the *least squares principle*, which was proposed in 1795 by German mathematician Carl Friedrich Gauss (1777–1855) while investigating the movements of celestial bodies, published in *Teoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*, Perthes and Besser, Hamburg, 1809. One should also mention that in 1805 French mathematician Adrien-Marie Legendre (1752–1833) was the first one to publish an algebraic procedure for the least squares method.

Exercise 2.8. Let $c^\star_{LS}$ be the arithmetic mean of the set $\mathcal{A} = \{a^1, \dots, a^m\} \subset \mathbb{R}$. Show that

$$\sum_{i=1}^{m} (c^\star_{LS} - a^i) = 0\,.$$

Check this property for data in Example 2.7.

Exercise 2.9. Let $\mathcal{A} = \{a^1, \dots, a^p\}$ and $\mathcal{B} = \{b^1, \dots, b^q\} \subset \mathbb{R}$ be disjoint sets, and let $a^\star_{LS}$ and $b^\star_{LS}$ be their arithmetic means. Show that the arithmetic mean of the union $\mathcal{C} = \mathcal{A} \cup \mathcal{B}$ equals

$$c^\star_{LS} = \tfrac{p}{p+q} a^\star_{LS} + \tfrac{q}{p+q} b^\star_{LS}\,.$$

Check the formula in several examples. What would the generalization of this formula for n data sets $\mathcal{A}_1, \dots, \mathcal{A}_n$ containing $p_1, \dots, p_n$ elements respectively, look like?

2.1.2 Best ℓ_1-representative

In case of the ℓ_1 metric, the function (2.1) becomes

$$F_1(x) := \sum_{i=1}^{m} |x - a^i|\,. \tag{2.9}$$

The next lemma shows that if $\mathcal{A}$ is a set of mutually different real numbers, the function F_1 attains its global minimum at the median of $\mathcal{A}$ (see e.g. [25, 38]). The case when some data might be equal will be considered in Section 2.1.3.

Lemma 2.10. *Let $\mathcal{A} = \{a^i \in \mathbb{R} : i = 1, \dots, m\}$ be a set of mutually different data points. The function F_1 given by (2.9) attains its global minimum at the median of the set $\mathcal{A}$.*

Proof. Without loss of generality, we may assume that $a^1 < a^2 < \cdots < a^m$. Note that F_1 is a convex piecewise linear function (see Figure 2.2c) and therefore it can attain its global minimum at a single point in $\mathcal{A}$ or at all points between two points in $\mathcal{A}$.

For $x \in (a^k, a^{k+1})$ we have

$$F_1(x) = \sum_{i=1}^{k} (x - a^i) - \sum_{i=k+1}^{m} (x - a^i) = (2k - m)\, x - \sum_{i=1}^{k} a^i + \sum_{i=k+1}^{m} a^i,$$

$$F_1'(x) = 2k - m.$$

Thus, the function F_1 decreases on intervals (a^k, a^{k+1}) for $k < \frac{m}{2}$, and increases for $k > \frac{m}{2}$.

Therefore we have to consider two cases:

- if m is odd, i. e. $m = 2p+1$, the function F_1 attains its global minimum at the middle datum a^p;
- if m is even, i. e. $m = 2p$, the function F_1 attains its global minimum at every point of the interval $[a^p, a^{p+1}]$.

Hence, a best ℓ_1-representative of the set $\mathcal{A} \subset \mathbb{R}$ is the *median* of $\mathcal{A}$. □

Note that the median of a set $\mathcal{A}$ may be either a set (a segment of real numbers) or a single real number. If the median of $\mathcal{A}$ is a set it will be denoted by $\operatorname{Med}\mathcal{A}$, and its elements by $\operatorname{med}\mathcal{A}$. The number $\operatorname{med}\mathcal{A}$ has the property (cf. Remark 2.3) that the sum of its absolute deviations to all data is minimal, i. e.

$$\sum_{i=1}^{m} |x - a^i| \geq \sum_{i=1}^{m} |\operatorname{med}\mathcal{A} - a^i|\,,$$

and the equality holds if and only if $x = \operatorname{med}\mathcal{A}$.[5]

Example 2.11. Given the data set $\mathcal{A} = \{2, 1.5, 2, 2.5, 5\}$, its median is $\operatorname{med}\mathcal{A} = 2$. What is the sum of absolute deviations?

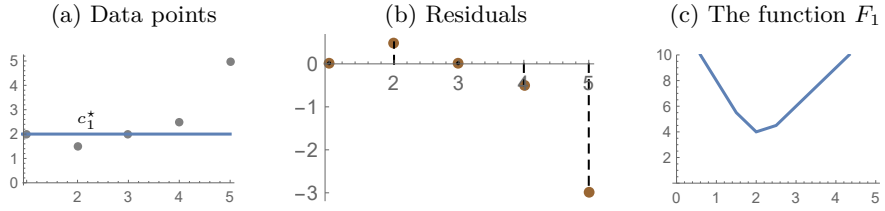


Figure 2.2: Median of the set $\mathcal{A} = \{2, 1.5, 2, 2.5, 5\}$

[5] The problem of finding the best ℓ_1-representative of a data set, appears in the literature as the *least absolute deviations principle*, ascribed to Croatian scholar Josip Ruđer Bošković (1711–1787), who posed it in 1757 in his article [2]. Due to complicated calculations, for a long time this principle was neglected in comparison to the Gauss least squares principle. Not until modern computers came about did this take an important place in scientific research, in particular because of its robustness: in contrast to the Gauss least squares principle, this principle ignores the outliers (strongly jutting data) in data sets. Scientific conferences devoted to ℓ_1 methods and applications are regularly held in Swiss city Neuchâtel, and the front page of the conference proceedings shows a Croatian banknote depicting the portrait of Josip Ruđer Bošković, [4].

Figure 2.2a shows the data and median $c_1^\star$, Figure 2.2b shows the residuals (numbers $c_1^\star - a^i$), and Figure 2.2c depicts the graph of the function F_1. Note that F_1 is a convex piecewise linear function and that $F_1(c_1^\star) = 4$.

How would the median of this set change if data contained an outlier? What would be the median if the datum 5 were replaced by 10, and what if it were replaced by 100?

To find the median of a set $\mathcal{A}$, first one has to sort the elements. Then, if the number of elements is odd, median is the middle element, and if the number of elements is even, median is any number between the two middle elements. For example,[6]

$$\begin{aligned}&\mathrm{Med}\{3,\ 1,\ 4,\ 5,\ 9\} = \{4\},\\ &\mathrm{Med}\{-1,\ 1,\ -2,\ 2,\ -5,\ 5,\ -9,\ 9\} = [-1, 1],\end{aligned}$$

but $\mathrm{med}\{3,\ 1,\ 4,\ 5,\ 9\} = 4$ and $\mathrm{med}\{-1,\ 1,\ -2,\ 2,\ -5,\ 5,\ -9,\ 9\} \in [-1, 1]$.

Remark 2.12. Note that the median of a data set $\mathcal{A}$ can always be chosen among the elements of $\mathcal{A}$ itself. This means that median, as the best ℓ_1-representative of a set, can always be an element of that set, contrary to the case of the arithmetic mean as the best LS-representative. In some applications this fact might be useful.

Note also that a half of elements of the set $\mathcal{A}$ are placed to the left, and the other half to the right of the median of $\mathcal{A}$.

As a measure of dispersion of a data set $\mathcal{A}$ around the median, in statistics literature [23, 24] one uses the *Median of Absolute Deviations from Median* (`MAD`):

$$\mathtt{MAD}\,\mathcal{A} = 1.483 \operatorname*{med}_{i=1,\dots,m} |a^i - \operatorname*{med}_{j=1,\dots,m} a^j|, \tag{2.10}$$

where the constant 1.483 was introduced in [23].

Example 2.13. The relative magnitudes of elements of the set

$$\mathcal{A} = \{9.05,\ 2.83,\ 3.00,\ 3.16,\ 4.12,\ 3.00,\ 3.50\}$$

can be better compared after mapping the set $\mathcal{A}$ to the unit interval $[0, 1]$ using the linear map

$$\varphi(x) = \frac{x - a}{b - a}, \qquad \text{where } a = \min \mathcal{A},\ b = \max \mathcal{A}. \tag{2.11}$$

[6] The median of a set can be obtained using *Mathematica* instruction `Median[]`. If the median of the given set happens to be an interval, the instruction `Median[]` will give the midpoint of that interval.

We get $\varphi(\mathcal{A}) = \{1., 0., 0.027, 0.053, 0.207, 0.027, 0.108\}$, and it is readily seen that $a^1 \in \mathcal{A}$ is by far the largest element in $\mathcal{A}$.

Following [23], this can be ascertained more exactly by first using (2.10) to find MAD=0.489 and define the new set

$$\begin{aligned}\widetilde{\mathcal{A}} &= \{\tilde{a}^i = |a^i - \operatorname*{med}_{j=1,\dots,m} a^j| \,/\, \texttt{MAD} : a^i \in \mathcal{A}\} \\ &= \{12.04,\ 0.67,\ 0.33,\ 0,\ 1.96,\ 0.33,\ 0.69\}.\end{aligned}$$

The element $a^i \in \mathcal{A}$ for which $\tilde{a}^i > 2.5$ is considered, according to [23], to be an outlier. So, in our example, only the element $a^1 = 9.05$ is an outlier in $\mathcal{A}$.

In statistics literature [**?**], median of a set $\mathcal{A}$ is tied to the *first quartile* (the element of $\mathcal{A}$ placed at $1/4$ of the sorted data) and the *third quartile* (the element of $\mathcal{A}$ placed at $3/4$ of the sorted data). What are the first and third quartile of the data set in previous example?

2.1.3 Best representative of weighted data

In practical applications it is sometimes necessary to equip the data with some weights. In this way we associate to each datum its impact or the frequency of occurrence. For example, to find the student's average grade point in the exams he passed, the data set is $\{2, 3, 4, 5\}$ and weights are the frequencies of occurrence of each grade.

As with data without weights, one can prove that the function

$$F_{LS}(x) = \sum_{i=1}^{m} w_i (x - a^i)^2$$

attains its global minimum at the unique point

$$c^\star_{LS} = \operatorname*{arg\,min}_{x\in\mathbb{R}} \sum_{i=1}^{m} w_i \, d_{LS}(x, a^i) = \tfrac{1}{W} \sum_{i=1}^{m} w_i \, a^i, \quad W = \sum_{i=1}^{m} w_i \,,$$

which we call *weighted arithmetic mean* [25].

In case of ℓ_1 metric function, the function (2.1) looks like

$$F_1(x) = \sum_{i=1}^{m} w_i \, |x - a^i| \,, \tag{2.12}$$

and it attains its global minimum at *weighted median* $\operatorname*{Med}_i(w_i, a^i)$ of the set $\mathcal{A}$, as shown by the following lemma.

Lemma 2.14 ([25]). *Let $a^1 < \cdots < a^m$ be a set of data points with weights $w_1, \ldots, w_m > 0$, and let $I = \{1, \ldots, m\}$ be the pertinent index set. Denote*

$$J := \{\nu \in I : \sum_{i=1}^{\nu} w_i \leq \sum_{i=\nu+1}^{m} w_i\},$$

and for $J \neq \emptyset$, denote $\nu_0 = \max J$. Then:

(i) If $J = \emptyset$, (i. e. $w_1 > \sum_{i=2}^{m} w_i$), then the minimum of F_1 is attained at the point $\alpha^\star = a^1$.

(ii) If $J \neq \emptyset$ and $\sum_{i=1}^{\nu_0} w_i < \sum_{i=\nu_0+1}^{m} w_i$, then the minimum of F_1 is attained at the point $\alpha^\star = a^{\nu_0+1}$.

(iii) If $J \neq \emptyset$ and $\sum_{i=1}^{\nu_0} w_i = \sum_{i=\nu_0+1}^{m} w_i$, then the minimum of F_1 is attained at every point $\alpha^\star$ in the segment $[a^{\nu_0}, a^{\nu_0+1}]$.

Proof. Notice that on each interval

$$(-\infty, a^1),\ [a^1, a^2), \ldots, [a^{m-1}, a^m),\ [a^m, \infty)$$

F is a linear function with slopes of these linear functions being consecutively d_ν, $\nu = 0, \ldots, m$, where

$$\begin{aligned} d_0 &= -\sum_{i=1}^{m} w_i\,, \\ d_\nu &= \sum_{i=1}^{\nu} w_i - \sum_{i=\nu+1}^{m} w_i = d_{\nu-1} + w_\nu + w_{\nu+1}\,, \quad \nu = 1, \ldots, m-1\,, \\ d_m &= \sum_{i=1}^{m} w_i\,. \end{aligned}$$

If $J = \emptyset$, then $2\sum_{i=1}^{\nu} w_i - \sum_{i=1}^{m} w_i > 0$ for every $\nu = 1, \ldots, m$, and $d_0 < 0 < d_\nu$, $\nu = 1, \ldots, m$. It follows that the function F_1 is strongly decreasing on $(-\infty, a^1)$ and strongly increasing on $(a^1, +\infty)$. Therefore the minimum of F_1 is attained for $\alpha^\star = a^1$.

If $J \neq \emptyset$, note that $\nu_0 = \max\{\nu \in I : d_\nu \leq 0\}$. Since $d_{\nu+1} - d_\nu = 2w_{\nu+1} > 0$, $d_0 < 0$, and $d_m > 0$, the sequence (d_ν) is increasing and satisfies

$$d_0 < d_1 \ldots < d_{\nu_0} \leq 0 < d_{\nu_0+1} < \ldots < d_m. \tag{2.13}$$

If $d_{\nu_0} < 0$, i. e. $2\sum_{i=1}^{\nu_0} \omega_i < \sum_{i=1}^{m} \omega_i$, from (2.13) it follows that F_1 is strongly decreasing on $(-\infty, a^{\nu_0+1})$ and strongly increasing on $(a^{\nu_0+1}, +\infty)$. Therefore the minimum of F_1 is attained for $\alpha^\star = a^{\nu_0+1}$.

If $d_{\nu_0} = 0$, i. e. $2\sum_{i=1}^{\nu_0} w_i = \sum_{i=1}^{m} w_i$, from (2.13) it follows that F_1 is strongly decreasing on $(-\infty, a^{\nu_0})$, it is constant on $[a^{\nu_0}, a^{\nu_0+1}]$, and strongly increasing on $(a^{\nu_0+1}, +\infty)$. Therefore the minimum of F_1 is attained at every point $\alpha^\star$ in the segment $[a^{\nu_0}, a^{\nu_0+1}]$. □

Hence, a best ℓ_1-representative of a weighted data set is the *weighted median* $\operatorname{Med}_i(w_i, a^i)$. Note that weighted median can also be a set (a segment of real numbers) or a single real number. Weighted median $\operatorname{med}_i(w_i, a^i)$ is any number with the property that the sum of weighted absolute deviations to all data is minimal, i. e.

$$\sum_{i=1}^{m} w_i \, |x - a^i| \geq \sum_{i=1}^{m} w_i \, |\operatorname{med}_j(w_j, a^j) - a^i| \,,$$

and the equality holds for $x = \operatorname{med}_j(w_j, a^j)$ (see also Remark 2.3).

The next corollary shows that Lemma 2.10 is just a special case of Lemma 2.14.

Corollary 2.15. *Let $a^1 \leq a^2 \leq \cdots \leq a^m$, $m > 1$, be a set of data points with weights $w_1 = \cdots = w_m = 1$. Then:*

(i) if m is odd ($m = 2k + 1$), then the minimum of the function F_1 is attained at the point $\alpha^\star = a^{k+1}$;

(ii) if m is even ($m = 2k$), the minimum of the function F_1 is attained at every point $\alpha^\star$ of the segment $[a^k, a^{k+1}]$.

Proof. First, note that in this case the set J from Lemma 2.14 is always nonempty.

Let $m = 2k + 1$. According to Lemma 2.14 (*ii*),

$$\nu_0 = \max\{\nu \in I : 2\nu - m \leq 0\} = \max\{\nu \in I : \nu \leq k + \tfrac{1}{2}\} = k,$$
$$d_{\nu_0} = d_k = 2k - m = 2k - 2k - 1 < 0,$$

and therefore $\alpha^\star = a^{k+1}$.

Let $m = 2k$. According to Lemma 2.14 (*iii*),

$$\nu_0 = \max\{\nu \in I : 2\nu - m \leq 0\} = \max\{\nu \in I : \nu = k\} = k,$$
$$d_{\nu_0} = d_k = 2k - m = 2k - 2k = 0\,.$$

It follows that the minimum of the function F_1 is attained at every point $\alpha^\star$ of the segment $[a^k, a^{k+1}]$. □

In general, determining weighted median is a very complicated numerical procedure [9, 25]. For this purpose there are numerous algorithms in the literature [9].

Example 2.16. Weighted median of a set $\mathcal{A} \subset \mathbb{R}$ with weights being positive integers, can be determined similarly as for median of a set without weights. First we sort the elements of the set $\mathcal{A}$. Next we form the *multiset* where each element of $\mathcal{A}$ is repeated according to its weight, and then we take the middle element of that multiset. A set $\mathcal{A} = \{a^1, \ldots, a^m\}$ with weights $w := \{w^1, \ldots, w^m\}$ will be called a *weighted set*, and its median will be denoted by $\operatorname{med}_{\mathrm{w}} \mathcal{A}$. For example, the weighted median of the set $\mathcal{A} = \{3, 1, 4, 5, 9\}$ with weights $\mathrm{w} = \{3, 1, 3, 2, 2\}$ is the middle element of the multiset (here written as a finite sequence)

$$1, 3, 3, 3, 4, 4, 4, 5, 5, 9, 9.$$

In our example, the weighted median of weighted set $\mathcal{A}$ with weights w is $\operatorname{med}_{\mathrm{w}} \mathcal{A} = 4$. What is the first and third quartile of the weighted set $\mathcal{A}$?

2.1.4 Bregman divergences

Let us consider yet another class of distance-like functions which is important in applications. Following [17, 37] we introduce the following definition:

Definition 2.17. Let $D \subseteq \mathbb{R}$ be a convex set (i. e. an interval) and let $\phi\colon D \to \mathbb{R}_+$ be a strictly convex continuously differentiable function on $\operatorname{int} D \neq \emptyset$. The function $d_\phi\colon D \times \operatorname{int} D \to \mathbb{R}_+$ defined by

$$d_\phi(x, y) = \phi(x) - \phi(y) - \phi'(y)(x - y) \tag{2.14}$$

is called the ***Bregman divergence***.

It is not difficult to see that such a function is indeed a distance-like function. Geometrically, for a given $x \in D$, $d_\phi(x, y)$ represents the difference between the value $\phi(x)$ and the value of the linear function, whose graph is the tangent line at the point $(y, \phi(y))$, $y \in D$, at the point x (see Figure 2.3). In the past ten years, distance-like functions of this kind have been intensely investigated and applied in operations research, information theory, nonlinear analysis, machine learning, wireless sensor network, etc. (see e.g. [**? ?**]).

Exercise 2.18. Show that for $\phi\colon \mathbb{R} \to \mathbb{R}_+$, $\phi(x) := x^2$, the Bregman divergence becomes the LS distance-like function.

