The mathematics of Principal Component Analysis

Study notes for Part 2 of the lecture

CS429 Machine Learning Algorithms and Practice

Companion to the slides "Principal Component Analysis (PCA)"

GP

Contents

How to use these notes 3

1 The running example and the notation 3
1.1 Six patients, two measurements . . . 3
1.2 The mean and the centred data . . . 3
1.3 From variance and covariance to the covariance matrix . . . 4

2 The ultimate aim, and the road to it 5

3 Step 1: the score of a patient along a direction 6
3.1 A direction is a unit vector . . . 6
3.2 The score is a projection . . . 6
3.3 The mean of the scores is zero . . . 6

4 Step 2: the variance along a direction as one formula 7

5 Step 3: why the length of $\mathbf{a}_1$ must be fixed 8

6 Step 4: from "the top of the curve" to a derivative 9
6.1 Only the angle is free . . . 9
6.2 The direction of turning . . . 10
6.3 The slope of the variance curve . . . 10

7 Step 5: what "flat" tells us, from the derivative to an eigenvector 11
7.1 The tool: splitting a vector along two perpendicular axes . . . 12
7.2 Applying the tool to $\mathbf{Sa}$ . . . 12
7.3 Using the condition . . . 13
7.4 The general case: any number of measurements . . . 13

8 Step 6: what the eigenvalue tells us, and the choice of PC1 13
8.1 How to read $\mathbf{Sa} = \lambda\mathbf{a}$ . . . 14

9 Step 7: computing PC1 by hand 15
9.1 The eigenvalues . . . 15
9.2 The eigenvector . . . 17

10 Step 8: the next best direction 18

11 Step 9: all the components together 18
11.1 The matrix $\mathbf{A}$: all the scores at once . . . 19
11.2 $\mathbf{A}$ is a rotation . . . 19
11.3 The covariance of the new variables . . . 19
11.4 The eigendecomposition $\mathbf{S} = \mathbf{A}\Lambda\mathbf{A}^\top$ . . . 20
11.5 The total spread, and the budget we will split . . . 20

12 Step 10: how many components to keep 20

13 Step 11: most spread is least loss 21

14 Step 12: the algorithm, and the SVD 22

A The same result with a Lagrange multiplier 23

B All the equations on one page 24

Further reading 24

How to use these notes

These notes follow, step by step, the mathematical part of the lecture on PCA (Part 2 of the slides). They do *not* try to follow a strict formalism. Their purpose is to elaborate on each step, linking the intuition to the mathematics, and they stay very close to the examples discussed during the lectures: the same six flu patients, the same pictures and the same numbers. The slides show the equations and a few words beside each one; here every step is spelled out and, above all, comes with an answer to the question *why are we doing this?*

The language is often informal. Words such as "shadow", "turning" or "stretching" are used where a textbook would use more precise terms, and some statements are made for two measurements and then extended in words. This deliberately sacrifices some accuracy in favour of a first understanding. For a more formal treatment, see the readings suggested at the end of these notes (Further reading).

Each section has the same structure.

- A blue box, **Where we are and where we are going**, recalls the ultimate aim and states the intermediate aim of the section: the specific problem that the section solves.
- The mathematics, with every computation written out.
- Grey boxes, **Why we do this**, which explain the purpose of a step that might otherwise look arbitrary.
- Green boxes, **Check with our patients**, which repeat each result with the numbers of the running example, so that you can verify everything with a calculator.

1 The running example and the notation

1.1 Six patients, two measurements

We use the same small data set as in the slides: six flu patients, for whom we record the number of coughs per hour and the number of sneezes per hour.

Patient i	1	2	3	4	5	6
Coughs per hour	13	12	10	5	2	0.5
Sneezes per hour	4	6.5	7.5	3	7	1

Each patient is a vector with $p = 2$ entries, $\mathbf{x}_i = (\text{coughs}_i, \text{sneezes}_i)^\top$, and there are $n = 6$ patients. Everything in these notes works for any number of measurements p; we use $p = 2$ because we can draw it.

Notation. Bold lower-case letters ($\mathbf{a}$, $\mathbf{x}$, $\mathbf{d}$) are column vectors; bold capitals ($\mathbf{S}$, $\mathbf{A}$) are matrices; $^\top$ means transpose, so $\mathbf{a}^\top\mathbf{d}$ is the dot product of $\mathbf{a}$ and $\mathbf{d}$, a single number. The length (or *norm*) of a vector is $\|\mathbf{a}\| = \sqrt{\mathbf{a}^\top\mathbf{a}}$.

