Distances for Quantitative Features

A visual walk-through of textbook Section 4.2.1 (Handl and Kuhlenkasper)

Prof. Dr. Jörg Osterrieder

School of Engineering and Computer Sciences

A Puzzle Before We Start

Six students each told us two numbers: the age of their mother and the age of their father. Student 1 answered (58, 60); student 2 answered (61, 62).

How would you turn "how different are student 1 and student 2" into a single, fair number?

That single number is a **distance**. Section 4.2.1 builds it step by step for features measured on a numeric scale, using exactly these 6 students.

What should a good "distance between two students" number do, before we pick a formula?

1. The Goal: One Number per Pair of Objects

For objects described only by **quantitative features** (values on a numeric scale, like an age), we summarise how far apart two objects are by a single distance d_{ij}. We collect all pairwise distances of n objects in a **distance matrix** D:

- It is symmetric: $d_{ij} = d_{ji}$ (the distance does not depend on order).
- Its diagonal is all zeros: $d_{ii} = 0$ (an object has no distance to itself).

Everything in this section answers one question: which formula should fill the entries of D?

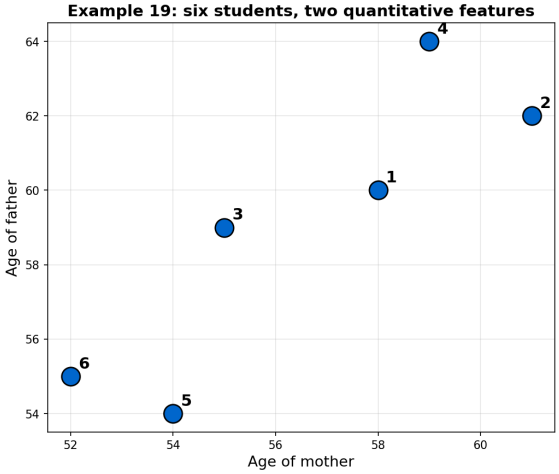
Why must a distance matrix be symmetric with a zero diagonal?

2. The Data: Example 19 (Six Students)

The textbook's running example, Table 4.1:

Student	Age mother	Age father
1	58	60
2	61	62
3	55	59
4	59	64
5	54	54
6	52	55

Two quantitative features, so each student is one point in a plane.



With two numeric features, what does one student look like geometrically?

3. Centering: Shift the Cloud, Keep the Distances

The textbook first **centers** the data: subtract each feature's mean. The means here are mother $= 56.5$ and father $= 59$.
Predict: does subtracting the mean change the distance between any two students?

3. Centering: Shift the Cloud, Keep the Distances

The textbook first **centers** the data: subtract each feature's mean. The means here are mother $= 56.5$ and father $= 59$. *Predict: does subtracting the mean change the distance between any two students?* No. Centering only slides the whole cloud; every pairwise distance is unchanged. It just makes the picture and the algebra simpler. Centered values (Table 4.2):

Student	1	2	3	4	5	6
mother	$+1.5$	$+4.5$	-1.5	$+2.5$	-2.5	-4.5
father	$+1.0$	$+3.0$	0.0	$+5.0$	-5.0	-4.0

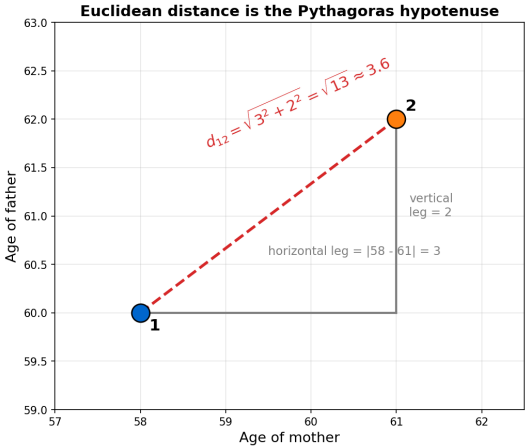
If centering leaves every distance unchanged, why bother doing it?

4. Euclidean Distance Is Just Pythagoras

The natural distance is the straight-line gap between the two points. In the plane, the Pythagoras theorem gives, for students i and j:

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2}$$

The horizontal leg is the difference in mothers' ages; the vertical leg is the difference in fathers' ages; d_{ij} is the hypotenuse.



Which two "legs" does the Euclidean distance between two students combine?

5. Worked Example: Students 1 and 2

Student 1 is $(58, 60)$ and student 2 is $(61, 62)$. *Predict the distance before the reveal.*

5. Worked Example: Students 1 and 2

Student 1 is $(58, 60)$ and student 2 is $(61, 62)$. *Predict the distance before the reveal.*

$$d_{12} = \sqrt{(58-61)^2 + (60-62)^2} = \sqrt{9+4} = \sqrt{13} \approx 3.6$$

Three years apart on mother's age, two years apart on father's age, combine (by Pythagoras) into one number: 3.6. Every value on the next slide is produced by this same calculation.

