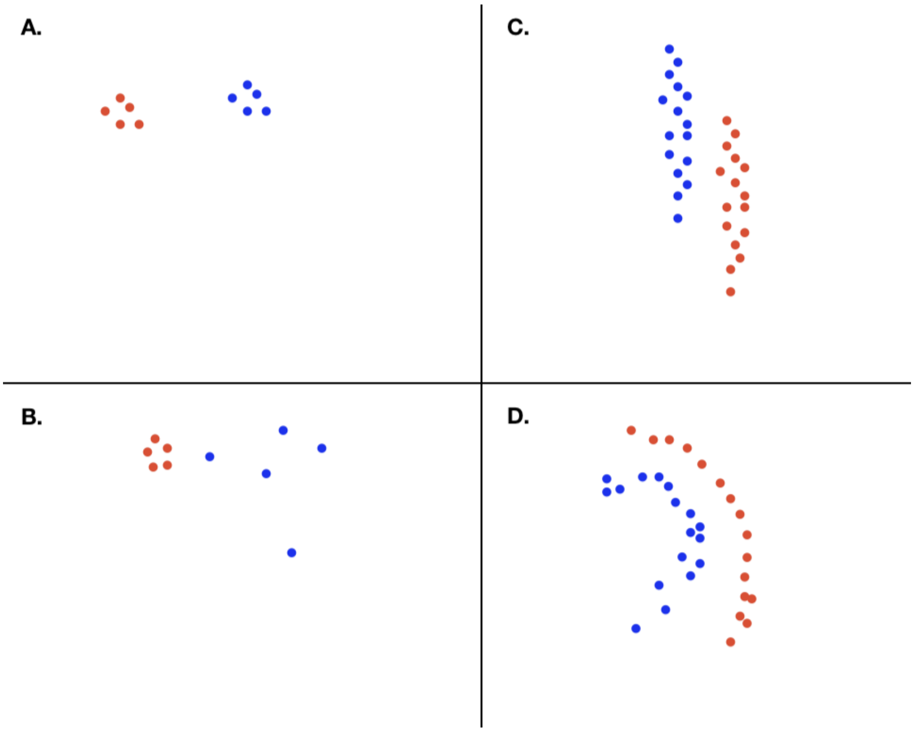
Multivariate Verfahren – Summer Semester 2024 **Exercise sheet 4**

Hannah Schulz-Kümpel June 20th 2024

Unsupervised Learning: Clustering

Question 1:

In the plot below, which of the following options could have produced each clustering (multiple answers are possible): *K-means, Single linkage (hierarchical clustering), Gaussian Mixture Models*.



Solution:

Plot A: K-means, Single linkage & Gaussian Mixture Models

Plot B: Gaussian Mixture Models

Plot C: Single linkage & Gaussian Mixture Models

Plot D: Single linkage

Question 2: Hierarchical Clustering

For four branches of a supermarket chain, the following values are obtained for the characteristics turnover and sales area, each measured in suitable units:

branch	1	2	3	4
turnover	8	5	10	4
sales area	24	22	25	21

Using the squared Euclidean distance as the distance between individual objects both times,

a) Perform a hierarchical clustering with the *Single Linkage* method

b) Perform a hierarchical clustering with the *Zentroid* method.

c) Draw the complete dendrograms for both methods.

Recap - Hierarchical Clustering:

- Given: n points $x_1, \ldots, x_n$
- Clustering: Forming suitable clusters / classes / groups
- Two possible approaches:
  - agglomerative: subclasses are successively combined
  - divisive: Start with all objects in 1 cluster, which is successively split up
- **Agglomerative procedure**: In the first step, all objects form their own cluster. Combine clusters based on distance dimensions until all objects are combined in one cluster.
- $d_{ij} = d(\boldsymbol{x}_i, \boldsymbol{x}_j) \widehat{=}$ distance between points i an j
- $D(C_i, C_j) \widehat{=}$ Distance between clusters C_i and C_j.
- $\mathcal{C}^\nu$ is defined as the partition in the ν-th step.
- $h_\nu \widehat{=}$ Distance between the two clusters merged in step ν (to be entered in the dendrogram).