1.2 The mean and the centred data

These objects come from the earlier lectures on variance, covariance and correlation.

The **mean** of each measurement:

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i, \qquad \bar{x}_{\text{coughs}} = \frac{42.5}{6} = 7.083, \qquad \bar{x}_{\text{sneezes}} = \frac{29}{6} = 4.833.$$

The **centred data**: each patient minus the mean, $\mathbf{d}_i = \mathbf{x}_i - \bar{\mathbf{x}}$.

Patient i	1	2	3	4	5	6
coughs $-\,7.083$	5.917	4.917	2.917	-2.083	-5.083	-6.583
sneezes $-\,4.833$	-0.833	1.667	2.667	-1.833	2.167	-3.833

Centring moves the origin to the centre of the cloud of points. It does not change the shape of the cloud, only where we measure it from.

1.3 From variance and covariance to the covariance matrix

We now build the covariance matrix in three small steps, starting from the definitions of the earlier lectures. The matrix form looks compact and slightly mysterious; the point of this subsection is to show that it contains nothing more than those familiar definitions.

Step (a): the general definitions. For a measurement x (say coughs) and a measurement y (say sneezes) recorded on the same n patients, the earlier lectures defined

$$\mathrm{var}(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad \mathrm{cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Two remarks. First, both formulas *already* subtract the means: variance and covariance are always computed from the deviations from the mean, whether or not we have centred the data beforehand. Second, the variance is the covariance of a measurement with itself, $\mathrm{var}(x) = \mathrm{cov}(x,x)$. (We divide by $n-1$ rather than n, as in the earlier lectures, because the data are a sample; nothing below depends on this choice.)

Step (b): name the deviations. The deviations $x_i - \bar{x}$ and $y_i - \bar{y}$ are exactly the two entries of the centred vector $\mathbf{d}_i = \mathbf{x}_i - \bar{\mathbf{x}}$ in the table above. Writing d_{i1} for the coughs entry and d_{i2} for the sneezes entry, the definitions read

$$\mathrm{var}(\text{coughs}) = \frac{1}{n-1}\sum_i d_{i1}d_{i1}, \qquad \mathrm{var}(\text{sneezes}) = \frac{1}{n-1}\sum_i d_{i2}d_{i2},$$
$$\mathrm{cov}(\text{coughs},\text{sneezes}) = \frac{1}{n-1}\sum_i d_{i1}d_{i2}.$$

So the reason we may use $\mathbf{d}_i$ is not a special assumption: the deviations from the mean are what the definitions of variance and covariance are made of, and $\mathbf{d}_i$ is simply their name. Centring the data, which we do in PCA anyway, just means that from now on we work with these deviations directly.

Step (c): collect all the variances and covariances in one matrix. With p measurements there are p variances and many covariances. We arrange them in a $p \times p$ table, the **covariance matrix S**, whose entry in row j and column k is the covariance of measurement j with measurement k:

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n} d_{ij}\, d_{ik}.$$

Now look at the *outer product* $\mathbf{d}_i\mathbf{d}_i^\top$: a column vector times a row vector, which is a $p \times p$ matrix whose entry (j,k) is $d_{ij}d_{ik}$. For patient 1:

$$\mathbf{d}_1\mathbf{d}_1^\top = \begin{pmatrix} 5.917 \\ -0.833 \end{pmatrix}\begin{pmatrix} 5.917 & -0.833 \end{pmatrix} = \begin{pmatrix} 5.917^2 & 5.917\cdot(-0.833) \\ -0.833\cdot 5.917 & (-0.833)^2 \end{pmatrix} = \begin{pmatrix} 35.01 & -4.93 \\ -4.93 & 0.69 \end{pmatrix}.$$

Each entry is one patient's contribution to one variance or covariance. Adding these matrices over the patients and dividing by $n-1$ therefore adds up, entry by entry, exactly the sums of Step (b):

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}\mathbf{d}_i\mathbf{d}_i^\top = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} = \begin{pmatrix} 28.24 & 5.82 \\ 5.82 & 6.67 \end{pmatrix}.$$

The formula $\frac{1}{n-1}\sum_i \mathbf{d}_i\mathbf{d}_i^\top$ is just a compact way of computing all the variances and covariances at once.