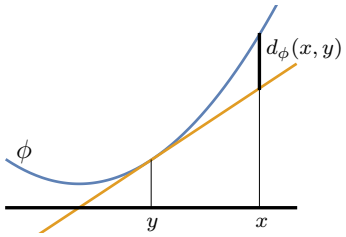


Figure 2.3: Geometric meaning of Bregman divergence

Example 2.19. Bregman divergence for $\phi\colon \mathbb{R}_{++} \to \mathbb{R}_+$, $\phi(x) = -\ln x$, is known as the *Itakura-Saito divergence* given by

$$d_{IS}(x,y) = \frac{x}{y} - \ln\frac{x}{y} - 1. \tag{2.15}$$

Let us find a corresponding best representative of the set $\mathcal{A} = \{a^i \in \mathbb{R}_{++} : i = 1,\dots,m\}$. In this case the function (2.1) becomes

$$F(x) = \sum_{i=1}^m d_{IS}(x,a^i) = \sum_{i=1}^m \Big(\frac{x}{a^i} - \ln\frac{x}{a^i} - 1\Big).$$

Since $F'(x) = \sum\limits_{i=1}^m \frac{1}{a^i} - \frac{m}{x}$, the point $c_{IS}^\star = m\big(\sum\limits_{i=1}^m \frac{1}{a^i}\big)^{-1}$ is the unique stationary point, and since $F''(x) = \frac{m}{x^2} > 0$, F is a convex function and $c_{IS}^\star$ is its only point of global minimum. Note that $c_{IS}^\star$ is the harmonic mean of the set $\mathcal{A}$.

Example 2.20. With the convention $0 \cdot \ln 0 = 0$, Bregman divergence for $\phi\colon \mathbb{R}_+ \to \mathbb{R}_+$, $\phi(x) = x\ln x$, is known as the *Kullback-Leibler divergence* given by

$$d_\phi(x,y) = x\ln\frac{x}{y} - x + y. \tag{2.16}$$

Let us find a corresponding best representative of the set $\mathcal{A} = \{a^i \in \mathbb{R}_+ : i = 1,\dots,m\}$. In this case the function (2.1) becomes

$$F(x) = \sum_{i=1}^m d_{KL} \sum_{i=1}^m d_{KL}(x,a^i) = \frac{x}{a^i} - \ln\frac{x}{a^i} - 1\,.$$

Since $F'(x) = \sum\limits_{i=1}^m \ln\frac{x}{a^i}$, the point $c_{KL}^\star = \sqrt[m]{\prod\limits_{i=1}^m a^i}$ is the unique stationary point, and since $F''(x) = \frac{m}{x} > 0$ for all $x \in \mathbb{R}_{++}$, F is a convex function, and the point $c_{KL}^\star$ is its only point of global minimum. Note that $c_{KL}^\star$ is the geometric mean of the set $\mathcal{A}$.

Remark 2.21. The above distance-like functions can be generalized to data sets with $n \geq 1$ features (see e.g. [17, 37]).

2.2 Representative of data sets with two features

A set with two features without weights is usually interpreted as a finite set $\mathcal{A} = \{a^i = (x_i, y_i) : i = 1, \ldots, m\} \subset \mathbb{R}^2$, and geometrically it can be visualized as a finite set of points in the plane. In the next section we give a short historical overview of looking for a best representative of a data set with two features, and possible applications.

2.2.1 Fermat–Torricelli–Weber problem

Let $A, B, C \in \mathbb{R}^2$ be three non-collinear points in the plane (see Figure 2.4).

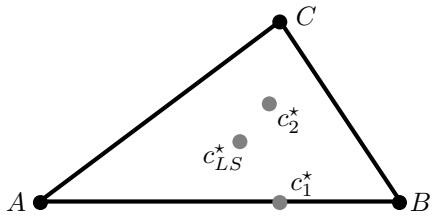


Figure 2.4: Fermat problem

The *Fermat's problem* consists in finding the point $c_2^\star \in \mathbb{R}^2$ with the property that the sum of its Euclidean, i. e. ℓ_2-distances to the vertices of the triangle $\triangle ABC$ is minimal. The point $c_2^\star$ is called the *geometric median* of the points A, B, C, and can be obtained (see e.g. [19]) as the intersection of the so-called Simpson's lines (see Figure 2.5a), or as the intersection of the so-called Torricelli's circles (see Figure 2.5b). The same problem can also be treated for a different distance-like functions: in the sense of physics — the *Torricelli's problem*, and in the sense of econometrics — the *Weber's problem* (see e.g. [6]).

The point $c_{LS}^\star \in \mathbb{R}^2$ (see Figure 2.4), with the property that the sum of its LS-distances (i. e. the sum of squared Euclidean distances) to vertices of the triangle $\triangle ABC$, is minimal, is called the *centroid* or the *Steiner point* (this is related to the centre of mass in physics), and is obtained as the intersection of medians of the triangle, i. e. line segments joining vertices to the midpoints of opposite sides.

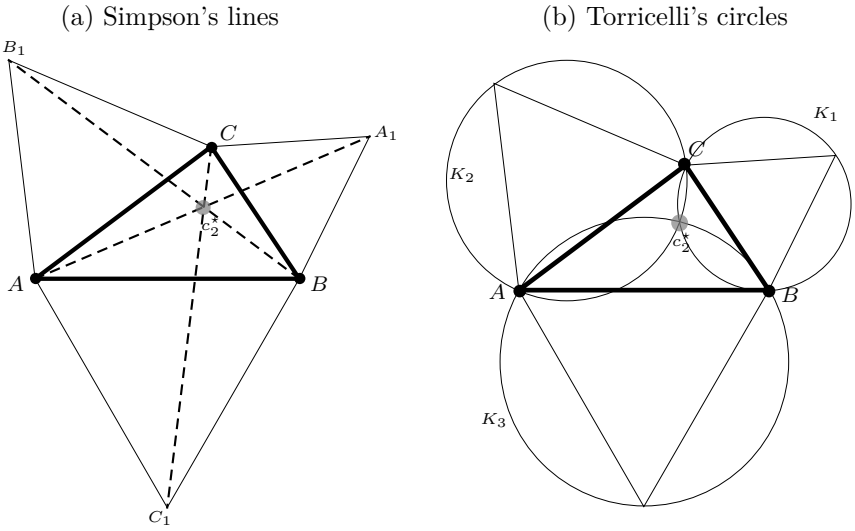


Figure 2.5: Fermat's problem

The point $c_1^\star \in \mathbb{R}^2$ (see Figure 2.4), with the property that the sum of its ℓ_1-distances to the vertices of the triangle $\triangle ABC$ is minimal, is called the *median* of the set $\{A, B, C\}$.

In general, one can consider a finite set of points in $\mathbb{R}^n$ and an arbitrary distance-like function d. The problem of finding best d-representative has many applications in various fields: telecommunication (optimal antenna coverage problem, discrete network location), public sector (optimal covering problem), economy (optimal location of consumer centers), hub location problems, robotics, optimal assignation problems, hourly forecast of natural gas consumption problem, etc. [6, 20, 28].

Exercise 2.22. Given the triangle $\triangle ABC$ with vertices $A = (0, 0)$, $B = (6, 0)$, and $C = (4, 3)$, find the vertices A_1, B_1, C_1 of the equilateral triangles constructed on the sides of triangle $\triangle ABC$, and find the intersection of line segments joining the points A–A_1, B–B_1, and C–C_1 as in Figure 2.5a.

Solution: $A_1 = (7.598, 3.232)$, $B_1 = (-0.598, 4.964)$, $C_1 = (3., -5.196)$;
Geometric median: $c_2^\star = (3.833, 1.630)$.

Exercise 2.23. Given the triangle $\triangle ABC$ with vertices $A = (0, 0)$, $B = (6, 0)$, and $C = (4, 3)$, find the vertices A_1, B_1, C_1 of the equilateral triangles constructed on the sides of triangle $\triangle ABC$, construct the circumcircles of these triangles, and find the intersection of these circles as in Figure 2.5b.

Solution: $A_1 = (7.598,\ 3.232)$, $B_1 = (-0.598,\ 4.964)$, $C_1 = (3.,\ -5.196)$;
$K_1 = ((5.866,\ 2.077), 2.082)$; $K_2 = ((1.134,\ 2.655), 2.887)$; $K_3 = ((3.,\ -1.732), 3.464)$;
Geometric median: $c_2^\star = (3.833, 1.630)$.

2.2.2 Centroid of a set in the plane

Let $\mathcal{A} = \{a^i = (x_i, y_i) \colon i = 1, \ldots, m\} \subset \mathbb{R}^2$ be a set without weights in the plane. The *centroid* $c_{LS}^\star$ of the set $\mathcal{A}$ is the solution to the optimization problem

$$\arg\min_{c \in \mathbb{R}^2} \sum_{i=1}^{m} d_{LS}(c, a^i)\,, \tag{2.17}$$

where $d_{LS}(a, b) = d_2^2(a, b) = \|a - b\|^2$. The point $c_{LS}^\star$ is the point at which the function

$$F_{LS}(x, y) = \sum_{i=1}^{m} \|c - a^i\|^2 = \sum_{i=1}^{m} \left((x - x_i)^2 + (y - y_i)^2\right), \quad c = (x, y).$$

attains its global minimum. $F_{LS}(x, y)$ is the sum of squared Euclidian, i. e. ℓ_2-distances from the point $c = (x, y)$ to the points $a^i \in \mathcal{A}$. From (2.7) it follows that

$$F_{LS}(x, y) = \sum_{i=1}^{m} \left((x - x_i)^2 + (y - y_i)^2\right) \geq \sum_{i=1}^{m} (\bar{x} - x_i)^2 + \sum_{i=1}^{m} (\bar{y} - y_i)^2, \tag{2.18}$$

where

$$\bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i\,, \qquad \bar{y} = \frac{1}{m} \sum_{i=1}^{m} y_i\,,$$

and the equality holds if and only if $x = \bar{x}$ and $y = \bar{y}$. Therefore, the solution to the global optimization problem (2.17) is the centroid of the set $\mathcal{A}$, i. e. the point $c_{LS}^\star = (\bar{x}, \bar{y})$.

Hence, the centroid of a finite set $\mathcal{A}$ of points in the plane is the point whose first and second coordinates are the arithmetic means of the first and second coordinates of points in $\mathcal{A}$, respectively.

Example 2.24. Check that for the given points $a^1 = (0, 0)$, $a^2 = (6, 0)$, and $a^3 = (4, 3)$, the centroid of the set $\{a^1, a^2, a^3\}$ is the point $c_{LS}^\star = (\frac{10}{3}, 1)$.

2.2.3 Median of a set in the plane

Median of a set of points $\mathcal{A} = \{a^i = (x_i, y_i) : i = 1, \dots, m\} \subset \mathbb{R}^2$ without weights, is a solution to the optimization problem

$$\arg\min_{c \in \mathbb{R}^2} \sum_{i=1}^{m} d_1(c, a^i)\,. \tag{2.19}$$

This is every point at which the function

$$F_1(x, y) = \sum_{i=1}^{m} \|c - a^i\|_1 = \sum_{i=1}^{m} (|x - x_i| + |y - y_i|), \quad c = (x, y),$$

attains the global minimum. $F_1(x, y)$ is the sum of ℓ_1-distances from $c = (x, y)$ to the points $a^i \in \mathcal{A}$. From (2.9) it follows that

$$F_1(x, y) = \sum_{i=1}^{m} (|x - x_i| + |y - y_i|) \geq \sum_{i=1}^{m} |\operatorname*{med}_k x_k - x_i| + \sum_{i=1}^{m} |\operatorname*{med}_k y_k - y_i|\,, \tag{2.20}$$

and the equality holds if and only if $x = \operatorname*{med}_k x_k$ i $y = \operatorname*{med}_k y_k$. Therefore, the solution to the global optimization problem(2.19) is a median of the set $\mathcal{A}$, which is a point

$$(\operatorname*{med}_k x_k, \operatorname*{med}_k y_k). \tag{2.21}$$

Hence, the median of a finite set $\mathcal{A}$ of points in the plane is any point whose first and second coordinates are medians of the first and second coordinates of points in $\mathcal{A}$, respectively.

Example 2.25. Check that for the three points $A_1 = (0, 0)$, $A_2 = (6, 0)$, and $A_3 = (4, 3)$, median of the set $\{A_1, A_2, A_3\}$ is the point $c_1^\star = (4, 0)$.

Example 2.26. Median of the set $\mathcal{A} = \{(1, 1), (1, 3), (2, 2), (3, 1), (3, 4), (4, 3)\} \subset \mathbb{R}^2$ is any point in the square $[2, 3] \times [2, 3]$ (see Figure 2.6), since median of the first coordinates of the data is $\operatorname{med}\{1, 1, 2, 3, 3, 4\} \in [2, 3]$, and median od the second coordinates is $\operatorname{med}\{1, 3, 2, 1, 4, 3\} \in [2, 3]$.

Exercise 2.27. Change the position of just one point of the set $\mathcal{A}$ in previous example in such a way that median becomes a single point, a segment or a rectangle.

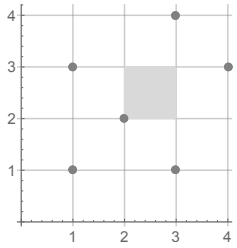


Figure 2.6: Median of the set $\mathcal{A} = \{(1,1),(1,3),(2,2),(3,1),(3,4),(4,3)\}$

2.2.4 Geometric median of a set in the plane

Geometric median $c^\star$ of the set $\mathcal{A} = \{a^i = (x_i, y_i) \colon i = 1, \ldots, m\} \subset \mathbb{R}^2$ without weights is the solution to the global optimization problem

$$c^\star = \arg\min_{c \in \mathbb{R}^2} \sum_{i=1}^{m} d_2(c, a^i). \tag{2.22}$$

The point $c^\star$ is the point at which the function

$$F_2(x,y) = \sum_{i=1}^{m} \|c - a^i\| = \sum_{i=1}^{m} \sqrt{(x-x_i)^2 + (y-y_i)^2}, \quad c = (x,y) \tag{2.23}$$

attains the global minimum. $F_2(x,y)$ is the sum of ℓ_2-distances between the point $c = (x,y) \in \mathbb{R}^2$ and points $a^i \in \mathcal{A}$, and in this case the variables x and y cannot be separated. Therefore the solution to the global optimization problem (2.22) cannot be written down explicitly.

Example 2.28. In order to find the geometric median of the set of three points $a^1 = (0,0)$, $a^2 = (6,0)$, and $a^3 = (4,3)$, one has to solve the following optimization problem:

$$\arg\min_{(x,y) \in \mathbb{R}^2} F_2(x,y),$$
$$F_2(x,y) = \sqrt{x^2 + y^2} + \sqrt{(x-6)^2 + y^2} + \sqrt{(x-4)^2 + (y-3)^2}\,.$$

Using *Mathematica* computation system, we can solve this optimization problem like this: first define the function

```
In[1]:= F2[x_, y_]:= Sqrt[x^2 + y^2] + Sqrt[(x-6)^2 + y^2]
                     + Sqrt[(x-4)^2 + (y-3)^2]
```

We can try to solve our problem as a global optimization problem using the *Mathematica*-module

```
In[2]:= NMinimize[F2[x, y], {x, y}]
```

According to [40], the module `NMinimize[]` can sometime find only a local minimum. In such a case we can try to solve the problem as a local optimization problem by using the *Mathematica*-module

```
In[2]:= FindMinimum[F2[x, y], {x, 1}, {y, 2}]
```

using some good initial approximation close to the solution. In our case we obtain $c_2^\star = (3.833,\ 1.630)$.

Remark 2.29. The best known algorithm for searching for the geometric median by solving the optimization problem (2.22) is the *Weiszfeld algorithm* (see [12, 33**?**]). This is an iterative procedure which arose as a special case of simple-iteration method for solving systems of nonlinear equations (see e.g. [**? ? ?**]).

First we find partial derivatives of the function (2.23), and make them equal to zero:

$$\frac{\partial F_2}{\partial x} = \sum_{i=1}^{m} \frac{x - x_i}{\|c - a^i\|} = x \sum_{i=1}^{m} \frac{1}{\|c - a^i\|} - \sum_{i=1}^{m} \frac{x_i}{\|c - a^i\|} = 0,$$
$$\frac{\partial F_2}{\partial y} = \sum_{i=1}^{m} \frac{y - y_i}{\|c - a^i\|} = y \sum_{i=1}^{m} \frac{1}{\|c - a^i\|} - \sum_{i=1}^{m} \frac{y_i}{\|c - a^i\|} = 0,$$

which can be written as

$$x = \Phi(x, y), \quad y = \Psi(x, y), \tag{2.24}$$

where

$$\Phi(x, y) = \frac{\sum\limits_{i=1}^{m} \frac{x_i}{\|c-a^i\|}}{\sum\limits_{i=1}^{m} \frac{1}{\|c-a^i\|}}, \quad \Psi(x, y) = \frac{\sum\limits_{i=1}^{m} \frac{y_i}{\|c-a^i\|}}{\sum\limits_{i=1}^{m} \frac{1}{\|c-a^i\|}}. \tag{2.25}$$

After choosing an initial approximation (x_0, y_0) from the convex hull of the set $\mathcal{A}$, $(x_0, y_0) \in \mathrm{conv}(\mathcal{A})$, the system (2.24) can be solved by successive iteration method

$$x_{k+1} = \Phi(x_k, y_k), \quad y_{k+1} = \Psi(x_k, y_k), \quad k = 0, 1, \ldots \tag{2.26}$$

Example 2.30. Let the data set $\mathcal{A} \subset \mathbb{R}^2$ be defined like this:

```
In[1]:= SeedRandom[13]
        sig = 1.5; m = 50; cen = {4,5};
        podT = Table[cen + RandomReal[NormalDistribution[0, sig], {2}],
```

```
                {i, m}];
podW = RandomReal[{0,1}, m];
Show[Table[ListPlot[{podT[[i]]},
     PlotStyle -> {PointSize[podW[[i]]/20], Gray}], {i,m}],
      PlotRange -> {{0,8},{0,8}}, AspectRatio -> Automatic]
```

Each datum in Figure 2.7 is equipped with weight according to the point-size (small disc) representing the datum. Check that the centroid is the point $c^\star_{LS} = (4.151,\ 4.676)$, median $c^\star_1 = (4.350,\ 4.750)$, and the geometric median $c^\star_2 = (4.251,\ 4.656)$.

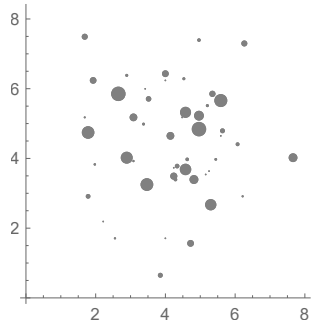


Figure 2.7: Set of weighted data $\mathcal{A}$

Exercise 2.31. Let the set $\mathcal{A} = \{(x_i, y_i) \in \mathbb{R}^2 : i = 1, \ldots, 10\}$, be given by the table

i	1	2	3	4	5	6	7	8	9	10
x_i	9	6	8	1	1	4	4	3	9	10
y_i	5	5	5	2	5	8	1	8	8	4

Depict the set $\mathcal{A}$ in the coordinate plane and find its centroid, median, and geometric median.
Hint: Use the following *Mathematica*-program

```
In[1]:= SeedRandom[2]
        A = RandomInteger[{1, 10}, {10, 2}]
            ListPlot[A, ImageSize -> Small]
                Print["Centroid = ", Mean[A]]
                Print["Median = ", Median[A]]
        Psi[x_, y_] := Sum[Norm[{x, y} - A[[i]]], {i, Length[A]}]
                Print["Geometric median:"]
                      NMinimize[Psi[x, y], {x, y}]
```

Solution: $c^\star_{LS} = (5.5,\ 5.1)$, $c^\star_1 = (5, 5)$, $c^\star_2 = (6, 5)$.