Solution:

a) Single-Linkage with squared euclidean distance $d_{ij} = ||x_i - x_j||^2$: In step ν, we merge those clusters $C_i,\, C_j \in \mathcal{C}^{(\nu-1)}$ for which the following applies:

$$D(C_i,\, C_j) = (h_\nu =) \min_{l \neq k} D(C_l,\, C_k) = \min_{l \neq k} \left\{ \min_{r \in C_l,\, s \in C_k} \{d_{rs}\} \right\}$$

(1) Distance-matrix of partition $\mathcal{C}^{(0)} = \{\{1\}, \{2\}, \{3\}, \{4\}\}$:

e.g.. $d_{12} = \left|\left| \begin{pmatrix} 8 \\ 24 \end{pmatrix} - \begin{pmatrix} 5 \\ 22 \end{pmatrix} \right|\right|^2 = \left|\left| \begin{pmatrix} 3 \\ 2 \end{pmatrix} \right|\right|^2 = 3^2 + 2^2 = 13$

	1	2	3	4
1	0	13	5	25
2		0	34	(2)
3			0	52
4				0

$\Rightarrow \quad h_1 = \min_{l \neq k} \left\{ \min_{r \in C_l,\, s \in C_k} \{d_{rs}\} \right\} = 2 \widehat{=} D(\{2\}, \{4\})$

$\Rightarrow$ Step 1: Merge $\{2\}$ and $\{4\}$

$\Rightarrow \mathcal{C}^{(1)} = \{\{1\}, \{2,4\}, \{3\}\}$

(2) Distance-matrix of partition $\mathcal{C}^{(1)}$:

$$\begin{array}{c|ccc} & 1 & 2,4 & 3 \\ \hline 1 & 0 & 13 & \textcircled{5} \\ 2,4 & & 0 & 34 \\ 3 & & & 0 \end{array} \quad \Rightarrow \quad h_2 = \min_{l \neq k} \left\{ \min_{r \in C_l,\, s \in C_k} \{d_{rs}\} \right\} = 5 \widehat{=} D(\{1\}, \{3\})$$

$\Rightarrow$ Step 2: Merge $\{1\}$ and $\{3\}$

$\Rightarrow \mathcal{C}^{(2)} = \{\{1,3\}, \{2,4\}\}$

(3) Distance between $\{1,3\}$ and $\{2,4\}$:

$h_3 = \min_{r \in \{1,3\},\, s \in \{2,4\}} \{d_{rs}\} = 13 \widehat{=} D(\{1,3\}, \{2,4\})$

$\Rightarrow$ Step 3: Merge $\{1,3\}$ and $\{2,4\}$

$\Rightarrow \mathcal{C}^{(3)} = \{\{1,2,3,4\}\}$

b) Zentroid-procedure with squared euclidean distance: In step ν, we merge those clusters $C_i,\, C_j \in \mathcal{C}^{(\nu-1)}$ for which the following applies:

$$D(C_i,\, C_j) = (h_\nu =) \min_{l \neq k} D(C_l,\, C_k) = \min_{l \neq k} ||\bar{x}_l - \bar{x}_k||^2 \,, \quad \text{where} \quad \bar{x}_r = \frac{1}{n_r} \sum_{s \in C_r} x_s$$

(1) Distance-matrix of partition $\mathcal{C}^{(0)} = \{\{1\}, \{2\}, \{3\}, \{4\}\}$:

$$\begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & 13 & 5 & 25 \\ 2 & & 0 & 34 & \textcircled{2} \\ 3 & & & 0 & 52 \\ 4 & & & & 0 \end{array} \quad \Rightarrow \quad h_1 = \min_{l \neq k} ||\bar{x}_l - \bar{x}_k||^2 = 2 \widehat{=} D(\{2\}, \{4\})$$

$\Rightarrow$ Step 1: Merge $\{2\}$ and $\{4\}$

$\Rightarrow \mathcal{C}^{(1)} = \{\{1\}, \{2,4\}, \{3\}\}$

Cluster centroids:

$$\bar{x}_{\{2,4\}} = \frac{1}{2} \left(\begin{pmatrix} 5 \\ 22 \end{pmatrix} + \begin{pmatrix} 4 \\ 21 \end{pmatrix} \right) = \begin{pmatrix} 4,5 \\ 21,5 \end{pmatrix}$$

$$\Rightarrow \bar{X}^{(1)} = \begin{pmatrix} 8 & 4,5 & 10 \\ 24 & 21,5 & 25 \end{pmatrix}$$

$$\{1\} \quad \{2,4\} \quad \{3\}$$

(2) Distance-matrix of partition $\mathcal{C}^{(1)}$:

e.g. $D(\{1\}, \{2,4\}) = \left|\left| \begin{pmatrix} 8 \\ 24 \end{pmatrix} - \begin{pmatrix} 4,5 \\ 21,5 \end{pmatrix} \right|\right|^2 = \left|\left| \begin{pmatrix} 3,5 \\ 2,5 \end{pmatrix} \right|\right|^2 = 3,5^2 + 2,5^2 = 18,5$

$$\begin{array}{c|ccc} & 1 & 2,4 & 3 \\ \hline 1 & 0 & 18,5 & \textcircled{5} \\ 2,4 & & 0 & 42,5 \\ 3 & & & 0 \end{array} \quad \Rightarrow \quad h_2 = \min_{l \neq k} ||\bar{x}_l - \bar{x}_k||^2 = 5 \widehat{=} D(\{1\}, \{3\})$$

$\Rightarrow$ Step 2: Merge $\{1\}$ and $\{3\}$

$\Rightarrow \mathcal{C}^{(2)} = \{\{1,3\},\{2,4\}\}$

Cluster centroids:

$$\Rightarrow \bar{X}^{(2)} = \begin{pmatrix} 9 & 4,5 \\ 24,5 & 21,5 \end{pmatrix}$$

$$\{1,3\} \quad \{2,4\}$$

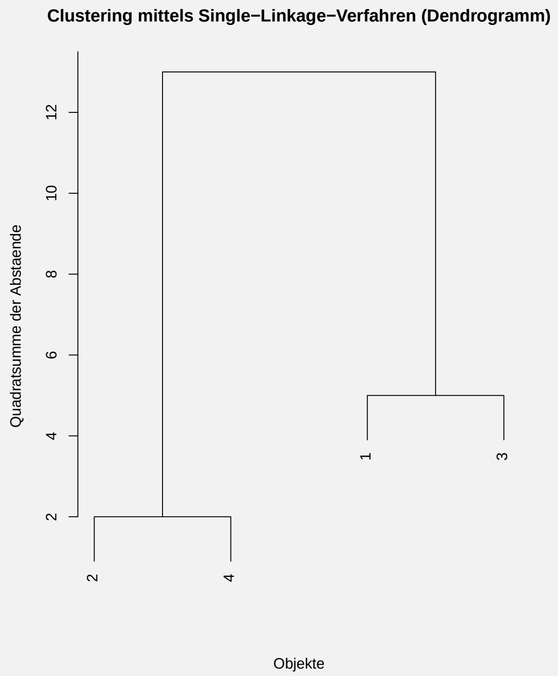
(3) Distance between $\{1,3\}$ and $\{2,4\}$:

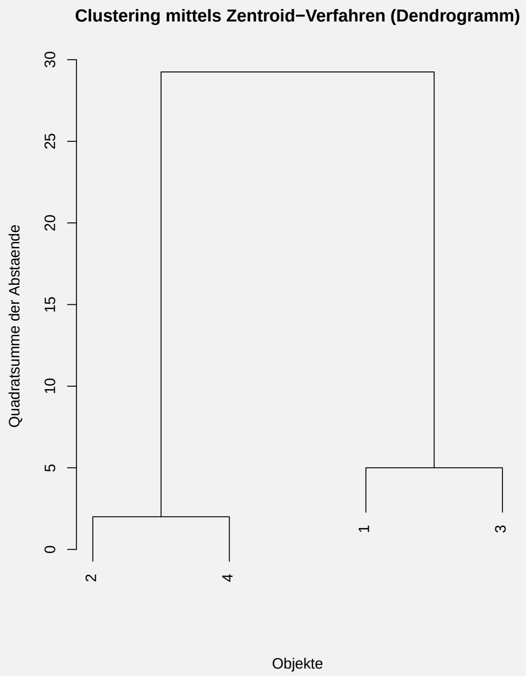
$h_3 = ||\bar{x}_{\{1,3\}} - \bar{x}_{\{2,4\}}||^2 = 4,5^2 + 3^2 = 29,25 \widehat{=} D(\{1,3\},\{2,4\})$

$\Rightarrow$ Step 3: Merge $\{1,3\}$ and $\{2,4\}$

$\Rightarrow \mathcal{C}^{(3)} = \{\{1,2,3,4\}\}$

c) The dendograms resulting from Single-Linkage and Zentroid procedures, respectively, are given by the following:





Question 3:

a) For a set of points $(x_i)_{i=1}^n$ in $\mathbb{R}^m$, show that the arithmetic mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i$ is the solution to the optimization problem

$$\hat{\mu} = \underset{\mu \in \mathbb{R}^m}{\operatorname{argmin}} \sum_{i=1}^n \|x_i - \mu\|^2$$

I.e. for a set of points, their mean can be characterized as the point which is, on average, closest to all the other points with respect to the squared euclidean distance.

b) Consider the following six points in $\mathbb{R}^2$:

$$x_1 = \begin{pmatrix}0\\0\end{pmatrix};\ x_2 = \begin{pmatrix}0\\1\end{pmatrix};\ x_3 = \begin{pmatrix}-1\\2\end{pmatrix};\ x_4 = \begin{pmatrix}2\\0\end{pmatrix};\ x_5 = \begin{pmatrix}3\\0\end{pmatrix};\ x_6 = \begin{pmatrix}4\\-1\end{pmatrix}.$$

Use Lloyd's algorithm and "random" initialization $\{x_1; x_6\}$ to perform **both** *k-means* and *k-medoids* (also with squared euclidean distance) clustering for $K = 2$.

Solution:

a) This immediately follows from the lecture slide's lemma in the subsection "Non-probabilistic methods" of chapter 7.1, whereby

$$\sum_{i=1}^{n} \|x_i - z\|^2 \geq \sum_{i=1}^{n} \|x_i - \hat{\mu}\|^2 \quad \forall z \in \mathbb{R}^m$$

in this setting.

b) For both k-means and k-medoids, we start by computing the squared euclidean distance between all points:

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0	1	5	4	9	17
x_2	1	0	2	5	10	20
x_3	5	2	0	12	20	34
x_4	4	5	12	0	1	5
x_5	9	10	20	1	0	2
x_6	17	20	34	5	2	0

This also gives us the following distances between the initialization points and all others:

	x_1	x_2	x_3	x_4	x_5	x_6
$\mu_1 = x_1$	0	1	5	4	9	17
$\mu_2 = x_6$	17	20	34	5	2	0

- **Then, for k-means:**

  <u>Iteration 1</u>: Looking at rows of the distance matrix corresponding to the centers

	x_1	x_2	x_3	x_4	x_5	x_6
$\mu_1 = x_1$	0	1	5	4	9	17
$\mu_2 = x_6$	17	20	34	5	2	0

  Then the partitions are $P_1 = \{x_1, x_2, x_3, x_4\}$ and $P_2 = \{x_5, x_6\}$. To find the new cluster centers, we have to compute the means:

$$\mu_1' = \frac{1}{|P_1|} \sum_{x \in P_1} = \frac{1}{4}\left(\begin{pmatrix}0\\0\end{pmatrix} + \begin{pmatrix}0\\1\end{pmatrix} + \begin{pmatrix}-1\\2\end{pmatrix} + \begin{pmatrix}2\\0\end{pmatrix}\right) = \frac{1}{4}\begin{pmatrix}1\\3\end{pmatrix}$$

$$\mu_2' = \frac{1}{|P_2|} \sum_{x \in P_2} = \frac{1}{2}\left(\begin{pmatrix}3\\0\end{pmatrix} + \begin{pmatrix}4\\-1\end{pmatrix}\right) = \frac{1}{2}\begin{pmatrix}7\\-1\end{pmatrix}$$

Iteration 2: We compute the squared euclidean distances to the new cluster centers:

	x_1	x_2	x_3	x_4	x_5	x_6
μ_1	0.625	0.125	3.125	3.625	8.125	17.125
μ_2	12.500	14.500	26.500	2.500	0.500	0.500

Then the partitions are $P_1 = \{x_1, x_2, x_3\}$ and $P_2 = \{x_4, x_5, x_6\}$. To find the new cluster centers, we have to compute the means:

$$\mu_1' = \frac{1}{|P_1|} \sum_{x \in P_1} = \frac{1}{3}\left(\begin{pmatrix}0\\0\end{pmatrix} + \begin{pmatrix}0\\1\end{pmatrix} + \begin{pmatrix}-1\\2\end{pmatrix}\right) = \frac{1}{3}\begin{pmatrix}-1\\3\end{pmatrix}$$

$$\mu_2' = \frac{1}{|P_2|} \sum_{x \in P_2} = \frac{1}{3}\left(\begin{pmatrix}2\\0\end{pmatrix} + \begin{pmatrix}3\\0\end{pmatrix} + \begin{pmatrix}4\\-1\end{pmatrix}\right) = \frac{1}{3}\begin{pmatrix}9\\-1\end{pmatrix}$$

Iteration 3: We compute the squared euclidean distances to the new cluster centers:

	x_1	x_2	x_3	x_4	x_5	x_6
μ_1	1.111	0.111	1.444	6.444	12.111	22.777
μ_2	9.111	10.777	21.444	1.111	0.111	1.444

Then the partitions are $P_1 = \{x_1, x_2, x_3\}$ and $P_2 = \{x_4, x_5, x_6\}$. As these are the same as in the previous iteration, the algorithm terminates.