On the diagonal are the variances of the two measurements: $s_{11} = 141.21/5 = 28.24$ for coughs and $s_{22} = 33.33/5 = 6.67$ for sneezes. Off the diagonal is their covariance, $s_{12} = s_{21} = 29.09/5 = 5.82$. $\mathbf{S}$ is **symmetric**, $\mathbf{S}^\top = \mathbf{S}$, because the covariance of coughs with sneezes is the covariance of sneezes with coughs (each outer product $\mathbf{d}_i\mathbf{d}_i^\top$ is symmetric, so their sum is too). We will use this symmetry several times.

2 The ultimate aim, and the road to it

The ultimate aim of PCA

Replace many measurements by a **few new axes** that keep **as much of the information as possible**.

What is "information" here? It is the *differences between patients*. If, along some direction, every patient had the same value, that direction would tell us nothing about who is who. The more the patients are spread out along a direction, the more that direction distinguishes them. So we measure information by **spread**, that is, by **variance**, and the aim becomes:

Find the direction along which the patients are **most spread out**. That direction is the first principal component, PC1. Then find the next best direction, and so on.

The rest of these notes is a chain of small problems. Each step exists because the previous step left a specific problem open. Keep this list in mind; it is the "progress strip" at the bottom of the slides.

1. **Score.** To compare directions, we need one number per patient along a line.
2. **Spread.** Measure how spread out those numbers are, as one formula in the direction $\mathbf{a}_1$.
3. **Fix length.** Stop the spread growing just because $\mathbf{a}_1$ is scaled up.
4. **Candidates.** Reduce infinitely many directions to a short list: the eigenvectors.
5. **Best: PC1.** Pick the candidate with the most spread: the largest eigenvalue.
6. **By hand.** Compute PC1 from $\mathbf{S}$: first the eigenvalues, then the eigenvector.
7. **Next PCs.** Add axes that bring new information, not repeats of PC1.
8. **All PCs.** See what the whole set of axes does to the data and to its spread.
9. **How many?** Decide how many axes to keep, and how much spread they hold.
10. **Least loss.** Confirm that keeping the most spread is the same as losing the least information.
11. **Algorithm.** Turn all of this into a procedure a computer can run.

3 Step 1: the score of a patient along a direction

Where we are and where we are going

Ultimate aim: the direction of greatest spread.
This step: we cannot compare directions until we can say, for a given direction, where each patient lies along it. So we need *one number per patient* for each direction: the score.

3.1 A direction is a unit vector

Take any line through the centre of the data. We describe its direction by a vector $\mathbf{a}_1$ of **length one**:

$$\|\mathbf{a}_1\| = \sqrt{a_{11}^2 + a_{21}^2} = 1.$$

We do not yet know the best $\mathbf{a}_1$: **finding it is the goal of everything that follows**. For now $\mathbf{a}_1$ is any candidate direction. (Section 5 explains why the length must be one.)

3.2 The score is a projection

The **score** of patient i along $\mathbf{a}_1$ is the dot product of $\mathbf{a}_1$ with the centred patient:

$$z_{i1} = \mathbf{a}_1^\top \mathbf{d}_i = \mathbf{a}_1^\top (\mathbf{x}_i - \bar{\mathbf{x}}) = a_{11}\,(\text{coughs}_i - 7.083) + a_{21}\,(\text{sneezes}_i - 4.833).$$

Why a dot product?

For any two vectors, $\mathbf{a}^\top \mathbf{d} = \|\mathbf{a}\|\,\|\mathbf{d}\| \cos\theta$, where θ is the angle between them (this is the geometric meaning of the dot product). Now drop a perpendicular from the point $\mathbf{d}$ onto the line of $\mathbf{a}$: it forms a right-angled triangle whose hypotenuse is $\|\mathbf{d}\|$ and whose side along the line is $\|\mathbf{d}\| \cos\theta$, the *shadow* (projection) of $\mathbf{d}$ on the line. When $\|\mathbf{a}\| = 1$, the dot product is exactly that length. So the score is exactly the position of the patient's projected point along the line, measured from the centre. This is the distance called c in the Pythagoras slides of Part 2.

Notice also what the score is as a formula: a *weighted sum* of the centred measurements, with weights a_{11} and a_{21}. These weights are what PCA calls the **loadings**: the "recipe" of the component.