Why is the combined distance 3.6 rather than $3 + 2 = 5$?

6. The Full Euclidean Distance Matrix (Eq. 4.1)

Applying the formula to all $\binom{6}{2}$ pairs gives the textbook's distance matrix D (rounded to one decimal):

	1	2	3	4	5	6
1	0.0	3.6	3.2	4.1	7.2	7.8
2	3.6	0.0	6.7	2.8	10.6	11.4
3	3.2	6.7	0.0	6.4	5.1	5.0
4	4.1	2.8	6.4	0.0	11.2	11.4
5	7.2	10.6	5.1	11.2	0.0	2.2
6	7.8	11.4	5.0	11.4	2.2	0.0

Symmetric, zero diagonal, exactly as a distance matrix must be. Students 5 and 6 are closest (2.2); students 2 and 5 are farthest (10.6).

Reading the matrix: which pair of students is the most similar, and how do you see it?

7. The Same Idea in p Dimensions

Nothing about Pythagoras needed exactly two features. For objects with feature vectors x_i and x_j of length p, the Euclidean distance is:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} = \sqrt{(x_i - x_j)'(x_i - x_j)}$$

- Sum the squared difference on every feature, then take the square root.
- With $p = 2$ this is exactly the triangle on slide 4.
- The matrix form $(x_i - x_j)'(x_i - x_j)$ is what software computes.

How does the two-feature formula extend to objects with p features?

8. Problem: Features on Very Different Scales

Euclidean distance adds raw squared differences, so a feature with a much larger spread dominates the result. The fix in Section 4.2.1 is the **scaled Euclidean distance**: divide each feature's difference by that feature's sample standard deviation s_k before combining:

$$d_{ij}^{s} = \sqrt{\sum_{k=1}^{p} \frac{(x_{ik} - x_{jk})^2}{s_k^2}} = \sqrt{\sum_{k=1}^{p} \left(\frac{x_{ik}}{s_k} - \frac{x_{jk}}{s_k} \right)^2}$$

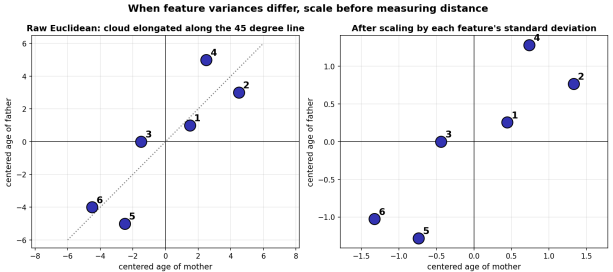
Now every feature contributes on a comparable footing.

Why can a single large-spread feature distort plain Euclidean distance?

9. Scaled Distance on Example 19

The sample variances (dividing by $n-1$) are $s_1^2 = 11.5$ (mother) and $s_2^2 = 15.2$ (father). Scaling by s_k gives the textbook's scaled distance matrix (Eq. 4.4):

	1	2	3	4	5	6
1	0.0	1.0	0.9	1.1	1.9	2.2
2	1.0	0.0	1.9	0.8	2.9	3.2
3	0.9	1.9	0.0	1.7	1.3	1.4
4	1.1	0.8	1.7	0.0	3.0	3.1
5	1.9	2.9	1.3	3.0	0.0	0.6
6	2.2	3.2	1.4	3.1	0.6	0.0



After scaling, which pair is closest, and did the ranking change versus the raw matrix?

10. Scaled Distance in Matrix Form

Collect the feature variances on the diagonal of a matrix V:

$$V = \operatorname{diag}(s_1^2, s_2^2, \ldots, s_p^2) \qquad \Longrightarrow \qquad d_{ij}^{s} = \sqrt{(x_i - x_j)' \, V^{-1} \, (x_i - x_j)}$$

- V^{-1} simply puts $1/s_k^2$ on the diagonal.
- This is the plain Euclidean distance computed on the standardised data $V^{-1/2}x$.
- Special case: if all variances are equal, V^{-1} is a constant and we are back to ordinary Euclidean distance.

What does V^{-1} contain, and what distance do we recover when all variances are equal?