- **and for k-medoids:**

Iteration 1: The partitions are $P_1 = \{x_1, x_2, x_3, x_4\}$ and $P_2 = \{x_5, x_6\}$. To find the new cluster centers, we sum the rows of the following sub-matrices of squared euclidean distances:

	x_1	x_2	x_3	x_4	Σ
x_1	0	1	5	4	10
x_2	1	0	2	5	8
x_3	5	2	0	12	19
x_4	4	5	12	0	21

and

	x_5	x_6	Σ
x_5	0	2	2
x_6	2	0	2

From this, we get that $\mu_1^{(1)} = x_2$ and $\mu_2^{(1)} = x_5$ or $\mu_2^{(1)} = x_6$ – we choose the former, i.e. $\mu_2^{(1)} = x_5$.

Iteration 2: Looking at rows of the distance matrix corresponding to the centers

	x_1	x_2	x_3	x_4	x_5	x_6
$\mu_1 = x_2$	1	0	2	5	10	20
$\mu_2 = x_5$	9	10	20	1	0	2

Then the partitions are $P_1 = \{x_1, x_2, x_3\}$ and $P_2 = \{x_4, x_5, x_6\}$. To find the new cluster centers, we sum the rows of the corresponding subtables:

	x_1	x_2	x_3	Σ
x_1	0	1	5	6
x_2	1	0	2	3
x_3	5	2	0	7

$\rightsquigarrow \mu_1' = x_2$

	x_4	x_5	x_6	Σ
x_4	0	1	5	6
x_5	1	0	2	3
x_6	5	2	0	7

$\rightsquigarrow \mu_2' = x_5$

The cluster centers are the same as before, so the algorithm terminates.

Question 4:

a) Outline the model assumptions used in the Gaussian Mixed Models (GMMs). How can a GMM be fit?

b) Consider a one-dimensional Gaussian Mixture Model with 2 clusters and parameters $(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \pi_1, \pi_2)$. Here (π_1, π_2) are the mixing weights, and $(\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2)$ are the centers and variances of the clusters. We are given a dataset $\mathcal{D} = \{x_1, x_2, x_3\} \subset \mathbb{R}$, and apply the EM-algorithm to find the parameters of the Gaussian mixture model. What is the complete log-likelihood that is being optimized for this problem?

c) Assume that the dataset $\mathcal{D}$ consists of the following three points, $x_1 = 1, x_2 = 10, x_3 = 20$. At some step in the EM-algorithm, we compute the expectation step which results in the following matrix: $\mathrm{T} = \begin{pmatrix} 1 & 0 \\ 0.4 & 0.6 \\ 0 & 1 \end{pmatrix}$, where τ_{ij} denotes the probability of x_i belonging to cluster j.

Given the above T for the expectation step, write the result of the following maximization step, specifically the

- mixing weights π_1, π_2
- centers μ_1, μ_2
- variance values σ_1^2, σ_2^2

Solution:

a)
- Observations: $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ with $\boldsymbol{x}_i \in \mathbb{R}^m$.
- Unknown group membership $r_1, \dots, r_n$
- For a given group membership, $\boldsymbol{x}_i$ is normally distributed:
  - $\boldsymbol{x}_i | r \sim \mathcal{N}(\boldsymbol{\mu}_r, \Sigma_r), \; r \in \{1, \dots, k\}$
  - $f_r(\boldsymbol{x}_i) = f(\boldsymbol{x}_i | r) = f(\boldsymbol{x}_i | \boldsymbol{\mu}_r, \Sigma_r)$
- Prior probability of group membership:

$$p(r), \; r \in \{1, \dots, k\}$$

- Assumption of mixture distribution:

$$f(\boldsymbol{x}) = \sum_{r=1}^{k} p(r) f(\boldsymbol{x} | r)$$

- Posteriori-probability of group membership:

$$\hat{p}(r|\boldsymbol{x}_i) = \frac{\hat{p}(r)\hat{f}(\boldsymbol{x}_i|r)}{\hat{f}(\boldsymbol{x}_i)} =: \hat{p}_{ir} \tag{1}$$