Check with our patients

Using the direction we found geometrically in Part 2, $\mathbf{a}_1 = (0.970,\ 0.245)^\top$, patient 1 has

$$z_{11} = 0.970 \times 5.917 + 0.245 \times (-0.833) = 5.740 - 0.204 = 5.53.$$

The six scores are $5.53,\ 5.18,\ 3.48,\ -2.47,\ -4.40,\ -7.33$.

3.3 The mean of the scores is zero

We want to measure the *spread* of the scores, and spread (variance) is always measured around the mean. So before anything else we need the mean of the scores.

$$\bar{z}_1 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{a}_1^\top (\mathbf{x}_i - \bar{\mathbf{x}}) = \mathbf{a}_1^\top \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i - \bar{\mathbf{x}} \right) = \mathbf{a}_1^\top (\bar{\mathbf{x}} - \bar{\mathbf{x}}) = 0.$$

Why each equality holds

The average of $\mathbf{x}_i - \bar{\mathbf{x}}$ is the average of the $\mathbf{x}_i$ minus $\bar{\mathbf{x}}$, and the average of the $\mathbf{x}_i$ is $\bar{\mathbf{x}}$ by definition. The result is not a coincidence of our data: it holds for *every* direction, because we centred the data. Centring puts the origin at the centre of the cloud, so the projected points are balanced around the origin on every line.

Check with our patients

$5.53 + 5.18 + 3.48 - 2.47 - 4.40 - 7.33 = 0$ (up to rounding).

4 Step 2: the variance along a direction as one formula

Where we are and where we are going

Ultimate aim: the direction of greatest spread.
This step: we cannot try every direction one by one. We want the spread of the scores written as a single formula in $\mathbf{a}_1$, in which the data appear only through a fixed matrix. Then "find the best direction" becomes "maximise a function of $\mathbf{a}_1$".

We apply the general definition of the variance, Step (a) of Section 1.3, to the scores $z_{11}, \ldots, z_{n1}$. Because their mean is zero (Step 1), each deviation $z_{i1} - \bar{z}_1$ is just z_{i1}, so the variance is simply the average square:

$$\operatorname{var}(z_1) = \frac{1}{n-1}\sum_{i=1}^{n}(z_{i1} - \bar{z}_1)^2 = \frac{1}{n-1}\sum_{i=1}^{n} z_{i1}^2.$$

This is also SS(distances)$/(n-1)$, where SS(distances) is the sum of squared distances of the projected points from the origin that we maximised by rotating the line in Part 2. The geometric and the algebraic pictures measure the same thing.

Now we rewrite each square so that $\mathbf{a}_1$ appears at the two ends:

$$\begin{aligned}
z_{i1}^2 &= z_{i1} \cdot z_{i1} && \text{a square is a number times itself} \\
&= z_{i1} \cdot z_{i1}^\top && \text{a single number equals its own transpose} \\
&= (\mathbf{a}_1^\top \mathbf{d}_i)(\mathbf{a}_1^\top \mathbf{d}_i)^\top && \text{substitute the score} \\
&= (\mathbf{a}_1^\top \mathbf{d}_i)(\mathbf{d}_i^\top \mathbf{a}_1) && (\mathbf{BC})^\top = \mathbf{C}^\top\mathbf{B}^\top \text{ and } (\mathbf{a}_1^\top)^\top = \mathbf{a}_1 \\
&= \mathbf{a}_1^\top \mathbf{d}_i \mathbf{d}_i^\top \mathbf{a}_1 && \text{regroup: matrix products are associative.}
\end{aligned}$$

The second line deserves a word. A single number can be regarded as a 1×1 matrix, and transposing a 1×1 matrix (swapping its rows and columns) changes nothing; we write the transpose only because it lets us use the rule $(\mathbf{BC})^\top = \mathbf{C}^\top\mathbf{B}^\top$ (transposing a product reverses the order of the factors and transposes each one). With $\mathbf{B} = \mathbf{a}_1^\top$ and $\mathbf{C} = \mathbf{d}_i$ this rule gives $(\mathbf{a}_1^\top \mathbf{d}_i)^\top = \mathbf{d}_i^\top (\mathbf{a}_1^\top)^\top = \mathbf{d}_i^\top \mathbf{a}_1$, and transposing twice gives back the original vector.

Why go to this trouble?

We could have written $(\mathbf{a}_1^\top \mathbf{d}_i)^2$ and stopped. The point of the rewriting is the last line: $\mathbf{a}_1$ now sits on the *outside*, at both ends, and everything that depends on the patient, $\mathbf{d}_i\mathbf{d}_i^\top$, sits in the middle. Since $\mathbf{a}_1$ is the same for every patient, it can leave the sum, and what remains in the middle is a quantity we already know.