2.3 Representative of data sets with several features

In practical applications data can have more then one or two features as was mentioned at the beginning of the Introduction, where several such examples were listed. Since the number of features represents the dimension of the data, it will be necessary to find representatives also for data of arbitrary high dimension.

We want to find a point in $\mathbb{R}^n$ which represents, as good as possible, a given set of points $\mathcal{A} = \{a^i = (a^i_1, \dots, a^i_n) \in \mathbb{R}^n : i = 1, \dots, m\}$ without weights.

In case of LS distance-like function, the best representative of the set $\mathcal{A}$ is its *centroid* (*barycenter*)[7]

$$c^\star_{LS} = \operatorname*{arg\,min}_{c\in\mathbb{R}^n} \sum_{i=1}^{m} d_{LS}(c, a^i) = \operatorname*{arg\,min}_{c\in\mathbb{R}^n} \sum_{i=1}^{m} \|c - a^i\|^2 = \tfrac{1}{m} \sum_{i=1}^{m} a^i,$$

and the corresponding minimizing function is

$$F_{LS}(c) = \sum_{i=1}^{m} \|c - a^i\|^2 .$$

In case of ℓ_1 metric function, a best representative of the set $\mathcal{A}$ is its *median*

$$c_1 = \operatorname*{med}_i a^i = (\operatorname*{med}_i a^i_1, \dots, \operatorname*{med}_i a^i_n) \in \operatorname{Med} \mathcal{A} = \operatorname*{arg\,min}_{c\in\mathbb{R}^n} \sum_{i=1}^{m} \|c - a^i\|_1,$$

and the corresponding minimizing function is

$$F_1(c) = \sum_{i=1}^{m} \|c - a^i\|_1.$$

2.3.1 Representative of weighted data

Let $\mathcal{A}$ be the set of data points with weights $w_1, \dots, w_m > 0$. If d is the LS distance-like function, the best representative of the set $\mathcal{A}$ with weights $w_1, \dots, w_m > 0$ is its *weighted centroid* (*barycenter*)

$$c^\star_{LS} = \operatorname*{arg\,min}_{c\in\mathbb{R}^n} \sum_{i=1}^{m} w_i\, d_{LS}(c, a^i) = \operatorname*{arg\,min}_{c\in\mathbb{R}^n} \sum_{i=1}^{m} w_i \|c - a^i\|^2 = \tfrac{1}{W} \sum_{i=1}^{m} w_i\, a^i,$$

[7] Recall that $\|\ \|$ denotes the Euclidean, i. e. ℓ_2-norm.

i. e.

$$c^\star_{LS} = \Big(\tfrac{1}{W}\sum_{i=1}^{m} w_i\, a^i_1, \ldots, \tfrac{1}{W}\sum_{i=1}^{m} w_i\, a^i_n\Big) \qquad \text{[coordinate-wise]}, \tag{2.27}$$

where $W = \sum_{i=1}^{m} w_i$, and the corresponding minimizing function is

$$F_{LS}(c) = \sum_{i=1}^{m} w_i \,\|c - a^i\|^2 \,. \tag{2.28}$$

If d is the ℓ_1 metric function, a best representative of the set $\mathcal{A}$ with weights $w_1, \ldots, w_m > 0$ is its *weighted median*

$$\begin{aligned} c^\star_1 &= \operatorname*{med}_i(w_i, a^i) = \big(\operatorname*{med}_i(w_i, a^i_1), \ldots, \operatorname*{med}_i(w_i, a^i_n)\big) \in \operatorname{Med}\mathcal{A} \\ &= \operatorname*{arg\,min}_{c\in\mathbb{R}^n} \sum_{i=1}^{m} w_i \,\|c - a^i\|_1, \end{aligned} \tag{2.29}$$

and the corresponding minimizing function is

$$F_1(c) = \sum_{i=1}^{m} w_i \,\|c - a^i\|_1. \tag{2.30}$$

Namely,

$$\begin{aligned} F_1(c) &= \sum_{i=1}^{m} w_i \,\|c - a^i\|_1 = \sum_{i=1}^{m} w_i \Big(\sum_{k=1}^{n} |c_k - a^i_k|\Big) \\ &= \sum_{k=1}^{n} \Big(\sum_{i=1}^{m} w_i \,|c_k - a^i_k|\Big) = \sum_{k=1}^{n}\sum_{i=1}^{m} w_i \,|c_k - a^i_k| \\ &\geq \sum_{k=1}^{n}\sum_{i=1}^{m} w_i \,|\operatorname*{med}_j(w_j, a^j_k) - a^i_k| = \sum_{i=1}^{m}\sum_{k=1}^{n} w_i \,|\operatorname*{med}_j(w_j, a^j_k) - a^i_k| \\ &= \sum_{i=1}^{m} w_i \,\|c^\star_1 - a^i\|_1 = F(c^\star_1), \end{aligned}$$

where $\operatorname*{med}_j(w_j, a^j_k)$ is the weighted median of data $\{a^1_k, \ldots, a^m_k\}$ with weights $w_1, \ldots, w_m > 0$.

Exercise 2.32. Show, similarly as was shown for the weighted median, that the function F_{LS} given by (2.28), attains its global minimum at the weighted centroid $c^\star_{LS}$ given by (2.27).

2.4 Representative of periodic data

Often it is the case that one has to find best representative of a data set describing events which are periodic in nature, and this is indeed frequently discussed in the literature. For instance, air temperature at certain measuring point during a year, water-level of a river at certain measuring place, seismic activities in specific area over some extended period of time, the illuminance i. e. the measure of the amount of light during a day, etc., are examples of such phenomena. Mathematically speaking, one has to deal with data sets on a circle. Namely, if we represent such a data set on the real line, as we did before, then the data corresponding to the beginning and the end of the same year, for example, would appear far away from each other, although they belong to the same year period. One has to define a distance-like function for such data sets also, and find the center of such data.

Example 2.33. Let $t_i \in \mathcal{A}$ represent the position of the small clock handle on a clock with 12 marks (see Figure 2.8a). Distances in $\mathcal{A}$ will be measured as the *elapsed time* from the moment t_1 to t_2:

$$d(t_1, t_2) = \begin{cases} t_2 - t_1, & \text{if } t_1 \leq t_2 \\ 12 + (t_2 - t_1), & \text{if } t_1 > t_2 \end{cases}.$$

For example, $d(2, 7) = 5$, but $d(7, 2) = 12 + (-5) = 7$. Note that this function is not symmetric.

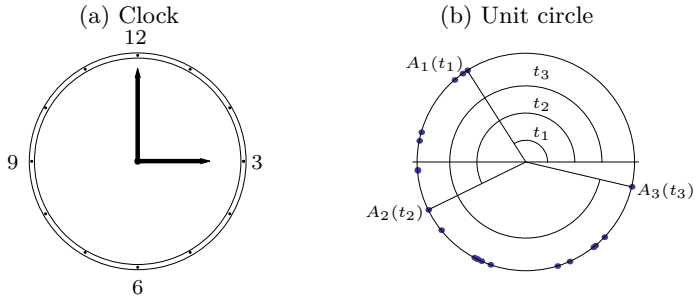


Figure 2.8: Data set on a circle

Example 2.34. Let $t_i \in \mathcal{A}$ represent the position of the small clock handle on a clock with 12 marks (see Figure 2.8a). Define the function measuring the distances on $\mathcal{A}$ as the *length of the time interval* from moment t_1 to t_2:

$$d(t_1, t_2) = \begin{cases} |t_2 - t_1|, & \text{if } |t_2 - t_1| \leq 6 \\ 12 - |t_2 - t_1|, & \text{if } |t_2 - t_1| > 6 \end{cases}.$$

For example, $d(2,9) = 12 - 7 = 5$ and $d(2,7) = 7 - 2 = 5$. Check whether this defines a metric function on the set $\mathcal{A}$.

2.4.1 Representative of data on the unit circle

In general, let (T_i, w_i), $i = 1, \ldots, m$, be a data set where T_i denotes the moment over $M \geq 1$ successive years during which the event we investigate occurred, and let $w_i > 0$ denote the intensity of the event at the moment T_i. The time moments T_i can denote days (for example for water level of a river at some point), hours (air temperatures at some place), or seconds (earthquake moments). We want to identify the moment at which this event is most notable. See [5, 18] for various aspects and applications of such data.

If the moments $T_1, \ldots, T_m$ were considered as simple time series, then would the data from, say the beginning of a year and the end of the same year, be far apart, although they belong to the same season, i. e. time of the year. Therefore, to each year we allot an interval of length 2π, and to a sequence of M successive years the interval $[0, 2\pi M]$. In this way the sequence $T_1, \ldots, T_m$ is transformed into the sequence $T'_1, \ldots, T'_m \in [0, 2\pi M]$.

In our discussion, important is only the moment of the year, and not the particular year in which the event occurred. Therefore, instead of the sequence (T'_i) we define a new sequence $t_i \in [0, 2\pi]$, $i = 1, \ldots, m$, where

$$t_i = 2\,\pi\,T'_i \,(\mathrm{mod}\, 2\,\pi), \quad i = 1, \ldots, m, \tag{2.31}$$

(the remainder of dividing $2\,\pi\,T'_i$ by $2\,\pi$). The number $t_i \in [0, 2\pi]$ represents the moment which is $t_i/2\pi$-th part of a year apart from January 1.

Using the sequence (2.31) we define the following data set:

$$\mathcal{A} = \{a(t_i) = (\cos t_i,\, \sin t_i) \in \mathbb{R}^2 : t_i \in [0, 2\pi],\ i = 1, \ldots, m\} \subset K\,, \tag{2.32}$$

where $K = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$ is the unit circle.

In the following lemma we define a metric on the unit circle and prove its basic properties (see also [14, 18])

Lemma 2.35. *Let $K = \{a(t) = (\cos t, \sin t) \in \mathbb{R}^2 : t \in [0, 2\pi]\}$ be the unit circle in the plane. The function $d_K \colon K \times K \to \mathbb{R}_+$ defined by*

$$d_K(a(t_1), a(t_2)) = \begin{cases} |t_1 - t_2|, & \text{if } \ |t_1 - t_2| \leq \pi\,, \\ 2\pi - |t_1 - t_2|, & \text{if } \ |t_1 - t_2| > \pi\,, \end{cases} \tag{2.33}$$

is a metric on K, and can equivalently be defined as

$$d_K(a(t_1), a(t_2)) = \pi - \big|\, |t_1 - t_2| - \pi \,\big|, \quad t_1, t_2 \in [0, 2\pi]\,. \tag{2.34}$$

Proof. It is straightforward to see that (2.33) and (2.34) are equivalent definitions of the function d_K. Let us show that d_K is a metric on K.

First we show that $d_K(a(t_1), a(t_2)) \geq 0$ for all $t_1, t_2 \in [0, 2\pi]$. Let $t_1, t_2 \in [0, 2\pi]$. Then

$$0 \leq |t_1 - t_2| \leq 2\pi \;\Rightarrow\; -\pi \leq |t_1 - t_2| - \pi \leq \pi \;\Rightarrow\; \big|\, |t_1 - t_2| - \pi \,\big| \leq \pi$$
$$\Rightarrow\; d_K(a(t_1), a(t_2)) = \pi - \big|\, |t_1 - t_2| - \pi \,\big| \geq 0\,.$$

Next we show that $d_K(a(t_1), a(t_2)) = 0$ if and only if $a(t_1) = a(t_2)$:
If $a(t_1) = a(t_2)$ then either $t_1 = t_2$ or $|t_1 - t_2| = 2\pi$. In both cases $d_K(a(t_1), a(t_2)) = 0$.
Conversely, if $d_K(a(t_1), a(t_2)) = 0$, then

$$\pi = \big|\, |t_1 - t_2| - \pi \,\big|\,. \tag{2.35}$$

If $|t_1 - t_2| \leq \pi$, then from (2.35) it follows that $\pi = \pi - |t_1 - t_2|$, and hence $t_1 = t_2$, thus $a(t_1) = a(t_2)$. If $|t_1 - t_2| \geq \pi$, then from (2.35) it follows that $\pi = |t_1 - t_2| - \pi$, i. e. $|t_1 - t_2| = 2\pi$, which is possible if and only if $a(t_1) = a(t_2)$.

Finally, $d_K(a(t_1), a(t_2)) \leq d_K(a(t_1), a(t_3)) + d_K(a(t_3), a(t_2))$ for all $t_1, t_2, t_3 \in K$. The equality holds if $a(t_3)$ lies on the arc between $a(t_1)$ and $a(t_2)$. Otherwise, the strict inequality holds true. □

Using the metrics (2.33), we define the *best representative* of the set $\mathcal{A}$ on the unit circle, as follows:

Definition 2.36. The best representative of the set $\mathcal{A} = \{a(t_i) \in K : t_i \in [0, 2\pi],\ i = 1, \ldots, m\}$ and weights $w_1, \ldots, w_m > 0$, with respect to the metric d_K defined by (2.33), is the point $c^\star(t^\star) = (\cos t^\star, \sin t^\star) \in K$, where

$$t^\star = \operatorname*{arg\,min}_{\tau \in [0, 2\pi]} \sum_{i=1}^{m} w_i\, d_K(a(\tau), a(t_i)), \quad a(\tau) = (\cos\tau, \sin\tau) \in K, \tag{2.36}$$

i. e. $t^\star \in [0, 2\pi]$ is the point at which the function $\Phi\colon [0, 2\pi] \to \mathbb{R}_+$ defined by

$$\Phi(\tau) = \sum_{i=1}^{m} w_i\, d_K(a(\tau), a(t_i)) \tag{2.37}$$

attains its global minimum.

Note that the function Φ does not have to be neither convex nor differentiable, and generally it may have several local minima. Therefore, this becomes a complex global optimization problem. In order to solve the GOP (2.36) one can apply the optimization algorithm DIRECT [8, 15, 29].

Example 2.37. Let $t_1, \ldots, t_m$ be a random sample from Gaussian normal distribution $\mathcal{N}(4, 1.2)$, and let $\mathcal{A} = \{a(t_i) = (\cos t_i, \sin t_i) \in \mathbb{R}^2 : i = 1, \ldots, m\}$ with weighs $w_i > 0$, $i = 1, \ldots, m$, be a data set. The set $\mathcal{A}$ is depicted as black points in Figure 2.9a, and the function $\tau \mapsto d_K(a(\tau), a(t_1))$ and the corresponding function Φ are shown in Figures 2.9b and 2.9c.

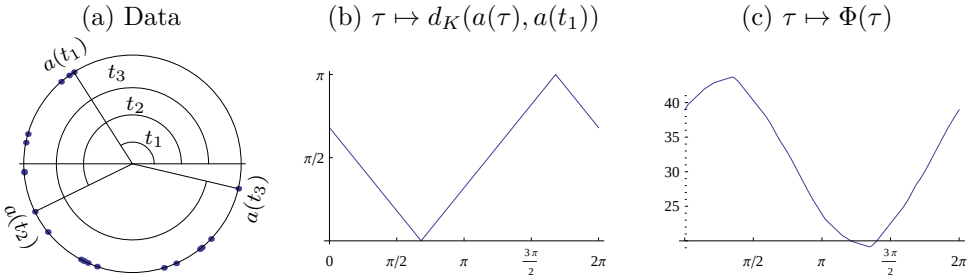


Figure 2.9: Data and the distances on the unit circle

2.4.2 Burn diagram

In order to graphically represent periodic events, it is appropriate to use *Burn diagram* (see e.g. [34]). In the Burn diagram, points are represented as $T = r\,(\cos t, \sin t)$, where (r, t) are the polar coordinates, i. e. t is the angle (in radians) between the x-axes and the radius vector of T, and r is the distance from T to the origin.

Example 2.38. Figure 2.10a shows earthquake positions in wider area of Osijek since 1880. Points in the Burn diagram (Figure 2.10b) identify individual earthquakes, where the distance to the origin represents the year when the earthquake happened, position on the circle reflects the day of the year, and the size of the point (small disc) corresponds to the magnitude. Figure 2.10 shows that the last stronger earthquake in close vicinity of Osijek happened by the end of winter 1922, located at geographic position (18.8, 45.7) (close to village Lug, some twenty kilometers to the northeast of Osijek).

a) Earthquake positions b) Time distribution of earthquakes

Figure 2.10: Earthquakes in wider area of Osijek since 1880

Chapter 3

Data clustering

Definition 3.1. Let $\mathcal{A}$ be a set of $m \geq 2$ elements. A partition of the set $\mathcal{A}$ into $1 \leq k \leq m$ disjoint nonempty subsets $\pi_1, \ldots, \pi_k$ such that

$$\bigcup_{j=1}^{k} \pi_j = \mathcal{A}, \qquad \pi_r \cap \pi_s = \emptyset,\ r \neq s, \qquad |\pi_j| \geq 1,\ j = 1, \ldots, k, \tag{3.1}$$

is called a ***k*-partition** of the set $\mathcal{A}$, and will be denoted by $\Pi = \{\pi_1, \ldots, \pi_k\}$. The elements of a partition are called ***clusters***, and the set of all partitions of $\mathcal{A}$ containing k clusters satisfying (3.1) will be denoted by $\mathcal{P}(\mathcal{A}; k)$.

Whenever we are going to talk about a partition of some set $\mathcal{A}$ we will always assume that it consists of subsets as described in Definition 3.1.

Theorem 3.2. *The number of all partitions of the set $\mathcal{A}$ consisting of k clusters is equal to the Stirling number of the second kind*

$$|\mathcal{P}(\mathcal{A}, k)| = \frac{1}{k!} \sum_{j=1}^{k} (-1)^{k-j} \binom{k}{j} j^m. \tag{3.2}$$

In the proof of Theorem 3.2 we are going to use the well-known *inclusion–exclusion principle* (see e.g. [10, p. 156]) written in the following form:

Lemma 3.3 (Inclusion–exclusion formula)**.** *Let $X_1, \ldots, X_k$ be subsets of a finite set X. The number of elements of X not belonging to any of the subsets $X_1, \ldots, X_k$ equals*

$$\Big| \bigcap_{i=1}^{k} \overline{X_i} \Big| = |X| - \sum_{1 \leq i \leq k} |X_i| + \sum_{1 \leq i < j \leq k} |X_i \cap X_j| - \cdots + (-1)^k |X_1 \cap \cdots \cap X_k|$$

where $\overline{X_i}$ denotes the complement $X \setminus X_i$.