11. The Problem Scaling Cannot Fix

Scaling made the two features equally spread, but in Example 19 the cloud is still **tilted**: it stretches along the 45 degree line because mother's age and father's age move together (they are correlated). *Predict: does dividing each feature by its standard deviation straighten a tilted cloud?*

11. The Problem Scaling Cannot Fix

Scaling made the two features equally spread, but in Example 19 the cloud is still **tilted**: it stretches along the 45 degree line because mother's age and father's age move together (they are correlated). *Predict: does dividing each feature by its standard deviation straighten a tilted cloud?* No. Scaling only resizes the axes; it cannot turn a slanted cloud upright. So two students can sit close on each axis yet be far apart along the cloud's own long direction. The scaled distances (Eq. 4.4, from variances $s_1^2 = 11.5$ and $s_2^2 = 15.2$) still do not capture this.

Why can two students look close on each feature yet still be far apart overall?

12. The Fix in One Picture: Turn the Ruler

Think of slanted handwriting. You do not measure it letter by letter on the page grid; you **tilt your head** until the lines run straight, then read. Same idea here. The cloud is like a rugby ball lying diagonally:

- A ruler along the page axes overstates how close two points are.
- Turn the ruler to follow the ball's own **long** and **short** axes, then rescale each so the ball becomes a round circle.
- Distances measured in that turned, rescaled view are finally fair.

That is the whole trick: **rotate first, then scale**.

In the rugby-ball picture, what goes wrong if you measure distance without turning the ruler first?

13. What “Rotate” Really Means

Rotating the cloud just builds **new axes** that are weighted blends of the old features:

$$\text{new axis} = a \cdot (\text{age mother}) + b \cdot (\text{age father})$$

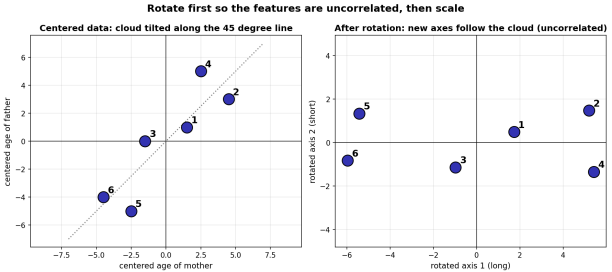
(a *linear combination*, the same kind of weighted sum you already use for a course grade). The weights are chosen so the two new features are **uncorrelated**: one axis follows the cloud's long direction, the other its short direction. *Predict: which new axis has the bigger spread?*

13. What “Rotate” Really Means

Rotating the cloud just builds **new axes** that are weighted blends of the old features:

$$\text{new axis} = a \cdot (\text{age mother}) + b \cdot (\text{age father})$$

(a *linear combination*, the same kind of weighted sum you already use for a course grade). The weights are chosen so the two new features are **uncorrelated**: one axis follows the cloud's long direction, the other its short direction. *Predict: which new axis has the bigger spread?* The long one (by far).



Why is a rotated axis just a weighted sum of the original two features?

14. Scale on the Rotated Axes (Example 19)

On the cloud's own axes the spreads are very different: the long axis has variance $s_1^2 = 25.1$, the short axis only $s_2^2 = 1.6$. Divide each rotated feature by its standard deviation, then take ordinary Euclidean distance. This is the textbook's Section 4.2.1 result: the rotated-then-scaled distance matrix (Eq. 4.5).

	1	2	3	4	5	6
1	0.00	1.04	1.40	1.63	1.58	1.85
2	1.04	0.00	2.40	2.23	2.12	2.87
3	1.40	2.40	0.00	1.29	2.16	1.03
4	1.63	2.23	1.29	0.00	3.04	2.31
5	1.58	2.12	2.16	3.04	0.00	1.72
6	1.85	2.87	1.03	2.31	1.72	0.00

What two operations, in order, turned the raw ages into these fair distances?

15. This Construction Has a Name

You just built a distance by **rotating to uncorrelated axes and then scaling**. That exact construction is the **Mahalanobis distance**, the subject of the next sections. The progression for quantitative features:

- Euclidean: straight-line distance (Pythagoras).
- Scaled Euclidean: divide by each feature's spread.
- Rotate then scale: handle correlated features.

In one sentence, what is the difference between scaled Euclidean and the Mahalanobis distance?

Rotation, Visually: What the New Axes Are

A tilted cloud has two natural directions of its own:

- a **long axis**, where the points spread the most,
- a **short axis**, at right angles, where they spread least.