Summing over patients:

$$\text{var}(z_1) = \frac{1}{n-1}\sum_{i=1}^{n} \mathbf{a}_1^\top \mathbf{d}_i \mathbf{d}_i^\top \mathbf{a}_1 = \mathbf{a}_1^\top \left[\frac{1}{n-1}\sum_{i=1}^{n} \mathbf{d}_i \mathbf{d}_i^\top\right] \mathbf{a}_1 = \mathbf{a}_1^\top \mathbf{S}\, \mathbf{a}_1.$$

The middle step takes $\mathbf{a}_1^\top$ out of the sum on the left and $\mathbf{a}_1$ out on the right, by the distributive law for matrices ($\mathbf{BC}_1\mathbf{D} + \mathbf{BC}_2\mathbf{D} = \mathbf{B}(\mathbf{C}_1 + \mathbf{C}_2)\mathbf{D}$). The bracket that remains is, by Section 1.3, exactly the covariance matrix: this is why we took the trouble to bring $\mathbf{a}_1$ to the outside.

$$\text{var}(z_1) = \mathbf{a}_1^\top \mathbf{S}\, \mathbf{a}_1$$

The variance of the scores along any direction is a formula in $\mathbf{a}_1$ alone; the data enter only through $\mathbf{S}$.

Check with our patients

Along the coughs axis, $\mathbf{a} = (1, 0)^\top$: $\mathbf{a}^\top \mathbf{S}\mathbf{a} = s_{11} = 28.24$, the variance of coughs, as it should be.
Along $\mathbf{a}_1 = (0.970, 0.245)^\top$: $\mathbf{S}\mathbf{a}_1 = (28.24 \cdot 0.970 + 5.82 \cdot 0.245,\ 5.82 \cdot 0.970 + 6.67 \cdot 0.245)^\top = (28.82,\ 7.28)^\top$, and $\mathbf{a}_1^\top(\mathbf{S}\mathbf{a}_1) = 0.970 \cdot 28.82 + 0.245 \cdot 7.28 = 29.7$.
Directly from the scores: $(5.53^2 + 5.18^2 + 3.48^2 + 2.47^2 + 4.40^2 + 7.33^2)/5 = 148.7/5 = 29.7$.

So the direction we found by rotating the line in Part 2 has variance 29.7, more than the coughs axis. The question is now: *which* $\mathbf{a}_1$ *makes* $\mathbf{a}_1^\top \mathbf{S}\mathbf{a}_1$ *as large as possible?*

5 Step 3: why the length of $\mathbf{a}_1$ must be fixed

Where we are and where we are going

Ultimate aim: the direction of greatest spread.
This step: before maximising, we must make sure the maximisation is about the *direction* of $\mathbf{a}_1$ and not about its size.

Suppose we tried to maximise $\mathbf{a}^\top \mathbf{S}\mathbf{a}$ with no restriction. Multiply $\mathbf{a}$ by a number c:

$$(c\,\mathbf{a})^\top \mathbf{S}\,(c\,\mathbf{a}) = c\,\mathbf{a}^\top \mathbf{S}\, c\,\mathbf{a} = c^2\,\mathbf{a}^\top \mathbf{S}\,\mathbf{a},$$

because a number can be moved freely through matrix products ($(c\mathbf{a})^\top = c\,\mathbf{a}^\top$, and $\mathbf{S}(c\mathbf{a}) = c\,\mathbf{S}\mathbf{a}$). Doubling $\mathbf{a}$ multiplies the value by 4; ten times longer multiplies it by 100. Since c can be as large as we like, the "maximum" is infinite and the problem has no answer.

What is going on: the scale-dependence of variance

Doubling $\mathbf{a}$ keeps the same mix of coughs and sneezes, so it is the same line through the data. But it doubles every score. This is exactly the experiment of the covariance lecture, where we counted coughs over two hours instead of one: every value doubled, the relationship was unchanged, and the variance grew four times. **Variance depends on the scale.** The size of $\mathbf{a}$ plays the role of the unit of measurement of the scores: a larger $\mathbf{a}$ is a smaller ruler. If we let the size vary, directions measured with a smaller ruler would win for the wrong reason.

There are two ways to remove the scale, and they give the same answer.