Instead of a proof of Lemma 3.3, let us look at an example. Consider the set X containing 16 elements, and its three subsets: X_1 (7 elements inside the red circle), X_2 (7 element inside the blue circle), and X_3 (8 elements inside the green circle), as shown in Figure 3.1a. The intersections $X_1 \cap X_2$ and $X_1 \cap X_3$ have 4 elements each, and the intersection $X_2 \cap X_3$ has 5 elements (see Figures 3.1b, c, d). Finally, the intersection $X_1 \cap X_2 \cap X_3$ has 3 elements (see Figure 3.1a). Therefore

$$\begin{aligned} |\overline{X}_1 \cap \overline{X}_2 \cap \overline{X}_3| &= |X| - (|X_1| + |X_2| + |X_3|) + (|X_1 \cap X_2| + |X_1 \cap X_3| + |X_2 \cap X_3|) - \\ &\quad - |X_1 \cap X_2 \cap X_3| \\ &= 16 - (7 + 7 + 8) + (4 + 4 + 5) - 3 = 4. \end{aligned}$$

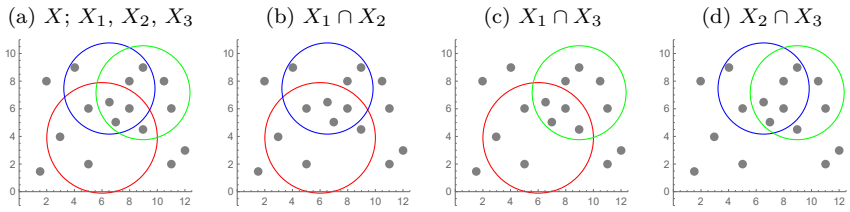


Figure 3.1: Number of elements of the set X not belonging to any of the subsets X_1, X_2, X_3

Proof of Theorem 3.2. Without loss of generality, let $\mathcal{A} = \{1, \ldots, m\}$, and let $\Pi^{(k)} = \{\pi_1, \ldots, \pi_k\}$ be its k-partition where $\pi_j \subset \mathcal{A}$ are disjoint nonempty subsets of $\mathcal{A}$. Let us denote the number of all such partitions by $|\mathcal{P}(\mathcal{A}; k)|$ and define the functions $f\colon \mathcal{A} \to J$, $J = \{1, \ldots, k\}$, by

$$f(x) = j, \quad \text{for } x \in \pi_j.$$

The number of these functions equals $|\mathcal{P}(\mathcal{A}; k)|$, and by permuting these k sets we obtain the number of all surjections from $\mathcal{A}$ onto J:

$$k!\,|\mathcal{P}(\mathcal{A}; k)|. \tag{3.3}$$

On the other hand, the number of all surjections from $\mathcal{A}$ onto J equals the number of all functions from $\mathcal{A}$ to J, minus the number of those functions which are not surjective.

Let $X = J^{\mathcal{A}}$ be the set of all functions from $\mathcal{A}$ to J. The number of all such functions is $|X| = k^m$ — the number of ways of selecting k, not necessarily distinct items from a collection of m items.

A function from $\mathcal{A}$ to J is not surjective if:

1. its image does not contain one element of J. The set X_i of all functions whose image does not contain the element $i \in J$ consists of exactly $(k-1)^m$ functions (the number of ways of selecting $k-1$, not necessarily distinct items from a collection of m items), and the set $\bigcup_{1\le i\le k} X_i$ of all functions missing exactly one element of J consists of $\binom{k}{1}(k-1)^m$ functions;

2. its image does not contain two distinct elements of J. The set $\bigcup_{1\le i<j\le k}(X_i \cap X_j)$ of all such functions contains $\binom{k}{2}(k-2)^m$ elements (the number of ways of selecting $k-2$, not necessarily distinct items from a collection of m items, for every pair of distinct elements of J);

   etc.

A function from $\mathcal{A}$ to J is surjective if and only if it does not belong to any of the sets $X_1, \dots, X_k$, i.e. if and only if it belongs to the set $\bigcap_{i=1}^{k} \overline{X_i}$. Using Lemma 3.3 we obtain the following number of all surjective functions from $\mathcal{A}$ to J:

$$\begin{aligned}
\Big|\bigcap_{i=1}^{k} \overline{X_i}\Big| &= |X| - \sum_{1\le i\le k} |X_i| + \sum_{1\le i<j\le k} |X_i \cap X_j| - \dots + (-1)^k |X_1 \cap \dots \cap X_k| \\
&= k^m - \binom{k}{1}(k-1)^m + \binom{k}{2}(k-2)^m - \dots + (-1)^k \binom{k}{k}(k-k)^m \\
&= \sum_{j=0}^{k} (-1)^j \binom{k}{j}(k-j)^m && [s := k-j] \\
&= \sum_{s=k}^{0} (-1)^{k-s} \binom{k}{k-s} s^m && [\text{for } s=0,\ s^m = 0] \\
&= \sum_{s=1}^{k} (-1)^{k-s} \binom{k}{k-s} s^m && [\text{since } \tbinom{n}{r} = \tbinom{n}{n-r}] \\
&= \sum_{s=1}^{k} (-1)^{k-s} \binom{k}{s} s^m \overset{[j:=s]}{=} \sum_{j=1}^{k} (-1)^{k-j} \binom{k}{j} j^m.
\end{aligned}$$

Using (3.3) we obtain (3.2), proving the theorem. □

In particular, Theorem 3.2 gives:

$$\begin{aligned}
\text{for } k=2: &\quad |\mathcal{P}(\mathcal{A};2)| = \tfrac{1}{2}(2^m - 2) = 2^{m-1} - 1, \\
\text{for } k=3: &\quad |\mathcal{P}(\mathcal{A};3)| = \tfrac{1}{2}\left(1 - 2^m + 3^{m-1}\right).
\end{aligned}$$

Example 3.4. The number of all k-partitions of a set $\mathcal{A}$, as described in Definition 3.1, can be rather huge. Table 3.1 shows the approximate number of all k-partitions of the set $\mathcal{A}$ for $m = 5, 10, 50, 1200, 10^6$, and $k = 2, 3, 4, 5, 6, 8, 10$.

$\approx \lvert\mathcal{P}(\mathcal{A}; k)\rvert$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=8$	$k=10$
$m=5$	15	25	10	1	–	–	–
$m=10$	511	9330	34105	42525	22827	750	1
$m=50$	10^{15}	10^{23}	10^{29}	10^{33}	10^{36}	10^{41}	10^{44}
$m=1200$	10^{361}	10^{572}	10^{721}	10^{837}	10^{931}	10^{1079}	10^{1193}
$m=10^6$	10^{301030}	10^{477120}	10^{602058}	10^{698968}	10^{778148}	10^{903085}	10^{10^6}

Table 3.1: Approximate number of all k-partitions for various numbers $m = |\mathcal{A}|$ and numbers $k = 2, 3, 4, 5, 6, 8, 10$ of clusters.

Example 3.5. Consider the set $\mathcal{A} \subset \mathbb{R}^2$, shown in Figure **??**a, containing $m = 1200$ elements. Table 3.1 shows the approximate number of all its k-partitions with $k = 2, 3, 4, 5, 6, 8$ and 10 clusters.

Provided that one defines the criterion that the *better* partition is the one whose clusters are *more compact* and *better separated*, one could ask the question of defining the *globally optimal (i.e. best) partition.*

3.1 Optimal k-partition

Let $\mathcal{A}$ be a set of $m \geq 2$ elements with $n \geq 1$ features. Since each feature is usually expressed by a number, one can, without loss of generality, always assume that a set $\mathcal{A}$ with $n \geq 1$ features is a subset of $\mathbb{R}^n$, $\mathcal{A} \subset \mathbb{R}^n$. For example, consider a group of 100 students with respect to their gender and height. Assigning the number 0 to males and 1 to females, and expressing heights in centimeters, one can identify this set of students with a subset of $\mathbb{R}^2$.

If we defined some distance-like function $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, we could have defined a *measure of compactness* and of *good separation* of clusters in a partition $\Pi = \{\pi_1, \ldots, \pi_k\}$ of the set $\mathcal{A} \subset \mathbb{R}^n$, as follows:

1. find a center $c_j \in \arg\min\limits_{x \in \mathbb{R}^n} \sum\limits_{a^i \in \pi_j} d(x, a^i)$ in every cluster π_j;
2. for every cluster π_j determine its total *dissipation* (the sum of distances from the points of π_j to the center c_j) $\mathcal{F}(\pi_j) = \sum\limits_{a^i \in \pi_j} d(c_j, a^i)$;

3. the sum $\sum_{j=1}^{k} \mathcal{F}(\pi_j)$ defines a measure of compactness and of good separation of clusters in the partition Π, and represents an objective function in this optimization problem (see (3.5)).

Figure **??** shows a partition for $k = 2, 3, 4, 5, 6, 7$, and 8 clusters and corresponding values of the LS-objective function. Notice that enlarging the number of clusters, decreases the objective function value. For example, Figure **??**c shows one of many (see Table 3.1) 3-partitions of the set $\mathcal{A}$. For this partition the objective function $\mathcal{F}_{LS}$ attains the value 19 860. It is plausible to ask whether this is the best 3-partition, or could one find a 3-partition with a smaller objective function value?

In general, we could pose at least some of the following questions:

1. Are the said objective functions the most appropriate ones for this example?

2. What is the most appropriate number of clusters in a partition?

3. Do the partitions shown in Figure **??** have the smallest objective function values among all possible partitions with those numbers of clusters?

From the previous example it is evident that the answers to the above questions won't be easy ones. The question of the choice of the objective function, as well as of the appropriate number of clusters in a partition, depends on the previous statistical analysis of the data. For objective functions, in this textbook, we are mostly going to use the LS distance-like function and the ℓ_1 metric function. In Section **??** we will deal with the choice of the most appropriate partition with spherical clusters, and in Section **??** with the choice of the most appropriate fuzzy-partition.

It needs to be said that the problem of finding an optimal partition is an NP-hard problem, [38], of a non-convex optimization of, in general, a non-differentiable function of several variables, which, in most cases, does have a substantial number of stationary points. In general, it won't be possible to carry-out the search for an optimal partition by searching the whole set $\mathcal{P}(\mathcal{A}; k)$. In the present textbook we are going to deal with finding the optimal partition with spherical clusters in Section **??**, finding the optimal partition with ellipsoidal clusters in Section **??**, and finding the optimal fuzzy-partition in Section **??**.

In general, given some distance-like function $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, where $\mathbb{R}_+ = [0, +\infty\rangle$ (see Section 2), to each cluster $\pi_j \in \Pi$ one can associate its

center

$$c_j \in \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \sum_{a \in \pi_j} d(x, a). \tag{3.4}$$

The quality of the partition determined by the objective function value $\mathcal{F}\colon \mathcal{P}(\mathcal{A}; k) \to \mathbb{R}_+$, is usually defined to be the sum over all clusters of the sums of distances from the points of clusters to their centers. The ***globally optimal k-partition (k-GOPart)*** is then considered to be the solution to the following ***global optimization problem*** `GOPart`:

$$\operatorname*{arg\,min}_{\Pi \in \mathcal{P}(\mathcal{A};k)} \mathcal{F}(\Pi), \qquad \mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a \in \pi_j} d(c_j, a). \tag{3.5}$$

Theorem 3.6. *Increasing the number of clusters in a partition does not increase the value of the objective function $\mathcal{F}$.*

For the proof of this theorem see page 60.

3.1.1 Minimal distance principle and Voronoi diagram

The minimal distance principle (see Algorithm 3.9 or (3.39)), is closely related to the so-called *Voronoi diagram* or *Dirichlet tessellation* (see e.g. [1, 21, 40]).

Let d be the usual Euclidean metric in the plane $\mathbb{R}^2$, and let us consider first the case of $k = 2$ clusters in the plane with centers c_1 and c_2. All elements $a \in \mathcal{A} \subset \mathbb{R}^2$ lying on the perpendicular bisector $\sigma(c_1, c_2)$ of the line segment $\overline{c_1 c_2}$ are equally distant from the centers c_1 and c_2. The line bisector $\sigma(c_1, c_2)$ is perpendicular to the segment $\overline{c_1 c_2}$ and divides the plane $\mathbb{R}^2$ into two half-planes — *Voronoi regions*:

$$VR(c_1) = \{x \in \mathbb{R}^2 : d(c_1, x) < d(c_2, x)\},$$
$$VR(c_2) = \{x \in \mathbb{R}^2 : d(c_1, x) > d(c_2, x)\}.$$

The perpendicular bisector $\sigma(c_1, c_2)$ represents the Voronoi diagram of the set of centers $\{c_1, c_2\}$ (see Figure 3.2a).

In the case of $k = 3$ clusters in the plane with centers c_1, c_2, and c_3, the perpendicular bisector $\sigma(c_1, c_2)$ of the line segment $\overline{c_1 c_3}$ defines two half-planes $M(c_1, c_2)$ and $M(c_2, c_1)$, the perpendicular bisector $\sigma(c_1, c_3)$ of the line segment $\overline{c_1 c_3}$ defines the half-planes $M(c_1, c_3)$ and $M(c_3, c_1)$, and the perpendicular bisector $\sigma(c_2, c_3)$ of the line segment $\overline{c_2 c_3}$ defines the half-planes $M(c_2, c_3)$ and $M(c_3, c_2)$. Voronoi regions with centers c_1, c_2, and c_3

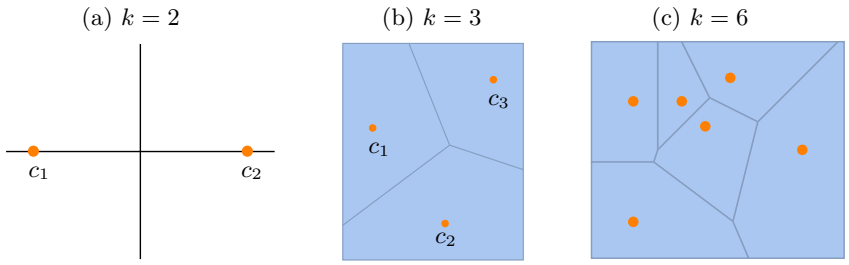


Figure 3.2: Minimal distance principle and Voronoi diagram

are defined as:

$$VR(c_1) = M(c_1, c_2) \cap M(c_1, c_3),$$
$$VR(c_2) = M(c_2, c_1) \cap M(c_2, c_3),$$
$$VR(c_3) = M(c_3, c_1) \cap M(c_3, c_2),$$

and the Voronoi diagram of the centers c_1, c_2, c_3 is defined as (see Figure 3.2b)

$$V(c_1, c_2, c_3) = (\overline{VR(c_1)} \cap \overline{VR(c_2)}) \bigcup (\overline{VR(c_1)} \cap \overline{VR(c_3)}) \bigcup (\overline{VR(c_2)} \cap \overline{VR(c_3)})$$

where $\overline{VR(c_1)}$ denotes the (topological) closure of $VR(c_1)$, and similarly for other regions.

In general, for k clusters with centers $c_1, \ldots, c_k$, the Voronoi regions are defined as

$$VR(c_j) = \bigcap_{s \neq j} M(c_j, c_s), \quad j = 1, \ldots, k,$$

and the Voronoi diagram is defined as the union of intersections of the closures of Voronoi regions

$$V(c_1, \ldots, c_k) = \bigcup_{s \neq j} \overline{VR(c_j)} \cap \overline{VR(c_s)}.$$

Note that the cluster $\pi(c_j)$, obtained by the minimal distance principle (3.39), lies in the Voronoi region $VR(c_j)$ bounded by the Voronoi diagram.

Figure 3.2c, generated by *Mathematica* computation system, shows the Voronoi diagram for six centers.

Exercise 3.7. Define and draw Voronoi diagrams in case of ℓ_1 and ℓ_∞ metric functions.

Exercise 3.8. Determine the Voronoi diagram for $k = 3$ by considering the circle circumscribed to the triangle $\triangle(c_1, c_2, c_3)$. Can such line of thought be applied for $k > 3$ also?

3.1.2 k-means algorithm I

There is no method for successfully solving the GOP (3.5). Nevertheless, there exists the well-known k-means algorithm giving locally optimal solution which heavily depends on the choice of initial approximation. Choosing an initial partition $\Pi^{(0)}$, the k-means algorithm finds in finitely many steps a locally optimal partition. The algorithm is usually set up in two steps which are iteratively successively repeated until the new partition does not differ from the previous one.

Algorithm 3.9 (k-means algorithm I)**.**

Step A: assignment step. Given a finite subset $\mathcal{A} \subset \mathbb{R}^n$ and the set of points $z_1, \ldots, z_k \in \mathbb{R}^n$, apply the minimal distance principle to determine clusters π_j, $j = 1, \ldots, k$, to get the partition $\Pi = \{\pi_1, \ldots, \pi_k\}$,

$$\pi_j := \pi_j(z_j) = \{a \in \mathcal{A} : d(z_j, a) \leq d(z_s, a) \text{ for all } s = 1, \ldots, k\}.$$

Step B: update step. For the given partition $\Pi = \{\pi_1, \ldots, \pi_k\}$ of the set $\mathcal{A}$ determine cluster centers $c_j \in \arg\min_{x \in \mathbb{R}^n} \sum_{a \in \pi_j} d(x, a)$, $j = 1, \ldots, k$, and calculate objective function value $\mathcal{F}(\Pi)$ according to (3.5);

Set $z_j = c_j$, $j = 1, \ldots, k$;

Remark 3.10. It might happen that in Step A some elements $a \in \mathcal{A}$ lie on the border between two or several clusters. The decision as to which cluster should such elements be designated, can drastically influence the further course of the iterative process (see [32]). An example of such a situation occurs in the problem of defining optimal electoral districts (see Example **??**). Almost always it becomes necessary to divide the electorate of a city into two or several electoral districts (in Croatia this is the case with the city of Zagreb). We are going to consider this problem later, when discussing fuzzy clustering of data in Section **??**.

The usual convention for simple clustering of data is to put the datum, which occurs on the border of two or several clusters, into the first cluster in order.

Example 3.11. Let us determine, using the k-means algorithm, the LS-optimal 3-partition of the set $\mathcal{A} = \{0, 2, 4, 8, 9, 10, 12, 16\}$ starting with the initial partition $\Pi^{(0)} = \{\{0, 2, 4\}, \{8, 9\}, \{10, 12, 16\}\}$.

Iteration	π_1	π_2	π_3	c_1	c_2	c_3	$\mathcal{F}_{LS}(\Pi)$
0	{0, 2, 4}	{8, 9}	{10, 12, 16}	2	8.5	12.67	27.17
1	{0, 2, 4}	{8, 9, 10}	{12, 16}	2	9	14	18
2	{0, 2, 4}	{8, 9, 10}	{12, 16}	2	9	14	18

Table 3.2: Searching for LS-optimal 3-partition of the set $\{0, 2, 4, 8, 9, 10, 12, 16\}$. Results of Step A are colored blue and of Step B orange.

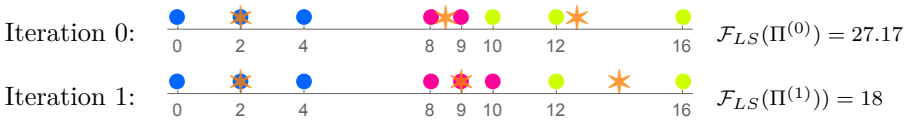


Figure 3.3: Searching for LS-optimal 3-partition of the set $\{0, 2, 4, 8, 9, 10, 12, 16\}$

Exercise 3.12. Using the k-means algorithm find the ℓ_1-optimal 3-partition of the set from Example 3.11 starting with the same initial partition.

The following theorem shows that the sequence of objective function values obtained by the k-means algorithm is monotonically decreasing (see also Theorem **??**).

Theorem 3.13. *Let $\mathcal{A} \subset \mathbb{R}^n$ be a set, $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ a distance-like function, and $\mathcal{F}$ the objective function given by (3.5). Applying the k-means algorithm, the value of objective function $\mathcal{F}$ is not going to increase.*

Proof. Let $\Pi^{(t)} = \{\pi_1^{(t)}, \ldots, \pi_k^{(t)}\}$ be a partition with centers $c^{(t)} = \{c_1^{(t)}, \ldots, c_k^{(t)}\}$ and $\mathcal{F}(\Pi^{(t)})$ be the corresponding objective function value.