- Group assignment via marginal, estimated Posteriori-probability:

$$\mathcal{C}_r = \{\boldsymbol{x}_i | \hat{p}_{ir} \geq \hat{p}_{is},\ r \neq s\},\ r \in \{1, \ldots, k\}$$

- Parameter estimation via *EM algorithm*, iterated until convergence:
  (1) E-step:
    - Given $\hat{p}(r),\ \hat{f}(\boldsymbol{x}|r),\ \hat{f}(\boldsymbol{x})$
    - Calculate $\hat{p}_{ir}$ according to (1)

  (2) M-step:
    - Given $\hat{p}_{ir}$, update $\hat{p}(r) = \frac{1}{n}\sum_{i=1}^{n} \hat{p}_{ir}$
    - $(\hat{\boldsymbol{\mu}}_r,\ \hat{\Sigma}_r) = \arg \max_{\boldsymbol{\mu}_r,\, \Sigma_r} \sum_{i=1}^{n} \hat{p}_{ir} \cdot \log\left(f_r(\boldsymbol{x}_i|\boldsymbol{\mu}_r,\ \Sigma_r)\right),\ r \in \{1, \ldots, g\}$ (weighted MLE)

b) The complete log-likelihood is given by

$$\begin{aligned}\log f\left(\mathcal{D} \mid \left(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \pi_1, \pi_2\right)\right) = &\log\left\{\pi_1\phi\left(x_1; \mu_1, \sigma_1\right) + \pi_2\phi\left(x_1; \mu_2, \sigma_2\right)\right\} + \\ &\log\left\{\pi_1\phi\left(x_2; \mu_1, \sigma_1\right) + \pi_2\phi\left(x_2; \mu_2, \sigma_2\right)\right\} + \\ &\log\left\{\pi_1\phi\left(x_3; \mu_1, \sigma_1\right) + \pi_2\phi\left(x_3; \mu_2, \sigma_2\right)\right\}\end{aligned}$$

where ϕ denotes the density of the one-dimensional normal distribution.

c) For the mixing weights, it holds that

$$\pi_j = \frac{1}{n}\sum_{i=1}^{n} \tau_{ij}\,,$$

so we get

$$\begin{aligned}\pi_1 &= \frac{1}{3}(1 + 0.4 + 0) = 1.4/3 \approx 0.47 \\ \pi_2 &= \frac{1}{3}(0 + 0.6 + 1) = 1.6/3 \approx 0.53\,.\end{aligned}$$

For the centers, it holds that

$$\mu_j = \frac{\sum_{i=1}^{n} \tau_{ij} x_i}{\sum_{i=1}^{n} \tau_{ij}}\,,$$

so we get

$$\begin{aligned}\mu_1 &= \frac{1}{1.4}(1 \cdot 1 + 0.4 \cdot 10 + 0 \cdot 20) = 5/1.4 \approx 3.57 \\ \mu_2 &= \frac{1}{1.6}(0 \cdot 1 + 0.6 \cdot 10 + 1 \cdot 20) = 26/1.6 \approx 16.25\,.\end{aligned}$$

For the variance values, it holds that

$$\sigma_j^2 = \frac{\sum_{i=1}^{n} \tau_{ij}\left(x_i - \mu_j\right)\left(x_i - \mu_j\right)^T}{\sum_{i=1}^{n} \tau_{ij}} \overset{\text{b/c one-dimensional}}{=} \frac{\sum_{i=1}^{n} \tau_{ij}\left(x_i - \mu_j\right)^2}{\sum_{i=1}^{n} \tau_{ij}}\,,$$

so we get

$$\begin{aligned}\sigma_1^2 &= \frac{1}{1.4}\left(1 \cdot \left(1 - \frac{5}{1.4}\right)^2 + 0.4 \cdot \left(10 - \frac{5}{1.4}\right)^2 + 0 \cdot \left(20 - \frac{5}{1.4}\right)^2\right) \approx 16.53 \\ \sigma_2^2 &= \frac{1}{1.6}\left(0 \cdot \left(1 - \frac{26}{1.6}\right)^2 + 0.6 \cdot \left(10 - \frac{26}{1.6}\right)^2 + 1 \cdot \left(20 - \frac{26}{1.6}\right)^2\right) \approx 23.44\,.\end{aligned}$$