- **Divide it out**, as correlation divided the covariance by the standard deviations: maximise the ratio $\dfrac{\mathbf{a}^\top \mathbf{S}\mathbf{a}}{\mathbf{a}^\top \mathbf{a}}$. Scaling $\mathbf{a}$ by c multiplies top and bottom by c^2, so the ratio depends only on the direction.

- **Fix it once and for all**: allow only vectors of length one. Then $\mathbf{a}^\top \mathbf{a} = 1$ and the two problems coincide.

$$\text{maximise } \mathbf{a}_1^\top \mathbf{S}\, \mathbf{a}_1 \quad \text{subject to} \quad \mathbf{a}_1^\top \mathbf{a}_1 = 1.$$

Why the number 1?

Any fixed length would rank the directions in the same order (length 2 would multiply every variance by 4 and leave the winner unchanged). We choose 1 because then the score is the true distance along the line, and the variance is SS(distances)$/(n-1)$, the very quantity we maximised geometrically.

A caution for later: fixing $\|\mathbf{a}_1\|$ removes the scale of $\mathbf{a}_1$, not the scale of the *measurements*. If coughs were counted over two hours, $\mathbf{S}$ itself would change and PC1 would tilt towards coughs. That is the reason for standardising the variables (Tip 1 in Part 3), which amounts to using the correlation matrix instead of $\mathbf{S}$.

6 Step 4: from "the top of the curve" to a derivative

Where we are and where we are going

Ultimate aim: the direction of greatest spread.
This step: with the length fixed, only the angle of $\mathbf{a}_1$ is free. The best direction is the top of the variance curve (Figure 1: the variance plotted against the angle of $\mathbf{a}_1$), and at the top the slope is zero. We compute that slope, so that we can set it to zero.

6.1 Only the angle is free

Every vector of length one in the plane can be written with a single angle. The tip of such a vector lies on the circle of radius one around the origin, and by the very definition of cosine and sine, the point of that circle at angle θ from the horizontal axis has coordinates $(\cos\theta, \sin\theta)$:

$$\mathbf{a}(\theta) = (\cos\theta,\ \sin\theta)^\top, \qquad \|\mathbf{a}(\theta)\|^2 = \cos^2\theta + \sin^2\theta = 1.$$

So the constraint is built in, and "choose a vector of length one" becomes "choose one number, θ". Substituting into the formula of Step 2, the variance becomes a function of one variable. Writing $c = \cos\theta$ and $s = \sin\theta$ and multiplying out,

$$\mathbf{Sa} = \begin{pmatrix} s_{11}c + s_{12}s \\ s_{12}c + s_{22}s \end{pmatrix},$$
$$\operatorname{var}(\theta) = \mathbf{a}^\top(\mathbf{Sa}) = c\,(s_{11}c + s_{12}s) + s\,(s_{12}c + s_{22}s) = s_{11}\cos^2\theta + 2s_{12}\sin\theta\cos\theta + s_{22}\sin^2\theta.$$

For our patients, $\operatorname{var}(\theta) = 28.24\cos^2\theta + 11.63\sin\theta\cos\theta + 6.67\sin^2\theta$, which is the curve plotted in Figure 1. We call it the **variance curve**.

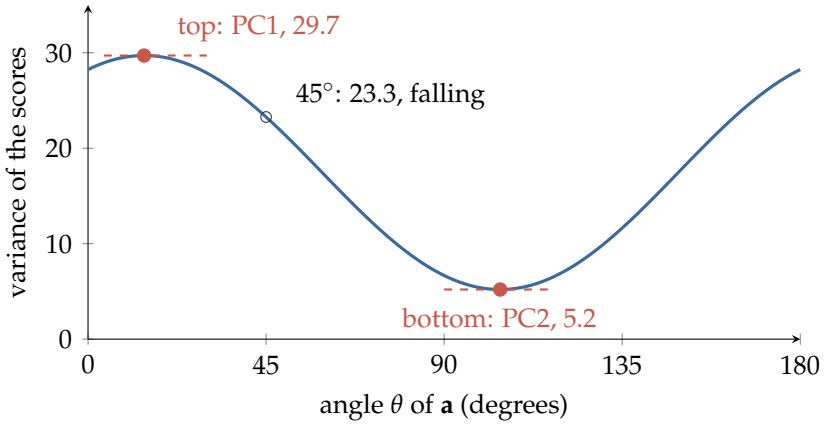


Figure 1: The **variance curve**: the variance of the patients' scores along the unit vector $\mathbf{a}(\theta) = (\cos\theta, \sin\theta)$, plotted against the angle θ. Its top (PC1) and its bottom (PC2) are the two points where the curve is flat. This curve is referred to throughout Steps 4 to 9.