Applying Step A (the minimal distance principle) to the set $\mathcal{A}$ with centers $c^{(t)}$, we obtain a new partition $\Pi^{(t+1)} = \{\pi_1^{(t+1)}, \ldots, \pi_k^{(t+1)}\}$ satisfying

$$\mathcal{F}(\Pi^{(t)}) = \sum_{j=1}^{k} \sum_{a \in \pi_j^{(t)}} d(c_j^{(t)}, a) \overset{\text{(Step A)}}{\geq} \sum_{j=1}^{k} \sum_{a \in \pi_j^{(t+1)}} d(c_j^{(t)}, a).$$

Next, applying Step B to each cluster $\pi_j^{(t+1)}$ (to determine new centers $c_j^{(t+1)}$), we obtain

$$\mathcal{F}(\Pi^{(t)}) \overset{\text{(Step A)}}{\geq} \sum_{j=1}^{k} \sum_{a \in \pi_j^{(t+1)}} d(c_j^{(t)}, a) \overset{\text{(Step B)}}{\geq} \sum_{j=1}^{k} \sum_{a \in \pi_j^{(t+1)}} d(c_j^{(t+1)}, a).$$

Therefore, $\mathcal{F}(\Pi^{(t)}) \geq \mathcal{F}(\Pi^{(t+1)})$. □

Example 3.14. Determine LS-optimal 2-partition of the set $\mathcal{A} = \{0, 2, 3\}$ by using the k-means algorithm with initial partition $\Pi^{(0)} = \{\{0, 2\}, \{3\}\}$.

Iteration	π_1	π_2	c_1	c_2	$\mathcal{F}_{LS}(\Pi)$
1	{0,2}	{3}	1	3	2
2	{0,2}	{3}	1	3	2

Table 3.3: Searching for the LS-optimal 2-partition of the set $\mathcal{A} = \{0, 2, 3\}$

Table 3.3 shows that using the LS distance-like function, the k-means algorithm sometimes cannot improve even the initial partition. However, in the previous example, a better partition is $\Pi^\star = \{\{0\}, \{2, 3\}\}$ because $\mathcal{F}_{LS}(\Pi^\star) = 0.5$. This simple example shows that the k-means algorithm gives a locally optimal partition. Choosing another initial partition we might have got the k-`GOPart`. Give it a try!

Besides the aforementioned shortcoming of the k-means algorithm to heavily depend on the initial partition and to produce only some locally optimal partition, as in Example 3.14, one should also point out yet another limitation: during the iterative process it might happen that some clusters become empty sets, i.e. it might happen that the number of clusters decreases (see Example **??**, page **??**).

3.2 Clustering data with one feature

Let $\mathcal{A}$ be a set of $m \geq 2$ elements with one feature. As we have remarked on page 30, such a set can be considered as a subset of $\mathbb{R}$, i.e. $\mathcal{A} = \{a^1, \ldots, a^m\} \subset \mathbb{R}$. The set $\mathcal{A}$ should be grouped into $1 \leq k \leq m$ clusters $\pi_1, \ldots, \pi_k$ conforming with Definition 3.1. For example, days of a year can be clustered into three clusters according to daily average temperatures expressed in °C: cluster of cold days, cluster of days with mild temperature, and cluster of warm days. According to the named feature, we are going to represent every element $a \in \mathcal{A}$ with a real number which we will also denote by a. Therefore, from now on, we are going to assume that $\mathcal{A} = \{a^1, \ldots, a^m\}$ is a *multiset* of data, i.e. some elements may appear multiple times in $\mathcal{A}$. So, in our example, the multiset $\mathcal{A}$ would have 365 elements, all being, say, 8, 15, or 22.

Given a distance-like function $d\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}_+$, one can associate to every cluster $\pi_j\in\Pi$ its center c_j as follows:

$$c_j \in \operatorname*{arg\,min}_{x\in\mathbb{R}} \sum_{a\in\pi_j} d(x,a), \quad j=1,\dots,k. \tag{3.6}$$

Furthermore, if we define an objective function $\mathcal{F}\colon \mathcal{P}(\mathcal{A};k)\to\mathbb{R}_+$ on the set $\mathcal{P}(\mathcal{A};k)$ of all partitions of $\mathcal{A}$, by

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k}\sum_{a\in\pi_j} d(c_j,a), \tag{3.7}$$

then the search for an optimal k-`GOPart` is done by solving the following optimization problem:

$$\operatorname*{arg\,min}_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{F}(\Pi). \tag{3.8}$$

Note that the k-`GOPart` will have the property that the sum of *dissipations* (the sum of deviations) of cluster elements to its center, is minimal. In this way we attempt to obtain as good the inner compactness and separation between clusters, as possible.

Remark 3.15. Number of all k-partitions of a set $\mathcal{A}$ with m elements can be rather huge (see Table 3.1). But in the case of data with one feature ($\mathcal{A}\subset\mathbb{R}$) it is obvious that the optimal partition can be expected among partitions where clusters follow one another. This means that all elements of cluster π_2 are on the right hand side of cluster π_1, all elements of cluster π_3 are on the right hand side of cluster π_2, etc. (see [26, p. 161]). The number of such partitions is considerably smaller as shown by the following proposition (see also Table 3.4).

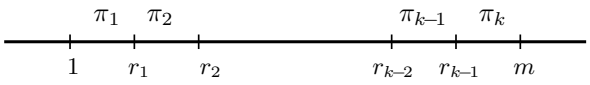
$\binom{m-1}{k-1}$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=8$	$k=10$
$m=10$	9	36	84	126	126	36	1
$m=30$	29	406	3 654	23 751	118 755	1 560 780	10 015 005
$m=50$	49	1 176	18 424	211 876	1 906 884	85 900 584	2 054 455 634

Table 3.4: Number of k-partitions of a set $\mathcal{A}\subset\mathbb{R}$ whose clusters follow one another

Proposition 3.16. *Let $\mathcal{A}=\{a^i\in\mathbb{R}\colon i=1,\dots,m\}$. The number of all k-partitions of the set $\mathcal{A}$ whose clusters $\pi_1,\dots,\pi_k$ follow one another equals*

$$\binom{m-1}{k-1}. \tag{3.9}$$

Proof. Without loss of generality, assume $\mathcal{A} = \{1, \ldots, m\}$.



Obviously the smallest element of cluster π_1 has to be $1 \in \mathcal{A}$, and the largest element of cluster π_k has to be $m \in \mathcal{A}$. Denote the largest elements of clusters $\pi_1, \ldots, \pi_{k-1}$ by $r_1, \ldots, r_{k-1}$. These numbers satisfy $1 \leq r_1 < r_2 < \cdots < r_{k-1} < m$. Therefore the question about number of all k-partitions of the set $\mathcal{A}$ whose clusters follow one another, boils down to the question of number of elements of the set

$$S = \{(r_1, \ldots, r_{k-1}) \in \mathcal{A}^{k-1} \colon 1 \leq r_1 < r_2 < \cdots < r_{k-1} < m\},$$

i.e. on the number of all subsets of the set $\{1, \ldots, m-1\}$ with $k-1$ elements. And this is the number of $(k-1)$-combinations of a set with $m-1$ elements. □

3.2.1 Application of LS distance-like function

Let $\Pi = \{\pi_1, \ldots, \pi_k\}$ be a k-partition of the set $\mathcal{A} \subset \mathbb{R}$, and $d_{LS} \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$, defined by $d_{LS}(x, y) = (x - y)^2$, be the LS-distace-like function. The centers of clusters $\pi_1, \ldots, \pi_k$ are called *centroids* and are determined as follows:

$$c_j = \operatorname*{arg\,min}_{x \in \mathbb{R}} \sum_{a \in \pi_j} (x - a)^2 = \frac{1}{|\pi_j|} \sum_{a \in \pi_j} a, \quad j = 1, \ldots, k, \tag{3.10}$$

and the objective function (3.7) is defined by

$$\mathcal{F}_{LS}(\Pi) = \sum_{j=1}^{k} \sum_{a \in \pi_j} (c_j - a)^2. \tag{3.11}$$

Example 3.17. Given the set $\mathcal{A} = \{2, 4, 8, 10, 16\}$, find all its 3-partitions satisfying Definition 3.1 and whose clusters follow one another. Determine also the corresponding centroids and the objective function $\mathcal{F}_{LS}$.

According to Stirling formula (3.2), the number of all 3-partitions of the set $\mathcal{A}$ is 25. But the number of 3-partitions of the same set with clusters following one another is only $\binom{5-1}{3-1} = \frac{4!}{2!\cdot 2!} = 6$, see Table 3.5. From this table we see that the LS-optimal 3-partition is $\Pi^\star = \{\{2, 4\}, \{8, 10\}, \{16\}\}$, the one in the fifth row, where the objective function $\mathcal{F}_{LS}$ attains its (global) minimum $\mathcal{F}_{LS}(\Pi^\star) = 4$, and $\Pi^\star$ is therefore the `LS 3-GOPart`.

π_1	π_2	π_3	c_1	c_2	c_3	$\mathcal{F}_{LS}(\Pi)$	$\mathcal{G}(\Pi)$
{2}	{4}	{8,10,16}	2	4	11.33	$0+0+34.67=34.67$	$36+16+33.33=85.33$
{2}	{4,8}	{10,16}	2	6	13	$0+8+18=26$	$36+8+50=94$
{2}	{4,8,10}	{16}	2	7.33	16	$0+18.67+0=18.67$	$36+1.33+64=101.33$
{2,4}	{8}	{10,16}	3	8	13	$2+0+18=20$	$50+0+50=100$
{2,4}	{8,10}	{16}	3	9	16	$2+2+0=4$	$50+2+64=116$
{2,4,8}	{10}	{16}	4.67	10	16	$18.67+0+0=18.67$	$33.33+4+64=101.33$

Table 3.5: All 3-partitions of $\mathcal{A}=\{2,4,8,10,16\}$ whose clusters follow one another

Exercise 3.18. What is the number of all 3-partitions, and the number of 3-partitions with clusters following one another, of the set $\mathcal{A}=\{1,4,5,8,10,12,15\}$? Write down all 3-partitions with clusters following one another, and find the LS-optimal one.

Solution: The number of all partitions is 301, and the number of partitions with clusters following one another is 15. The LS-optimal 3-partition is $\Pi^\star=\{\{1,4,5\},\{8,10\},\{12,15\}\}$, and $\mathcal{F}(\Pi^\star)=\frac{91}{6}\approx 15.1667$.

3.2.2 The dual problem

The following lemma shows that applying the LS distance-like function, the dissipation of the set $\mathcal{A}$ about its center c equals the sum of dissipations of clusters π_j, $j=1,\dots,k$ about their centers c_j, $j=1,\dots,k$, and the weighted sum of squared distances between c and c_j, where the weights are determined by the size of sets π_j.

Lemma 3.19. *Let $\mathcal{A}=\{a^1,\dots,a^m\}$ be a data set, let $\Pi=\{\pi_1,\dots,\pi_k\}$ be its k-partition with clusters $\pi_1,\dots,\pi_k$, and let*

$$c=\frac{1}{m}\sum_{i=1}^{m}a^i,\qquad c_j=\frac{1}{|\pi_j|}\sum_{a\in\pi_j}a,\quad j=1,\dots,k. \tag{3.12}$$

Then

$$\sum_{i=1}^{m}(c-a^i)^2=\mathcal{F}_{LS}(\Pi)+\mathcal{G}(\Pi), \tag{3.13}$$

where

$$\mathcal{F}_{LS}(\Pi) = \sum_{j=1}^{k} \sum_{a \in \pi_j} (c_j - a)^2, \tag{3.14}$$

$$\mathcal{G}(\Pi) = \sum_{j=1}^{k} |\pi_j| \, (c_j - c)^2. \tag{3.15}$$

Proof. Notice that for c_j we have $\sum\limits_{a^i \in \pi_j} (c_j - a^i) = 0$. Using this, for every $x \in \mathbb{R}$ we have

$$\begin{aligned} \sum_{a^i \in \pi_j} (x - a^i)^2 &= \sum_{a^i \in \pi_j} \left((x - c_j) + (c_j - a^i)\right)^2 \\ &= \sum_{a^i \in \pi_j} (x - c_j)^2 + 2 \sum_{a^i \in \pi_j} (x - c_j)(c_j - a^i) + \sum_{a^i \in \pi_j} (c_j - a^i)^2 \\ &= |\pi_j| \, (x - c_j)^2 + \sum_{a^i \in \pi_j} (c_j - a^i)^2, \end{aligned}$$

i.e.

$$\sum_{a^i \in \pi_j} (x - a^i)^2 = \sum_{a^i \in \pi_j} (c_j - a^i)^2 + |\pi_j| \, (c_j - x)^2, \quad j = 1, \ldots, k. \tag{3.16}$$

If we put $c = \frac{1}{m} \sum\limits_{i=1}^{m} a^i$ in (3.16) instead of x and sum all equations, we obtain (3.13). □

The objective function $\mathcal{F}_{LS}$ occurred naturally in formula (3.13), and this formula shows that the total dissipation of elements of the set $\mathcal{A}$ about its centroid c, can be described as the sum of two objective functions $\mathcal{F}_{LS}$ and $\mathcal{G}$.

In particular, the LS-optimal 3-partition of the set $\mathcal{A}$ in Example 3.17 is $\Pi^\star = \{\{2, 4\}, \{8, 10\}, \{16\}\}$, for which $\mathcal{F}(\Pi^\star) = 4$ (see Table 3.5). The obvious question is: what is $\mathcal{G}(\Pi^\star)$ in this example?

To answer this question, let us expand Table 3.5 by adding values of the function $\mathcal{G}$ for each partition (blue part of the table). Notice that the sum $\mathcal{F}_{LS}(\Pi) + \mathcal{G}(\Pi)$ is constant and equals $\sum\limits_{i=1}^{m} (c - a^i)^2 = 120$, which is in accordance with (3.13), and the maximal value of the function $\mathcal{G}$ is attained precisely at the LS-optimal 3-partition $\Pi^\star$ for which the objective function F_{LS} attains its minimal value.

Is this accidental?

To answer this question, let us first try to solve the following example which considers a similar problem. For this we need some foreknowledge from calculus(see e.g. [11, 13]).

Example 3.20. Let $\varphi, \psi \in C^2(\mathbb{R})$ be two functions such that $\varphi(x)+\psi(x) = \kappa$ for some constant $\kappa \in \mathbb{R}$. The function φ attains its local minimum at $x_0 \in \mathbb{R}$ if and only if the function ψ attains at x_0 its local maximum, and $\varphi(x_0) = \kappa - \psi(x_0)$.

If $\varphi'(x_0) = 0$, then $\psi'(x_0) = 0$, and vice versa. Also, if $\varphi''(x_0) > 0$, then $\psi''(x_0) < 0$, and vice versa. Therefore we have

- $x_0 \in \arg\min_{x\in\mathbb{R}} \varphi(x)$ if and only if $x_0 \in \arg\max_{x\in\mathbb{R}} \psi(x)$;
- $\min_{x\in\mathbb{R}} \varphi(x) = \kappa - \max_{x\in\mathbb{R}} \psi(x)$, i.e. $\varphi(x_0) = \kappa - \psi(x_0)$.

Check whether the two functions $\varphi(x) = x^2 - 1$ and $\psi(x) = -x^2 + 3$ satisfy these properties. Draw their graphs in the same coordinate system. Try to come up yourself with another example of a pair of functions φ, ψ satisfying the said properties.

The next theorem follows directly from Lemma 3.19 [35].

Theorem 3.21. *Using the notation from Lemma 3.19, there exists a partition* $\Pi^\star \in \mathcal{P}(\mathcal{A}; k)$ *such that*

(*i*) $\Pi^\star \in \arg\min_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{F}_{LS}(\Pi) = \arg\max_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{G}(\Pi)$,

(*ii*) $\min_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{F}_{LS}(\Pi) = \mathcal{F}_{LS}(\Pi^\star)$ *and* $\max_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{G}(\Pi) = \mathcal{G}(\Pi^\star)$,

where $\mathcal{G}(\Pi^\star) = \sum_{i=1}^{m} (c - a^i)^2 - \mathcal{F}_{LS}(\Pi^\star)$.

This means that in order to find the LS-optimal partition, instead of minimizing the function $\mathcal{F}_{LS}$ given by (3.11), one can maximize the dual function $\mathcal{G}$:

$$\arg\max_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{G}(\Pi), \qquad \mathcal{G}(\Pi) = \sum_{j=1}^{k} |\pi_j|(c_j - c)^2. \tag{3.17}$$

The optimization problem (3.17) is called the ***dual problem*** with respect to the optimization problem $\arg\min_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{F}_{LS}(\Pi)$.

One can say that the LS-optimal partition has the property that the sum of dissipations of cluster elements (sum over all clusters of the sums of

LS-distances from cluster elements to their centroids) minimal, and at the same time the centroids of clusters are distant from each another as much as possible. In this way one achieves the best inner compactness and best separation between clusters.

3.2.3 Least absolute deviation principle

Let $\Pi = \{\pi_1, \ldots, \pi_k\}$ be a k-partition of the set $\mathcal{A} \subset \mathbb{R}$, and $d_1 \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$, defined by $d_1(x, y) = |x - y|$, be the ℓ_1 metric function. The centers $c_1, \ldots, c_k$ of clusters $\pi_1, \ldots, \pi_k$ are determined by

$$c_j = \operatorname{med}(\pi_j) \in \operatorname{Med}(\pi_j) = \arg\min_{x \in \mathbb{R}} \sum_{a \in \pi_j} |x - a|, \quad j = 1, \ldots, k, \tag{3.18}$$

and the objective function (3.7) is defined by

$$\mathcal{F}_1(\Pi) = \sum_{j=1}^{k} \sum_{a \in \pi_j} |c_j - a|. \tag{3.19}$$

If one uses (3.20) from Exercise 3.24 then, in order to calculate the objective function (3.19), one doesn't need to know the centers of clusters (3.18), which speeds up the calculation process.

Example 3.22. Let $\mathcal{A} = \{2, 4, 8, 10, 16\}$ be the set as in Example 3.17. We want to find all of its 3-partitions satisfying Definition 3.1 with clusters following one another.

π_1	π_2	π_3	c_1	c_2	c_3	$\mathcal{F}_1(\Pi)$
{2}	{4}	{8,10,16}	2	4	10	$0+0+8=8$
{2}	{4,8}	{10,16}	2	6	13	$0+4+6=10$
{2}	{4,8,10}	{16}	2	8	16	$0+6+0=6$
{2,4}	{8}	{10,16}	3	8	13	$2+0+6=8$
{2,4}	{8,10}	{16}	3	9	16	2+2+0=4
{2,4,8}	{10}	{16}	4	10	16	6+0+0=6

Table 3.6: Partitions of the set $\mathcal{A}$ with clusters following one another

In addition, we want to determine the corresponding cluster centers and the values of the objective function $\mathcal{F}_1$ using the ℓ_1 metric function, and then find the globally ℓ_1-optimal 3-partition.

The number of all 3-partitions with clusters following one another is $\binom{m-1}{k-1} = 6$ and, as shown in Table 3.6, the ℓ_1-optimal 3-partition is $\Pi^\star = \{\{2,4\},\{8,10\},\{16\}\}$ since the objective function $\mathcal{F}_1$ defined by (3.19) attains at $\Pi^\star$ its lowest value (global minimum). Hence, partition $\Pi^\star$ is the ℓ_1-`GOPart`.

Exercise 3.23. Among all partitions of the set $\mathcal{A} = \{1,4,5,8,10,12,15\}$ from Exercise 3.18, find the ℓ_1-optimal 3-partition.