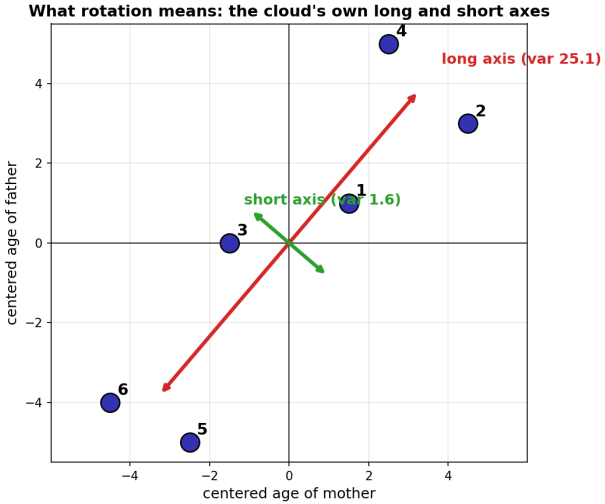
Rotating just means: stop using the page axes and use **these two axes** instead. The new features are the coordinates along them. *Predict: along which drawn axis is the cloud most spread out?*

Rotation, Visually: What the New Axes Are

A tilted cloud has two natural directions of its own:

- a **long axis**, where the points spread the most,
- a **short axis**, at right angles, where they spread least.

Rotating just means: stop using the page axes and use **these two axes** instead. The new features are the coordinates along them. *Predict: along which drawn axis is the cloud most spread out?* The long one: its spread (variance 25.1) dwarfs the short one (1.6).



What are the cloud's own two axes, and how do you spot the long one?

Rotation, Visually: Why It Is Needed

On the page axes the big spread runs along the 45 degree diagonal, so plain and scaled distance over-weight that one direction.

- Scaling alone leaves the cloud tilted: the problem is not fixed.
- Rotating onto the cloud's own axes measures distance along the directions that actually matter.

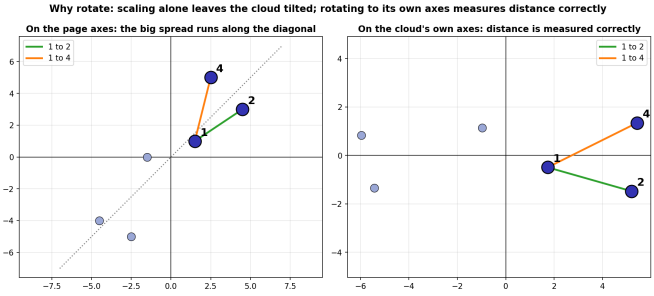
Predict: does rotating change which student is closest to student 1?

Rotation, Visually: Why It Is Needed

On the page axes the big spread runs along the 45 degree diagonal, so plain and scaled distance over-weight that one direction.

- Scaling alone leaves the cloud tilted: the problem is not fixed.
- Rotating onto the cloud's own axes measures distance along the directions that actually matter.

Predict: does rotating change which student is closest to student 1? No. Student 2 stays nearer student 1 than student 4 throughout; rotation only makes the measurement honest, it does not change who is closest.



Why does measuring on the page axes over-weight the diagonal direction?

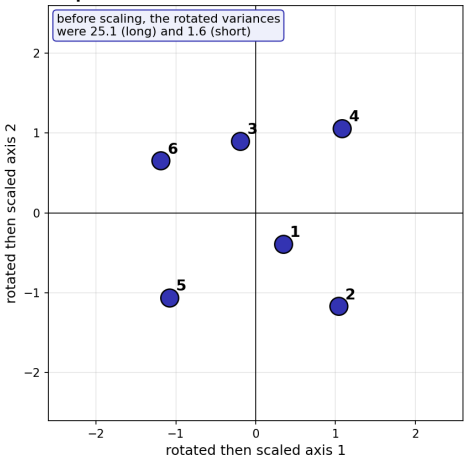
Rotation, Visually: What Changes

Predict the cloud's shape after we rotate and then divide each new axis by its spread.

Rotation, Visually: What Changes

Predict the cloud's shape after we rotate and then divide each new axis by its spread. It becomes **round**. Before scaling the rotated variances were 25.1 (long) and 1.6 (short); dividing each axis by its standard deviation makes both spreads equal to 1, so the cloud is a circular, uncorrelated blob. Plain Euclidean distance on this round cloud is finally fair: that is exactly the Mahalanobis distance named on the previous slide.

Consequence: after rotate then scale the cloud is round



After rotate then scale, why is plain Euclidean distance now the fair one to use?

The One Big Idea

For quantitative features, distance starts as **Pythagoras** (Euclidean), becomes **scaled Euclidean** when spreads differ, and needs **rotate then scale** when features are correlated.

- Distance matrix D: symmetric, zero diagonal, one number per pair.
- Euclidean: $d_{ij} = \sqrt{\sum_k (x_{ik} - x_{jk})^2}$; here $d_{12} = \sqrt{13} \approx 3.6$.
- Scaled: divide by s_k ($s_1^2 = 11.5$, $s_2^2 = 15.2$) so no feature dominates.
- Correlated features: rotate to uncorrelated axes, then scale.

Could you now choose between Euclidean and scaled Euclidean for a new dataset, and justify it?