The variance curve goes up and down and repeats every 180°, because θ and $\theta + 180°$ describe the same line. It does *not* grow for ever: the runaway problem of Step 3 is gone. Its highest point, at about 14.2° with variance 29.7, is PC1. Its lowest, at about 104.2° with variance 5.2, will turn out to be PC2.

6.2 The direction of turning

The variance depends on θ only through $\mathbf{a}(\theta)$, so to differentiate the variance we first need the derivative of $\mathbf{a}$ itself. Differentiating each component:

$$\mathbf{a}' = \frac{d\mathbf{a}}{d\theta} = \left(\frac{d\cos\theta}{d\theta}, \frac{d\sin\theta}{d\theta}\right)^\top = (-\sin\theta,\ \cos\theta)^\top.$$

This vector has a clear meaning. By the definition of the derivative, for a small change $\Delta\theta$ (in radians) the change in $\mathbf{a}$ is approximately the derivative times $\Delta\theta$, as for any function ($f(\theta + \Delta\theta) \approx f(\theta) + \Delta\theta\, f'(\theta)$):

$$\mathbf{a}(\theta + \Delta\theta) \approx \mathbf{a}(\theta) + \Delta\theta\, \mathbf{a}',$$

so $\mathbf{a}'$ is the **direction of turning**: the direction in which the tip of $\mathbf{a}$ moves when we turn it. It has length one, since $\sin^2\theta + \cos^2\theta = 1$, and it is perpendicular to $\mathbf{a}$, since their dot product is zero:

$$\mathbf{a}'^\top \mathbf{a} = -\sin\theta\cos\theta + \cos\theta\sin\theta = 0.$$

Turning moves $\mathbf{a}$ sideways, never making it longer or shorter, consistently with the fixed length. Geometrically, $\mathbf{a}'$ is $\mathbf{a}$ rotated by 90°: the tangent to the circle at the tip of $\mathbf{a}$.

Check with our patients

At 45°: $\mathbf{a} = (0.707, 0.707)$, $\mathbf{a}' = (-0.707, 0.707)$. Turning to 46° ($\Delta\theta = 0.01745$): exactly, $\mathbf{a}(46°) = (0.6947, 0.7193)$; with the approximation, $(0.7071, 0.7071) + 0.01745\,(-0.7071, 0.7071) = (0.6948, 0.7194)$.

6.3 The slope of the variance curve

$\text{var}(\theta) = \mathbf{a}^\top(\mathbf{S}\mathbf{a})$ is a dot product of two vectors that both change with θ: $\mathbf{u} = \mathbf{a}$ and $\mathbf{v} = \mathbf{S}\mathbf{a}$. The product rule works for dot products just as for ordinary products. Writing the dot product in components, $\mathbf{u}^\top\mathbf{v} = \sum_j u_j v_j$, and applying the usual product rule to each term,

$$\frac{d}{d\theta}\,\mathbf{u}^\top\mathbf{v} = \sum_j (u_j' v_j + u_j v_j') = \mathbf{u}'^\top\mathbf{v} + \mathbf{u}^\top\mathbf{v}'.$$

With $\mathbf{u} = \mathbf{a}$ we have $\mathbf{u}' = \mathbf{a}'$. With $\mathbf{v} = \mathbf{S}\mathbf{a}$ we need $\mathbf{v}'$: $\mathbf{S}$ contains only the data, so it does not depend on θ, and each entry of $\mathbf{S}\mathbf{a}$ is a fixed combination of the entries of $\mathbf{a}$ (for example $s_{11}a_1 + s_{12}a_2$); differentiating such a combination just differentiates the entries of $\mathbf{a}$, so $\mathbf{v}' = \mathbf{S}\mathbf{a}'$. Therefore

$$\frac{d\,\mathrm{var}}{d\theta} = \mathbf{a}'^{\top}\mathbf{S}\mathbf{a} + \mathbf{a}^{\top}\mathbf{S}\mathbf{a}'.$$