Exercise 3.24. Let $\mathcal{A} = \{a^1,\ldots,a^m\}$ be a finite increasing sequence of real numbers. Prove the following:

$$\sum_{i=1}^{m} |a^i - \operatorname{med}(\mathcal{A})| = \sum_{i=1}^{\lceil \frac{m}{2} \rceil} \left(a^{m-i+1} - a^i\right), \tag{3.20}$$

where $\lceil x \rceil$ equals x if x in an integer, and $\lceil x \rceil$ is the smallest integer larger then x if x is not an integer.[1] For example, $\lceil 20 \rceil = 20$, whereas $\lceil 20.3 \rceil = 21$.

3.2.4 Clustering weighted data

Let $\mathcal{A} = \{a^1,\ldots,a^m\} \subset \mathbb{R}$ be a data set of real numbers and to each datum $a^i \in \mathcal{A}$ a corresponding weight $w_i > 0$ is assigned. For example, in [30, Example 3.8], where the authors analyses the problem of high water levels of the river Drava at Donji Miholjac, data are days of the year and weights are the measured water levels.

In the case of weighted data the objective function (3.7) becomes

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a^i \in \pi_j} w_i \, d(c_j, a^i), \tag{3.21}$$

where

$$c_j \in \arg\min_{x \in \mathbb{R}} \sum_{a^i \in \pi_j} w_i \, d(x, a^i), \quad j = 1,\ldots,k. \tag{3.22}$$

In particular, when applying the LS distance-like function, the centers c_j of clusters π_j are weighted arithmetic means of data in π_j

$$c_j = \frac{1}{\kappa^j} \sum_{a^i \in \pi_j} w_i \, a^i, \qquad \kappa^j = \sum_{a^i \in \pi_j} w_i, \tag{3.23}$$

and when applying the ℓ_1 metric function, the centers c_j of clusters π_j are weighted medians of data in π_j [25, 39]

$$c_j = \operatorname*{med}_{a^i \in \pi_j}(w_i, a^i) \in \operatorname{Med}(w, \mathcal{A}). \tag{3.24}$$

[1] In *Mathematica* computation system, $\lceil x \rceil$ is obtained by `Ceiling[x]`, and $\lfloor x \rfloor$ by `Floor[x]`.

Example 3.25. Let us again consider the set $\mathcal{A} = \{1, 4, 5, 8, 10, 12, 15\}$ from Exercise 3.18. Assign to each but the last datum the weight 1, and to the last datum the weight 3. Now the LS-optimal 3-partition becomes $\Pi^\star = \{\{1, 4, 5\}, \{8, 10, 12\}, \{15\}\}$ with centroids $\frac{10}{3}$, 10, and 15, and the objective function value $\mathcal{F}(\Pi^\star) = \frac{50}{3} = 16.667$.

In order to determine the centers of clusters when applying the ℓ_1 metric function, one has to know how to calculate weighted median of the data. As mentioned in Section 2.1.3, this might turn out to be not a simple task. If the weights are integers, the problem can be reduced to finding the usual median of data (see Example 2.16). If the weights are not integers, then by multiplying with some appropriate number and taking approximations, one can reduce the weights to integers.

Exercise 3.26. Find the ℓ_1-optimal 3-partition of the set $\mathcal{A}$ from the previous example with all weights being equal 1, and in the case when the weights are assigned as in the previous example.

Exercise 3.27. Write down formulas for the centroid of the set $\mathcal{A}$, and for the objective functions $\mathcal{F}$ and $\mathcal{G}$ for the data set $\mathcal{A}$ with weights $w_1, \ldots, w_m > 0$.

Solution: $\mathcal{G}(\Pi) = \sum_{j=1}^{k} \big(\sum_{\pi_j} w_s\big)(c_j - c)^2$.

3.3 Clustering data with two or several features

Let $\mathcal{A}$ be a set of $m \geq 2$ elements with $n \geq 2$ features. As already said, such a set can be regarded as a subset of $\mathbb{R}^n$, i.e. $\mathcal{A} = \{a^i = (a^i_1, \ldots, a^i_n) \in \mathbb{R}^n : i = 1, \ldots, m\}$. The set $\mathcal{A}$ should be grouped in accordance with Definition 3.1, into $1 \leq k \leq m$ disjoint nonempty clusters. For example, elements of the set $\mathcal{A} \subset \mathbb{R}^2$ from Example 3.5 have two features — the abscissa and the ordinate, and the elements can be grouped into $2, 3, 4, 5, 6, 7, 8$ or more clusters (see Figure **??**).

Let $\Pi \in \mathcal{P}(\mathcal{A}; k)$ be a partition of the set $\mathcal{A}$. Given a distance-like function $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, to each cluster $\pi_j \in \Pi$ one can assign its center c_j in the following way:

$$c_j \in \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \sum_{a \in \pi_j} d(x, a), \quad j = 1, \ldots, k. \tag{3.25}$$

Analogously to the case of data with one feature, if we define the objective function $\mathcal{F}\colon \mathcal{P}(\mathcal{A}; k) \to \mathbb{R}_+$ on the sets of all partitions $\mathcal{P}(\mathcal{A}; k)$ of the set $\mathcal{A}$

consisting of k clusters, by

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a \in \pi_j} d(c_j, a), \tag{3.26}$$

then we search for the optimal k-partition by solving the following GOP:

$$\operatorname*{arg\,min}_{\Pi \in \mathcal{P}(\mathcal{A};k)} \mathcal{F}(\Pi). \tag{3.27}$$

Note that the optimal k-partition has the property that the dissipation (the sum of d-distances of the cluster elements to their centers) is minimal. In this way we attempt to achieve as good the inner compactness of clusters as possible.

3.3.1 Least squares principle

Let $\Pi = \{\pi_1, \ldots, \pi_k\}$ be a partition of the set $\mathcal{A} = \{a^i = (a_1^i, \ldots, a_n^i) \in \mathbb{R}^n : i = 1, \ldots, m\}$. The centers $c_1, \ldots, c_k$ of clusters $\pi_1, \ldots, \pi_k$ for the LS distance-like function $d_{LS} \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, defined as $d_{LS}(a, b) = \|a - b\|^2$, are called *centroids* and are obtained as follows:

$$\begin{aligned} c_j &= \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \sum_{a \in \pi_j} \|x - a\|^2 = \tfrac{1}{|\pi_j|} \sum_{a \in \pi_j} a \\ &= \Big(\tfrac{1}{|\pi_j|} \sum_{a \in \pi_j} a_1, \ldots, \tfrac{1}{|\pi_j|} \sum_{a \in \pi_j} a_n \Big), \qquad j = 1, \ldots, k, \end{aligned} \tag{3.28}$$

where $\sum\limits_{a \in \pi_j} a_\ell$, $\ell = 1, \ldots, n$, denotes the sum of ℓ-th components of all elements in cluster π_j. In this case the objective function (3.26) is defined by

$$\mathcal{F}_{LS}(\Pi) = \sum_{j=1}^{k} \sum_{a \in \pi_j} \|c_j - a\|^2. \tag{3.29}$$

Example 3.28. Consider the set $\mathcal{A} = \{a^1 = (1, 1), a^2 = (3, 1), a^3 = (3, 2), a^4 = (2, 2)\}$ in the plane. The number of its 2-partitions is $\mathcal{P}(\mathcal{A}; 2) = 2^{4-1} - 1 = 7$ and they are all listed in Table 3.7. Let us find the optimal 2-partition.

The elements of $\mathcal{A}$ have two features — the abscissa and the ordinate, so the set $\mathcal{A}$ can be simply recorded as $\mathcal{A} = \{a^i = (x_i, y_i) \in \mathbb{R}^2 : i = 1, \ldots, 4\}$, and depicted in the plane (see Figure 3.4).

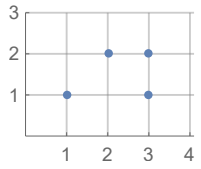


Figure 3.4: The set $\mathcal{A} \subset \mathbb{R}^2$

According to (3.2) the set $\mathcal{A}$ has 7 different 2-partitions. Let $\Pi = \{\pi_1, \pi_2\}$ be one of them. Its centroid is defined by

$$c_1 = \frac{1}{|\pi_1|} \sum_{a \in \pi_1} a, \qquad c_2 = \frac{1}{|\pi_2|} \sum_{a \in \pi_2} a,$$

and the corresponding LS-objective function is

$$\mathcal{F}_{LS}(\Pi) = \sum_{a \in \pi_1} \|c_1 - a\|^2 + \sum_{a \in \pi_2} \|c_2 - a\|^2.$$

So the value of the objective function $\mathcal{F}_{LS}$ is obtained by adding the sum of LS-distances of elements of cluster π_1 to its centroid c_1, and the sum of LS-distances of elements of cluster π_2 to its centroid c_2.

π_1	π_2	c_1	c_2	$\mathcal{F}_{LS}(\Pi)$	$\mathcal{G}(\Pi)$
$\{(1,1)\}$	$\{(2,2),(3,1,(3,2)\}$	$(1,1)$	$(2.67, 1.67)$	$0+1.33=1.33$	$1.82+0.60=2.42$
$\{(3,1)\}$	$\{(1,1),(2,2),(3,2)\}$	$(3,1)$	$(2.,1.67)$	$0+2.67=2.67$	$0.81+0.27=1.08$
$\{(3,2)\}$	$\{(1,1),(2,2),(3,1)\}$	$(3,2)$	$(2.,1.3)$	$0+2.67=2.67$	$0.81+0.27=1.08$
$\{(2,2)\}$	$\{(1,1),(3,1),(3,2)\}$	$(2,2)$	$(2.3,1.3)$	$0+3.33=3.33$	$0.31+0.10=0.42$
$\{(1,1),(3,1)\}$	$\{(2,2),(3,2)\}$	$(2,1)$	$(2.5,2.)$	$2+0.5=2.5$	$0.625+0.625=1.25$
$\{(1,1),(3,2)\}$	$\{(2,2),(3,1)\}$	$(2,1.5)$	$(2.5,1.5)$	$2.5+1.=3.5$	$0.125+0.125=0.25$
$\{(1,1),(2,2)\}$	$\{(3,1),(3,2)\}$	$(1.5,1.5)$	$(3.,1.5)$	$1+0.5=1.5$	$1.125+1.125=2.25$

Table 3.7: Partitions, centers and values of objective functions $\mathcal{F}_{LS}$ and $\mathcal{G}$ from Example 3.28

Table 3.7 lists all partitions of the set $\mathcal{A}$, the centroids of respective clusters, and values of objective function $\mathcal{F}_{LS}$. As one can see, the LS-optimal partition is $\Pi^\star = \{\{(1,1)\}, \{(2,2),(3,1),(3,2)\}\}$, since $\mathcal{F}_{LS}$ attains the global minimum at it (see also Figure 3.4).

3.3.2 Dual problem

The next lemma shows that in the case of LS distance-like function, dissipation of the set $\mathcal{A}$ about its center c equals the sum of dissipations of all clusters

π_j, $j = 1, \ldots, k$, about their centers c_j, $j = 1, \ldots, k$, and the weighted sum of squared distances between the center c and cluster centers c_j, where the weights are determined by sizes of sets π_j.

Lemma 3.29. *Let $\mathcal{A} = \{a^i \in \mathbb{R}^n : i = 1, \ldots, m\}$ be a data set, $\Pi = \{\pi_1, \ldots, \pi_k\}$ some k-partition with clusters $\pi_1, \ldots, \pi_k$, and let*

$$c = \frac{1}{m}\sum_{i=1}^{m} a^i, \qquad c_j = \frac{1}{|\pi_j|}\sum_{a^i \in \pi_j} a^i, \quad j = 1, \ldots, k \tag{3.30}$$

be the centroid of the set $\mathcal{A}$ and centroids of clusters $\pi_1, \ldots, \pi_k$, respectively. Then

$$\sum_{i=1}^{m} \|c - a^i\|^2 = \mathcal{F}_{LS}(\Pi) + \mathcal{G}(\Pi), \tag{3.31}$$

where

$$\mathcal{F}_{LS}(\Pi) = \sum_{j=1}^{k} \sum_{a^i \in \pi_j} \|c_j - a^i\|^2, \tag{3.32}$$

$$\mathcal{G}(\Pi) = \sum_{j=1}^{k} |\pi_j| \|c_j - c\|^2. \tag{3.33}$$

Proof. Note first, that c_j satisfies the arithmetic mean property

$$\sum_{a^i \in \pi_j} (c_j - a^i) = 0. \tag{3.34}$$

For an arbitrary $x \in \mathbb{R}^n$ we have

$$\begin{aligned}\sum_{a^i \in \pi_j} \|x - a^i\|^2 &= \sum_{a^i \in \pi_j} \|(x - c_j) + (c_j - a^i)\|^2 \\ &= \sum_{a^i \in \pi_j} \|x - c_j\|^2 + 2\sum_{a^i \in \pi_j} \langle x - c_j, c_j - a^i\rangle + \sum_{a^i \in \pi_j} \|c_j - a^i\|^2.\end{aligned}$$

Since $\sum\limits_{a^i \in \pi_j} \langle x - c_j, c_j - a^i\rangle = \langle x - c_j, \sum\limits_{a^i \in \pi_j} (c_j - a^i)\rangle \overset{(3.34)}{=} 0$, from the previous equality we obtain

$$\sum_{a^i \in \pi_j} \|x - a^i\|^2 = \sum_{a^i \in \pi_j} \|c_j - a^i\|^2 + |\pi_j| \, \|c_j - x\|^2, \quad j = 1, \ldots, k. \tag{3.35}$$

Substituting $c = \frac{1}{m}\sum\limits_{i=1}^{m} a^i$ for x into (3.35) and adding all equations, we obtain (3.31). □

The objective function $\mathcal{F}_{LS}$ occurs in (3.31) naturally, and this formula shows that the total dissipation of elements of $\mathcal{A}$ about its centroid c can be expressed as the sum of two objective functions $\mathcal{F}_{LS}$ and $\mathcal{G}$.

As in Section 3.2.2, using Lemma 3.29, one can show that the following theorem holds true [3, 35].

Theorem 3.30. *Using the notation as in Lemma 3.29, there exists a partition* $\Pi^\star \in \mathcal{P}(\mathcal{A};k)$ *such that*

$$(i) \quad \Pi^\star \in \operatorname*{arg\,min}_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{F}_{LS}(\Pi) = \operatorname*{arg\,max}_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{G}(\Pi),$$

$$(ii) \quad \min_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{F}_{LS}(\Pi) = \mathcal{F}_{LS}(\Pi^\star) \quad \textit{and} \quad \max_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{G}(\Pi) = \mathcal{G}(\Pi^\star),$$

where $\mathcal{G}(\Pi^\star) = \sum_{i=1}^{m} \|c - a^i\|^2 - \mathcal{F}_{LS}(\Pi^\star)$.

This means that in order to find the LS-optimal partition, instead of minimizing the function $\mathcal{F}_{LS}$ defined by (3.32), one can solve the problem of maximizing the function $\mathcal{G}$

$$\operatorname*{arg\,max}_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{G}(\Pi), \qquad \mathcal{G}(\Pi) = \sum_{j=1}^{k} |\pi_j|\, \|c_j - c\|^2. \tag{3.36}$$

The optimization problem (3.36) is called the ***dual problem*** for the optimization problem $\operatorname*{arg\,min}_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{F}_{LS}(\Pi)$.

One can say that the LS-optimal partition has the property that the sum of dissipation of cluster elements (the sum over all clusters of sums of LS-distances between cluster elements and respective centroids) is minimal, and at the same time the clusters a maximally separated. In this way one achieves the best inner compactness and separation between clusters.

Example 3.31. In Example 3.28 one can consider also the corresponding dual problem.

Particularly, in this case the formula (3.31) becomes

$$\sum_{i=1}^{m} \|c-a^i\|^2 = \Big(\sum_{a\in\pi_1} \|c_1-a\|^2 + \sum_{a\in\pi_2} \|c_2-a\|^2\Big) + \big(m_1\,\|c_1-c\|^2 + m_2\,\|c_2-c\|^2\big),$$

and the dual optimization problem (3.36) becomes

$$\operatorname*{arg\,max}_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{G}(\Pi), \qquad \mathcal{G}(\Pi) = m_1\,\|c_1 - c\|^2 + m_2\,\|c_2 - c\|^2.$$

For each 2-partition in Table 3.7, page 46, the values of the dual objective function $\mathcal{G}$ are shown in blue. As can be seen, $\mathcal{G}$ attains its maximal value at the partition $\Pi^\star = \{\{(1,1)\}, \{(2,2),(3,1),(3,2)\}\}$, the same one at which $\mathcal{F}_{LS}$, given by (3.29), attained the minimal value.

Example 3.32. The set $\mathcal{A} = \{a^i = (x_i, y_i) : i = 1, \ldots, 8\} \subset \mathbb{R}^2$ is given by the following table:

i	1	2	3	4	5	6	7	8
x_i	1	4	4	4	7	8	8	10
y_i	3	5	7	9	1	6	10	8

Applying the LS distance-like function for 2-partitions

$$\Pi_1 = \{\{a^1, a^2, a^5\}, \{a^3, a^4, a^6, a^7, a^8\}\},$$
$$\Pi_2 = \{\{a^1, a^2, a^3, a^5\}, \{a^4, a^6, a^7, a^8\}\},$$

depicted in 3.5 with clusters colored blue and red respectively, determine the centroids and corresponding values of objective functions $\mathcal{F}_{LS}$ and $\mathcal{G}$, and based on this, identify the partition being closer to the optimal one.

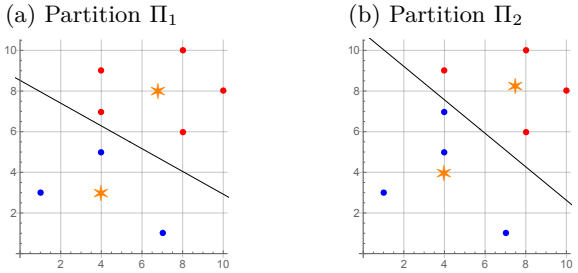


Figure 3.5: Two partitions of the set $\mathcal{A}$ from Example 3.32

As for the partition Π_1 we obtain $c_1 = (4,3)$, $c_2 = (6.8,8)$, $\mathcal{F}_{LS} = 26 + 38.8 = 64.8$ and $\mathcal{G} = 61.575$, and for partition Π_2, $c_1 = (4,4)$, $c_2 = (7.5, 8.25)$, $\mathcal{F}_{LS} = 38 + 27.75 = 65.75$ and $\mathcal{G} = 60.625$. Therefore the 2-partition Π_1 is closer to the LS-optimal one. Check, applying the *Mathematica*-module `WKMeans[]`, whether this is the globally LS-optimal 2-partition. Note (formula (3.2)) that in this case, in total there are $2^7 - 1 = 127$ different 2-partitions.

Exercise 3.33. Let the set $\mathcal{A} = \{a^i = (x_i, y_i) : i = 1, \ldots, m\}$, depicted in Figure 3.6, be given by the following table:

i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	1	2	4	4	5	6	7	8	8	8	9	10
y_i	3	1	5	9	7	1	5	2	6	10	4	8

Determine at which of the two 3-partitions shown in Figure 3.6 does the LS-objective function $\mathcal{F}_{LS}$, given by (3.29), attain the smaller value.

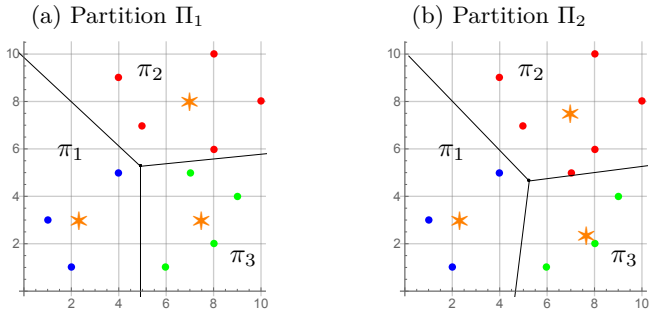


Figure 3.6: Comparison of the two partitions in Exercise 3.33

Solution:

$$\Pi_1 = \big\{\{a^1, a^2, a^3\},\ \{a^4, a^5, a^9, a^{10}, a^{12}\},\ \{a^6, a^7, a^8, a^{11}\}\big\} \qquad \text{(Figure 3.6a)}$$

$$\Pi_2 = \big\{\{a^1, a^2, a^3\},\ \{a^4, a^5, a^7, a^9, a^{10}, a^{12}\},\ \{a^6, a^8, a^{11}\}\big\} \qquad \text{(Figure 3.6b)}$$

Π_1 : $c_1 = (2.33, 3)$, $c_2 = (7, 8)$, $c_3 = (7.5, 3)$;
$\mathcal{F}_{LS} = 12.67 + 34 + 15 = 61.67$; $\mathcal{G} = 127.25$,

Π_2 : $c_1 = (2.33, 3)$, $c_2 = (7, 7.5)$, $c_3 = (7.67, 2.33)$;
$\mathcal{F}_{LS} = 12.67{+}41.5{+}9.33{=}63.5$; $\mathcal{G} = 125.42$.

Hence, smaller value of LS-objective function $\mathcal{F}_{LS}$ (and larger value of the dual function $\mathcal{G}$) is attained at the 3-partition Π_1, and therefore it is closer to the LS-optimal one. Try, using the *Mathematica*-module `WKMeans[]` with various initial partitions, to find a better 3-partition.

3.3.3 Least absolute deviation principle

Let $\mathcal{A} \subset \mathbb{R}^n$ be a set, $\Pi = \{\pi_1, \ldots, \pi_k\}$ some k-partition, and $d_1 \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, given by $d_1(x, y) = \|x - y\|_1$, the ℓ_1 metric function. The centers

$c_1, \ldots, c_k$ of clusters $\pi_1, \ldots, \pi_k$ are determined by

$$\begin{aligned} c_j = \operatorname{med}(\pi_j) &= \big(\operatorname*{med}_{a\in\pi_j} a_1, \ldots, \operatorname*{med}_{a\in\pi_j} a_n\big) \in \operatorname{Med}(\pi_j) \\ &= \big(\operatorname*{Med}_{a\in\pi_j} a_1, \ldots, \operatorname*{Med}_{a\in\pi_j} a_n\big) = \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \sum_{a\in\pi_j} \|x-a\|_1 \end{aligned} \tag{3.37}$$

where $\operatorname*{med}_{a\in\pi_j} a_\ell$ denote medians of the ℓ-th components of all clusters π_j, $\ell = 1, \ldots, n$. The ℓ_1 metric objective function is, in this case, defined as

$$\mathcal{F}_1(\Pi) = \sum_{j=1}^{k} \sum_{a\in\pi_j} \|c_j - a\|_1 \,. \tag{3.38}$$

Exercise 3.34. Show that using the ℓ_1 metric function, the globally optimal 2-partition from Example 3.28 is $\{\{(1,1),(3,2)\}, \{(2,2),(3,1)\}\}$ with cluster centers being $c_1 = (2, 1.5)$ and $c_2 = (2.5, 1.5)$, and the value of objective function $\mathcal{F}_1$ being 5.

Exercise 3.35. Use the least absolute deviation principle to the partitions in Exercise 3.33.

Example 3.36. The set $\mathcal{A} = \{a^i = (x_i, y_i) \in \mathbb{R}^2 : i = 1, \ldots, 10\}$ is given by the following table:

i	1	2	3	4	5	6	7	8	9	10
x_i	2	3	4	4	5	6	6	8	8	9
y_i	9	3	5	7	8	2	6	4	6	5

Determine at which of the following two 3-partitions does the ℓ_1-objective function (3.38) attain the smaller value.

$$\Pi_1 = \{\{a^2, a^3, a^6\}, \{a^1, a^4, a^5\}, \{a^7, a^8, a^9, a^{10}\}} \qquad \text{(Figure 3.7a)}$$

$$\Pi_2 = \{\{a^2, a^6\}, \{a^1, a^3, a^4, a^5, a^7\}, \{a^8, a^9, a^{10}\}\}\,. \qquad \text{(Figure 3.7b)}$$

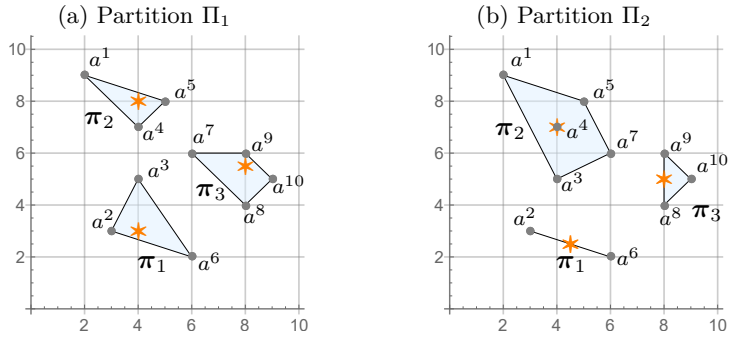


Figure 3.7: Comparison of the two partitions in Exercise 3.36

The following table lists ℓ_1-centers of clusters and values of the objective function for both partitions. One can see that Π_1 is the better partition because the objection function attains smaller value at Π_1 than at Π_2.

	c_1	c_2	c_3	$\mathcal{F}_1$
Π_1	$(4, 3)$	$(4, 8)$	$(8, 5.5)$	$(1+2+3)+(3+1+1)+(2.5+1.5+0.5+1.5) = 17$
Π_2	$(4.5, 2.5)$	$(4, 7)$	$(8, 5)$	$(2+2)+(4+2+0+2+3)+(1+1+1) = 18$

Remark 3.37. In a similar way as in Section 3.2.4, where we have considered the clustering problem for one-dimensional weighted data, we could proceed in the case of two- and more-dimensional data.

3.4 Objective function $F(c_1, \ldots, c_k) = \sum_{i=1}^{m} \min_{1 \le j \le k} d(c_j, a^i)$

The objective function GOP (3.5) is not suitable for applying standard optimization methods, since the independent variable is a partition. Therefore we are going to reformulate GOP (3.5) in such a way that the objective function becomes an ordinary function of several variables. For a survey of most popular methods for finding partitions, see [38].

As was noted in the k-means algorithm 3.9 on page 34, we are going, for the given set $c_1, \ldots, c_k \in \mathbb{R}^n$ of centers, to split the set $\mathcal{A}$ into k clusters $\pi(c_1), \ldots, \pi(c_k)$,[2] such that the cluster π_j contains those elements of the set $\mathcal{A}$ which are closest to the center c_j, so that for every $a^i \in \mathcal{A}$ we have

$$a^i \in \pi_j(c_j) \Leftrightarrow d(c_j, a^i) \le d(c_s, a^i) \text{ for all } s = 1, \ldots, k. \tag{3.39}$$

[2] Notice that the cluster $\pi(c_j)$ depends on neighboring clusters, and that the notation $\pi(c_j)$ refers to the fact that the cluster $\pi(c_j)$ is associated with the center c_j.

In addition, one has to take care that each element of $\mathcal{A}$ belongs to a single cluster. This principle, which we call *minimal distance principle*, results in a partition $\Pi = \{\pi_1, \dots, \pi_k\}$ with clusters $\pi_1, \dots, \pi_k$.

The problem of finding the optimal partition of the set $\mathcal{A}$ can therefore be reduced to the following `GOP` (see also [35]):

$$\arg\min_{c \in \mathbb{R}^{n \times k}} F(c), \qquad F(c) = \sum_{i=1}^{m} \min_{1 \le j \le k} d(c_j, a^i), \tag{3.40}$$

where $c \in \mathbb{R}^{n \times k}$ is the concatenation of vectors $c_1, \dots, c_k$. The function F is non-negative, symmetric, non-differentiable, non-convex, but Lipschitz-continuous.

The following theorem shows that, when using the LS distance-like function, the function F defined by (3.40) is Lipschitz-continuous. Similarly, in [27] it is shown that this function is Lipschitz-continuous also in the case of ℓ_1 metric function. This is an important property of the function F since it allows one to use the global optimization algorithm `DIRECT` [8, 15, 22, 29].

Theorem 3.38. *Let $\mathcal{A} = \{a^i \in \mathbb{R}^n : i = 1, \dots, m\} \subset \Delta$, where $\Delta = \{x \in \mathbb{R}^n : \alpha_i \le x_i \le \beta_i\}$, for some $\alpha = (\alpha_1, \dots, \alpha_n)$ and $\beta = (\beta_1, \dots, \beta_n) \in \mathbb{R}^n$. The function $F : \Delta^k \to \mathbb{R}_+$ defined by*

$$F(c) = \sum_{i=1}^{m} \min_{j=1,\dots,k} \|c_j - a^i\|^2$$

is Lipschitz continuous.

In order to prove this theorem, we are going to approximate the function F up to an $\epsilon > 0$ by a differentiable (smooth) function F_ϵ. To do this, we will need some theoretical preparations.

Lemma 3.39. *The function $\psi : \mathbb{R}^n \to \mathbb{R}$, defined by $\psi(x) = \ln(e^{x_1} + \cdots + e^{x_n})$, is a convex function.*

Proof. One has to show that for all $x, y \in \mathbb{R}^n$ and $\lambda \in [0, 1]$

$$\psi(\lambda\, x + (1 - \lambda)\, y) \le \lambda\, \psi(x) + (1 - \lambda)\, \psi(y), \tag{3.41}$$

i.e.

$$\psi(\alpha x + \beta y) \le \alpha\, \psi(x) + \beta\, \psi(y), \tag{3.42}$$

where $\alpha, \beta > 0$ are such that $\alpha + \beta = 1$.

Let the numbers $p := \frac{1}{\alpha}$ and $q := \frac{1}{\beta}$ be such that $\frac{1}{p} + \frac{1}{q} = 1$. Since $\alpha + \beta = 1$, one of the numbers α, β has to be smaller than 1, hence one of the numbers p, q has to be larger than 1. Let $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n) \in \mathbb{R}^n$. Applying the Hölder inequality[3] (see [36]) to vectors

$$a = (e^{\alpha x_1}, \dots, e^{\alpha x_n}) \text{ and } b = (e^{\beta y_1}, \dots, e^{\beta y_n}) \in \mathbb{R}^n,$$

we obtain

$$|\langle a, b\rangle| \le \Big(\sum_{i=1}^{n} (e^{\alpha x_i})^p\Big)^{1/p} \Big(\sum_{i=1}^{n} (e^{\beta y_i})^q\Big)^{1/q},$$

i.e.

$$\sum_{i=1}^{n} e^{\alpha x_i + \beta y_i} \le \Big(\sum_{i=1}^{n} e^{x_i}\Big)^{\alpha} \Big(\sum_{i=1}^{n} e^{y_i}\Big)^{\beta}.$$

By taking the logarithm we obtain the required inequality (3.41). □

Corollary 3.40. *Let $A \in \mathbb{R}^{n\times n}$ be a square matrix, $b \in \mathbb{R}^n$ a vector and $\psi\colon \mathbb{R}^n \to \mathbb{R}$ defined by $\psi(x) = \ln(e^{x_1} + \cdots + e^{x_n})$. Then $\Phi(x) = \psi(Ax + b)$ is a convex function.*

Proof. As in the proof of previous lemma, it suffices to show that for arbitrary $x, y \in \mathbb{R}^n$ and $\alpha, \beta > 0$, such that $\alpha + \beta = 1$, one has

$$\Phi(\alpha\, x + \beta\, y) \le \alpha\, \Phi(x) + \beta\, \Phi(y).$$

Since

$$\begin{aligned}
\Phi(\alpha\, x + \beta\, y) &= \psi(A(\alpha\, x + \beta\, y) + b) = \psi(\alpha\, Ax + \beta\, Ay + b)\\
&= \psi(\alpha\, Ax + \alpha\, b + \beta\, Ay + \beta\, b - (\alpha + \beta)\, b + b)\\
&= \psi(\alpha\, (Ax + b) + \beta\, (Ay + b))\\
&\le \alpha\, \Phi(x) + \beta\, \Phi(y),
\end{aligned}$$

the required inequality follows. □

[3] For two vectors $a, b \in \mathbb{R}^n$, and real numbers p and q such that $\frac{1}{p} + \frac{1}{q} = 1$, $p > 1$, the Hölder inequality states that $\sum_{i=1}^{n} |a_i b_i| \le \|a\|_p \, \|b\|_q$, i.e.

$$\sum_{i=1}^{n} |a_i b_i| \le \Big(\sum_{i=1}^{n} |a_i|^p\Big)^{1/p} \Big(\sum_{i=1}^{n} |b_i|^q\Big)^{1/q}.$$

In particular, for $p = q = 2$ this becomes the well-known Cauchy-Schwarz-Buniakowsky inequality.

Exercise 3.41. Show that $\psi\colon \mathbb{R}^n_+ \to \mathbb{R}$ defined by $\psi(x)=\ln(\frac{1}{x_1}+\cdots+\frac{1}{x_n})$ is a convex function.

Lemma 3.42. *For every $\epsilon>0$, the function $\psi_\epsilon\colon \mathbb{R}\to\mathbb{R}_+$ defined by*

$$\psi_\epsilon(x)=\epsilon\ln\left(e^{-\frac{x}{\epsilon}}+e^{\frac{x}{\epsilon}}\right)=\epsilon\ln\left(2\operatorname{ch}\tfrac{x}{\epsilon}\right) \tag{3.43}$$

is a convex function of class $C^\infty(\mathbb{R})$, and it satisfies

$$0<\psi_\epsilon(x)-|x|\le\epsilon\ln 2 \text{ for all } x\in\mathbb{R}, \tag{3.44}$$

$$\psi_\epsilon'(x)=\operatorname{th}\tfrac{x}{\epsilon},\qquad \psi_\epsilon''(x)=\frac{1}{\epsilon\operatorname{ch}^2\frac{x}{\epsilon}},\qquad \arg\min_{x\in\mathbb{R}}\psi_\epsilon(x)=0, \tag{3.45}$$

and the equality in (3.44) holds true if and only if $x=0$.

a) Functions $x\mapsto|x|$ and $x\mapsto\psi_\epsilon(x)$ b) Function $u\mapsto\frac{2\operatorname{ch}u}{e^{|u|}}$

Figure 3.8: Smooth approximation of the function $x\mapsto|x|$

Proof. Putting $n=2$, $x_1=-\frac{x}{\epsilon}$ and $x_2=\frac{x}{\epsilon}$, convexity of the function ψ_ϵ follows from Lemma 3.39. In order to prove (3.44), notice that

$$\begin{aligned}\psi_\epsilon(x)-|x| &= \epsilon\,\ln\left(2\operatorname{ch}\tfrac{x}{\epsilon}\right)-\epsilon\,\tfrac{|x|}{\epsilon}\\ &= \epsilon\left(\ln\left(2\operatorname{ch}\tfrac{x}{\epsilon}\right)-\ln\exp\tfrac{|x|}{\epsilon}\right)=\epsilon\ln\frac{2\operatorname{ch}\frac{x}{\epsilon}}{\exp\frac{|x|}{\epsilon}}.\end{aligned}$$

Since for every $u\in\mathbb{R}$ (see Exercise 3.43) $1<\frac{2\operatorname{ch}u}{\exp|u|}\le 2$, and since the logarithmic function is monotonous, the previous equality implies

$$\epsilon\ln 1<\psi_\epsilon(x)-|x|\le\epsilon\ln 2.$$

Formulas (3.45) follow directly □

Exercise 3.43. Prove that for every $u\in\mathbb{R}$ the following holds true (see Figure 3.8b)

$$1<\frac{2\operatorname{ch}u}{e^{|u|}}\le 2.$$

In view of (3.44), note that the function $x \mapsto |x|$, $x \in \mathbb{R}$, can be approximated by the function ψ_ϵ (see Figure 3.8a).

In general, the non-differentiable function $f\colon \mathbb{R}^k \to \mathbb{R}$, defined as $f(z) = \max\limits_{j=1,\dots,k} z_j$, can be approximated by the differentiable function

$$\psi_\epsilon(z) = \psi_\epsilon(z_1, \dots, z_k) = \epsilon \ln \sum_{j=1}^{k} \exp\left(\tfrac{z_j}{\epsilon}\right). \tag{3.46}$$

Namely,

$$\begin{aligned}
\psi_\epsilon(z) - f(z) &= \epsilon \ln \sum_{j=1}^{k} \exp\left(\frac{z_j}{\epsilon}\right) - \epsilon \frac{\max\limits_{i=1,\dots,k} z_i}{\epsilon} \\
&= \epsilon\Big(\ln \sum_{j=1}^{k} \exp\left(\frac{z_j}{\epsilon}\right) - \ln \exp \frac{\max z_i}{\epsilon}\Big) \\
&= \epsilon \ln \frac{\sum_{j=1}^{k} \exp\left(\frac{z_j}{\epsilon}\right)}{\exp \frac{\max z_i}{\epsilon}} = \epsilon \ln \sum_{j=1}^{k} \exp \frac{z_j - \max z_i}{\epsilon} \le \epsilon \ln \sum_{j=1}^{k} e^0 = \epsilon \ln k.
\end{aligned}$$

Moreover, since $\min\limits_{j=1,\dots,k} z_j = -\max\limits_{j=1,\dots,k}(-z_j)$, we can use this result to approximate the function $F(c_1, \dots, c_k) = \sum\limits_{i=1}^{m} \min\limits_{1 \le j \le k} d(c_j, a^i)$ by

$$F_\epsilon(c_1, \dots, c_k) = -\epsilon \sum_{i=1}^{m} \ln \sum_{j=1}^{k} \exp\left(-\tfrac{d(c_j, a^i)}{\epsilon}\right). \tag{3.47}$$

We are now ready to prove Theorem 3.38.

Proof of Theorem 3.38. In accordance with (3.47), define the auxiliary function

$$F_\varepsilon(u) = -\varepsilon \sum_{i=1}^{m} \ln \sum_{j=1}^{k} \exp\left(-\tfrac{\|c_j - a^i\|^2}{\varepsilon}\right).$$

Then, according to [17], the following holds true:

$$0 \le F(u) - F_\varepsilon(u) \le \varepsilon\, m \ln k$$

Therefore

$$\begin{aligned}
|F(u) - F(v)| &= |(F(u) - F_\varepsilon(u)) + (F_\varepsilon(v) - F(v)) + (F_\varepsilon(u) - F_\varepsilon(v))| \\
&\le |F(u) - F_\varepsilon(u)| + |F_\varepsilon(v) - F(v)| + |F_\varepsilon(u) - F_\varepsilon(v)| \\
&\le 2\,\varepsilon\, m \ln k + |F_\varepsilon(u) - F_\varepsilon(v)|.
\end{aligned} \tag{3.48}$$

Since

$$\frac{\partial F_\varepsilon(x)}{\partial x_p}=2\sum_{i=1}^{m}\frac{(x_p-a^i)\exp\left(-\frac{\|x_p-a^i\|^2}{\varepsilon}\right)}{\sum_{j=1}^{k}\exp\left(-\frac{\|x_j-a^i\|^2}{\varepsilon}\right)},$$

we obtain

$$\begin{aligned}\left\|\tfrac{\partial F_\varepsilon(x)}{\partial x_p}\right\| &\le 2\sum_{i=1}^{m}\|x_p-a^i\|\le 2\sum_{i=1}^{m}\max_{j=1,\dots,m}\|a^i-a^j\|\\ &\le 2\,m\max_{i,j\in\{1,\dots,m\}}\|a^i-a^j\|,\quad p=1,\dots,k,\end{aligned}$$

i.e. the gradient $\nabla F_\varepsilon(x)$ is continuous and bounded on Δ^k. Using the Lagrange intermediate value theorem for the function F_ϵ on Δ^k, we conclude that there exists an $L>0$ (not depending on ε) such that

$$|F_\varepsilon(u)-F_\varepsilon(v)|\le L\|u-v\|,\quad u,v\in\Delta^k.$$

Finally, for $\varepsilon\to 0^+$, (3.48) implies that $|F(u)-F(v)|\le L\,\|u-v\|$. □

The following lemma and theorem show the connection between the objective function $\mathcal{F}$ defined by (3.5) and the objective function F defined by (3.40).

Lemma 3.44. *Let $\mathcal{A}=\{a^i\in\mathbb{R}^n : i=1,\dots,m\}$ be a finite set in $\mathbb{R}^n$, $z_1,\dots,z_k\in\mathbb{R}^n$ a set of mutually distinct points, and $d\colon\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}_+$ a distance-like function. In addition, let $\Pi=\{\pi_1(z_1),\dots,\pi_k(z_k)\}$ be a partition whose clusters were obtained by minimal distance principle and let $c_j\in\operatorname{arg\,min}_{x\in\mathbb{R}^n}\sum_{a\in\pi_j} d(x,a)$, $j=1,\dots,k$, be their centers. Then*

$$F(z_1,\dots,z_k)\overset{(\star)}{\ge}\mathcal{F}(\Pi)\overset{(\star\star)}{\ge}F(c_1,\dots,c_k),\tag{3.49}$$

while inequalities $(\star)$ and $(\star\star)$ turn to equalities if and only if $z_j=c_j$ for every $j=1,\dots,k$.

Proof. In order to prove inequality $(\star)$ we split $\sum_{i=1}^{m}$ into k sums $\sum_{j=1}^{k}\sum_{a\in\pi_j}$.

$$\begin{aligned} F(z_1,\dots,z_k) &= \sum_{i=1}^{m}\min\{d(z_1,a^i),\dots,d(z_k,a^i)\} \\ &= \sum_{j=1}^{k}\sum_{a^i\in\pi_j}\min\{d(z_1,a^i),\dots,d(z_k,a^i)\} \\ &= \sum_{j=1}^{k}\sum_{a^i\in\pi_j} d(z_j,a^i) \\ &\overset{(\star)}{\geq} \sum_{j=1}^{k}\sum_{a^i\in\pi_j} d(c_j,a^i) = \mathcal{F}(\{\pi_1,\dots,\pi_k\}). \end{aligned}$$

To prove $(\star\star)$, first notice that for every $a\in\pi_j$ on has

$$d(c_j,a)\geq \min\{d(c_1,a),\dots,d(c_k,a)\}.$$

Therefore,

$$\begin{aligned} \mathcal{F}(\{\pi_1,\dots,\pi_k\}) &= \sum_{j=1}^{k}\sum_{a^i\in\pi_j} d(c_j,a^i) \\ &\geq \sum_{j=1}^{k}\sum_{a^i\in\pi_j}\min\{d(c_1,a^i),\dots,d(c_k,a^i)\} \\ &= \sum_{i=1}^{m}\min\{d(c_1,a^i),\dots,d(c_k,a^i)\} = F(c_1,\dots,c_k), \end{aligned}$$

showing $(\star\star)$. □

Theorem 3.45. *Let $\mathcal{A}=\{a^i\in\mathbb{R}^n : i=1,\dots,m\}\subset\mathbb{R}^n$. Then:*

(i) $c^\star=(c_1^\star,\dots,c_k^\star)^T\in \underset{c_1,\dots,c_k\in\mathbb{R}^n}{\arg\min} F(c_1,\dots,c_k)$ *if and only if*

$$\Pi^\star=\{\pi_1^\star(c_1^\star),\dots,\pi_k^\star(c_k^\star)\}\in \underset{\Pi\in\mathcal{P}(\mathcal{A};k)}{\arg\min}\,\mathcal{F}(\Pi)\,,$$

(ii) $\min_{c_1,\dots,c_k\in\mathbb{R}^n} F(c_1,\dots,c_k)=\min_{\Pi\in\mathcal{P}(\mathcal{A};k)}\mathcal{F}(\Pi)\,.$

Proof. *(a)* Let $c^\star=(c_1^\star,\dots,c_k^\star)^T\in \underset{c_1,\dots,c_k\in\mathbb{R}^n}{\arg\min} F(c_1,\dots,c_k)$. Denote by $\pi_j^\star$ the corresponding clusters obtained by minimal distance principle, and let $\Pi^\star=\{\pi_1^\star,\dots,\pi_k^\star\}$. According to Lemma 3.44

$$F(c^\star)=\mathcal{F}(\Pi^\star). \tag{3.50}$$

We claim that

$$\Pi^\star \in \operatorname*{arg\,min}_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{F}(\Pi). \tag{3.51}$$

Namely, if there existed a partition $\mathcal{N}^\star=\{\nu_1^\star,\dots,\nu_k^\star\}\in\mathcal{P}(\mathcal{A};k)$ with cluster centers $\zeta^\star=(\zeta_1^\star,\dots,\zeta_k^\star)$ such that $\mathcal{F}(\mathcal{N}^\star)<\mathcal{F}(\Pi^\star)$, we would have

$$F(\zeta^\star)\overset{\text{Lemma }3.44}{=}\mathcal{F}(\mathcal{N}^\star)<\mathcal{F}(\Pi^\star)\overset{\text{Lemma }3.44}{=}F(c^\star),$$

which is not possible since $c^\star\in\operatorname*{arg\,min}_{c\in\mathbb{R}^{n\times k}} F(c)$.

(*b*) Let $\Pi^\star=\{\pi_1^\star,\dots,\pi_k^\star\}\in\operatorname*{arg\,min}_{\Pi\in\mathcal{P}(\mathcal{A};k)}\mathcal{F}(\Pi)$. Denote by $c^\star=(c_1^\star,\dots,c_k^\star)$ the centers of clusters $\pi_1^\star,\dots,\pi_k^\star$. According to Lemma 3.44

$$F(c^\star)=\mathcal{F}(\Pi^\star). \tag{3.52}$$

We claim that

$$c^\star\in\operatorname*{arg\,min}_{c\in\mathbb{R}^{n\times k}} F(c). \tag{3.53}$$

Namely, if there existed a $\zeta^\star=(\zeta_1^\star,\dots,\zeta_k^\star)$ such that $F(\zeta^\star)<F(c^\star)$, then the partition $\mathcal{N}^\star(\zeta^\star)$ would satisfy

$$\mathcal{F}(\Pi^\star)\overset{\text{Lemma }3.44}{=}F(c^\star)>F(\zeta^\star)\overset{\text{Lemma }3.44}{=}\mathcal{F}(\mathcal{N}^\star),$$

which is not possible since $\Pi^\star\in\operatorname*{arg\,min}_{\Pi\in\mathcal{P}(\mathcal{A};k)}\mathcal{F}(\Pi)$. □

Example 3.46. Let $\mathcal{A}=\{1,3,4,8\}$ be a set with $m=4$ data. Table 3.8 lists some values of objective functions $\mathcal{F}_{LS}$ and F_{LS} supporting claims of Lemma 3.44 and Theorem 3.45. For the optimal partition the inequality $(\star)$ becomes equality, while z_1, z_2 coincide with cluster centers (the fourth row).

	z_1	z_2	$F_{LS}(z_1,z_2)$	π_1	π_2	c_1	c_2	$\mathcal{F}_{LS}$	$F_{LS}(c_1,c_2)$
1.	1	4	17	{1}	{3,4,8}	1	5	14	14
2.	1	5	14	{1,3}	{4,8}	2	6	10	10
3.	3	7	6	{1,3,4}	{8}	$\frac{8}{3}$	8	$\frac{14}{3}$	$\frac{14}{3}$
4.	$\frac{8}{3}$	8	$\frac{14}{3}$	{1,3,4}	{8}	$\frac{8}{3}$	8	$\frac{14}{3}$	$\frac{14}{3}$

Table 3.8: Comparing values of objective functions $\mathcal{F}_{LS}$ and F_{LS} for $\mathcal{A}=\{1,3,4,8\}$

Example 3.47. Let $\mathcal{A} = \{16, 11, 2, 9, 2, 8, 15, 19, 8, 17\}$ be a set with $m = 10$ data. Table 3.9 lists some values of objective functions $\mathcal{F}_1$ and F_1 supporting claims of Lemma 3.44 and Theorem 3.45. In particular, pay attention to the third row showing sharp inequality $(\star\star)$.

	z_1	z_2	$F_1(z_1, z_2)$	π_1	π_2	(c_1, c_2)	$\mathcal{F}_1$	$F_1(c_1, c_2)$
1.	2	6	55	{2,2}	{8,8,9,11,15,16,17,19}	{2,13}	31	31
2.	2	13	31	{2,2}	{8,8,9,11,15,16,17,19}	{2,13}	31	31
3.	3	15	29	{2,2,8,8,9}	{11,15,16,17,19}	{8,16}	23	21
4.	6	16	25	{2,2,8,8,9,11}	{15,16,17,19}	$\{8, \frac{33}{2}\}$	21	21
5.	8	16	21	{2,2,8,8,9,11}	{15,16,17,19}	$\{8, \frac{33}{2}\}$	21	21

Table 3.9: Comparing values of objective functions $\mathcal{F}_1$ and F_1

Exercise 3.48. Carry out a similar verification as in Example 3.46 using the ℓ_1 metric function, and also a similar verification as in Example 3.47 using the LS distance-like function.

Using Theorem 3.45 we are now ready to prove Theorem 3.6, stating that increasing the number of clusters does not increase the value of the objective function $\mathcal{F}$.

Proof of Theorem 3.6. Let $\hat{c} = (\hat{c}_1, \ldots, \hat{c}_{k-1})$ be the centers of the optimal $(k-1)$-partition $\Pi^{(k-1)}$, and $c^\star = (c_1^\star, \ldots, c_k^\star)$ be the centers of the optimal k-partition $\Pi^{(k)}$. Take a $\zeta \in \mathbb{R}^n \setminus \{\hat{c}_1, \ldots, \hat{c}_{k-1}\}$ and let

$$\delta_{k-1}^i := \min_{1 \le s \le k-1} d(\hat{c}_s, a^i), \quad i = 1, \ldots, m.$$

Then

$$\begin{aligned}
\mathcal{F}(\Pi^{(k-1)}) &\overset{\text{Thm 3.45}}{=} F(\hat{c}) = \sum_{i=1}^m \min\{d(\hat{c}_1, a^i), \ldots, d(\hat{c}_{k-1}, a^i)\} = \sum_{i=1}^m \delta_{k-1}^i \\
&\ge \sum_{i=1}^m \min\{\delta_{k-1}^i, d(\zeta, a^i)\} \qquad [\Pi^{(k)} \text{ being optimal } k\text{-partition}] \\
&\ge \sum_{i=1}^m \min\{d(c_1^\star, a^i), \ldots, d(c_k^\star, a^i)\} \\
&= F(c^\star) \overset{\text{Thm 3.45}}{=} \mathcal{F}(\Pi^{(k)}),
\end{aligned}$$

asserting that increasing the number of clusters in the optimal partition does not increase the value of the objective function. □

Remark 3.49. The above proof of Theorem 3.6 implicitly shows that F is a monotonous function.

Lemma 3.44 and Theorem 3.45 motivates the following definition.

Definition 3.50. Let $\mathcal{A} = \{a^i \in \mathbb{R}^n : i = 1, \ldots, m\}$ be a finite set in $\mathbb{R}^n$, $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ a distance-like function and $\hat{\Pi} = \{\hat{\pi}_1, \ldots, \hat{\pi}_k\}$ a partition whose cluster centers $\hat{c}_1, \ldots, \hat{c}_k$ are such that the function F attains a local minimum at $(\hat{c}_1, \ldots, \hat{c}_k)$. The partition $\hat{\Pi}$ is called a ***locally optimal k-partition*** (`LOPart`) of the set $\mathcal{A}$ provided that

$$\mathcal{F}(\hat{\Pi}) = F(\hat{c}_1, \ldots, \hat{c}_k). \tag{3.54}$$

Bibliography

[1] F. Aurenhammer, R. Klein, *Voronoi diagrams*, In: J. Sack, G. Urrutia, editors, *Handbook of Computational Geometry, Chapter V*. Elsevier Science Publishing, 2000, 201–290.

[2] R. J. Boscovich, *De litteraria expeditione per pontificiam ditionem, et synopsis amplioris operis, ac habentur plura eius ex exemplaria etiam sensorum impressa*, Bononienci Scientiarum et Artium Znstituto Atque Academia Commentarrii, **4**(1757) 353–396.

[3] I. S. Dhillon, Y. Guan, B. Kulis, *Kernel k-means, spectral clustering and normalized cuts*, In: *Proceedings of the 10-th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), August 22–25, 2004, Seattle, Washington, USA*, 2004, 551–556.

[4] Y. Dodge, editor, *Statistical data analysis based on the L1-norm and related methods, Proceedings of the Third International Conference on Statistical Data Analysis Based on the L1-norm and Related Methods*. Elsevier, 1997.

[5] J. Dortet-Bernadet, N. Wicker, *Model-based clustering on the unit sphere with an illustration using gene expression profiles*, Biostatistics, **9(1)**(2008) 66–80.

[6] Z. Drezner, H. W. Hamacher, *Facility Location: Applications and Theory*, Springer, 2004.

[7] B. S. Everitt, S. Landau, M. Leese, *Cluster analysis*, Wiley, London, 2001.

[8] R. Grbić, E. K. Nyarko, R. Scitovski, *A modification of the DIRECT method for Lipschitz global optimization for a symmetric function*, Journal of Global Optimization, **57**(2013) 1193–1212.

[9] C. Gurwitz, *Weighted median algorithms for l_1 approximation*, BIT, **30**(1990) 301–310.

[10] J. Harris, J. L. Hirst, M. Mossinghoff, *Combinatorics and Graph Theory*, Undergraduate Texts in Mathematics. Springer, 2008.

[11] E. M. T. Hendrix, B. G. Tóth, *Introduction to Nonlinear and Global Optimization*, Springer, 2010.

[12] C. Iyigun, A. Ben-Israel, *A generalized Weiszfeld method for the multi-facility location problem*, Operations Research Letters, **38**(2010) 207–214.

[13] F. Jarre, J. Stoer, *Optimierung*, Springer Verlag, Berlin, Heidelberg, 2004.

[14] M. Jiang, *On the sum of distances along a circle*, Discrete Mathematics, **308**(2008) 2038–2045.

[15] D. R. Jones, C. D. Perttunen, B. E. Stuckman, *Lipschitzian optimization without the Lipschitz constant*, Journal of Optimization Theory and Applications, **79**(1993) 157–181.

[16] L. Kaufman, P. J. Rousseeuw, *Finding groups in data: An introduction to cluster analysis*, John Wiley & Sons, Chichester, UK, 2005.

[17] J. Kogan, *Introduction to Clustering Large and High-dimensional Data*, Cambridge University Press, New York, 2007.

[18] K. V. Mardia, P. E. Jupp, *Directional Statistics*, Wiley, 2000.

[19] D. J. Maširević, S. Miodragović, *Geometric median in the plane*, Elemente der Mathematik, **70**(2015) 21–32.

[20] B. Mirkin, *Data clustering for Data Mining*, Chapman & Hall/CRC, 2005.

[21] A. Okabe, B. Boots, K. Sugihara, *Spatial Tessellations: Concepts and Applications of Voronoi diagrams*, John Wiley & Sons, Chichester, UK, 2000.

[22] R. Paulavičius, J. Žilinskas, *Simplicial Global Optimization*, volume X of *Series: Springer Briefs in Optimization*, Springer-Verlag, Berlin, 2014.

[23] P. J. Rousseeuw, M. Hubert, *Robust statistics for outlier detection*, Wiley Interdiscip. Rev. Data Min. Knowl. Discov., **1**(2011) 73–79, DOI: 10.1002/widm.2.

[24] P. J. Rousseeuw, A. M. Leroy, *Robust Regression and Outlier Detection*, Wiley, New York, 2003.

[25] K. Sabo, R. Scitovski, *The best least absolute deviations line – properties and two efficient methods*, ANZIAM Journal, **50**(2008) 185–198.

[26] K. Sabo, R. Scitovski, I. Vazler, *Grupiranje podataka - klasteri*, Osječki matematički list, **10**(2010) 149–178.

[27] K. Sabo, R. Scitovski, I. Vazler, *One-dimensional center-based l_1-clustering method*, Optimization Letters, **7**(2013) 5–22.

[28] K. Sabo, R. Scitovski, I. Vazler, M. Zekić-Sušac, *Mathematical models of natural gas consumption*, Energy Conversion and Management, **52**(2011) 1721–1727.

[29] R. Scitovski, *A new global optimization method for a symmetric lipschitz continuous function and application to searching for a globally optimal partition of a one-dimensional set*, Journal of Global Optimization, (2017), DOI: 10.1007/s10898-017-0510-4.

[30] R. Scitovski, M. B. Alić, *Grupiranje podataka*, Odjel za matematiku, Sveučilište u Osijeku, 2016.

[31] R. Scitovski, S. Kosanović, *Rate of change in economics research*, Economics analysis and workers management, **19**(1985) 65–75.

[32] R. Scitovski, K. Sabo, *Analysis of the k-means algorithm in the case of data points occurring on the border of two or more clusters*, Knowledge-Based Systems, **57**(2014) 1–7.

[33] R. Scitovski, K. Sabo, D. Grahovac, *Globalna optimizacija*, Odjel za matematiku, 2017, https://www.mathos.unios.hr/images/homepages/scitowsk/GOP.pdf.

[34] R. Scitovski, S. Scitovski, *A fast partitioning algorithm and its application to earthquake investigation*, Computers & Geosciences, **59**(2013) 124–131.

[35] H. Späth, *Cluster-Formation und Analyse*, R. Oldenburg Verlag, München, 1983.

[36] J. M. Steele, *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*, Mathematical Association of America, 2004.

[37] M. Teboulle, *A unified continuous optimization framework for center-based clustering methods*, Journal of Machine Learning Research, **8**(2007) 65–102.

[38] S. Theodoridis, K. Koutroumbas, *Pattern Recognition*, Academic Press, Burlington, 2009, 4^{th} edition.

[39] I. Vazler, K. Sabo, R. Scitovski, *Weighted median of the data in solving least absolute deviations problems*, Communications in Statistics - Theory and Methods, **41:8**(2012) 1455–1465.

[40] I. Wolfram Research, *Mathematica*, Wolfram Research, Inc., Champaign, Illinois, 2016, version 11.0 edition.

Index