We want a single term to set to zero, and the two terms are in fact equal. The second is a single number, so it equals its transpose; transposing a product of three factors reverses their order and transposes each one, $(\mathbf{BCD})^{\top} = \mathbf{D}^{\top}\mathbf{C}^{\top}\mathbf{B}^{\top}$; and transposing $\mathbf{a}^{\top}$ gives back $\mathbf{a}$:

$$\mathbf{a}^{\top}\mathbf{S}\mathbf{a}' = (\mathbf{a}^{\top}\mathbf{S}\mathbf{a}')^{\top} = \mathbf{a}'^{\top}\mathbf{S}^{\top}(\mathbf{a}^{\top})^{\top} = \mathbf{a}'^{\top}\mathbf{S}^{\top}\mathbf{a} = \mathbf{a}'^{\top}\mathbf{S}\mathbf{a},$$

where the last step uses the symmetry $\mathbf{S}^{\top} = \mathbf{S}$ (Section 1.3). Therefore:

$$\frac{d\,\mathrm{var}}{d\theta} = 2\,\mathbf{a}'^{\top}(\mathbf{S}\mathbf{a}), \qquad \text{so the top of the variance curve is where} \quad \mathbf{a}'^{\top}(\mathbf{S}\mathbf{a}) = 0.$$

Check with our patients

At 45°: $\mathbf{S}\mathbf{a} = (28.24 \cdot 0.707 + 5.82 \cdot 0.707,\ 5.82 \cdot 0.707 + 6.67 \cdot 0.707) = (24.08, 8.83)$ and $\mathbf{a}'^{\top}(\mathbf{S}\mathbf{a}) = -0.707 \cdot 24.08 + 0.707 \cdot 8.83 = -10.79$. The slope is -21.6: not zero, and negative, so the variance is falling as θ grows past 45°, exactly as Figure 1 shows.
At 14.2°: $\mathbf{a} = (0.970, 0.245)$, $\mathbf{a}' = (-0.245, 0.970)$, $\mathbf{S}\mathbf{a} = (28.82, 7.28)$ and $\mathbf{a}'^{\top}(\mathbf{S}\mathbf{a}) = -0.245 \cdot 28.82 + 0.970 \cdot 7.28 = -7.06 + 7.06 = 0$. Flat: the top.

7 Step 5: what "flat" tells us, from the derivative to an eigenvector

Where we are and where we are going

Ultimate aim: the direction of greatest spread.
This step: we know that at the top $\mathbf{a}'^{\top}(\mathbf{S}\mathbf{a}) = 0$. We now rewrite this condition as an equation that involves only $\mathbf{a}$, $\mathbf{S}$ and a number, without $\mathbf{a}'$. That equation, $\mathbf{S}\mathbf{a} = \lambda\mathbf{a}$, has a standard method of solution and works with any number of measurements.

Why not just solve $\mathbf{a}'^{\top}(\mathbf{S}\mathbf{a}) = 0$ for θ?

In two dimensions we could. With $c = \cos\theta$, $s = \sin\theta$, $\mathbf{a}' = (-s, c)$ and $\mathbf{S}\mathbf{a} = (s_{11}c + s_{12}s,\ s_{12}c + s_{22}s)$,

$$\mathbf{a}'^{\top}(\mathbf{S}\mathbf{a}) = -s\,(s_{11}c + s_{12}s) + c\,(s_{12}c + s_{22}s) = (s_{22} - s_{11})\,sc + s_{12}\,(c^2 - s^2).$$

The double-angle identities $2sc = \sin 2\theta$ and $c^2 - s^2 = \cos 2\theta$ turn the condition $\mathbf{a}'^{\top}(\mathbf{S}\mathbf{a}) = 0$ into $\frac{1}{2}(s_{22} - s_{11})\sin 2\theta + s_{12}\cos 2\theta = 0$, that is,

$$\tan 2\theta = \frac{2s_{12}}{s_{11} - s_{22}} = \frac{11.63}{21.58} = 0.539, \qquad 2\theta = 28.3°, \quad \theta = 14.2°.$$

(This formula is classical: it appears, for example, in Kelley's 1935 method for rotating two variables.) The trouble is that this route does not extend. With p measurements a direction needs $p - 1$ angles, it can be turned in $p - 1$ perpendicular directions, and the condition becomes a system of trigonometric equations that nobody would want to solve. Moreover, $\mathbf{a}'$ is only a by-product of how we chose to write $\mathbf{a}$ (with an angle); it is a tool for finding the condition, not part of the answer. So we rewrite the condition without it.

7.1 The tool: splitting a vector along two perpendicular axes

In general, any vector in the plane is the sum of its parts along two perpendicular unit vectors. On the usual axes this is familiar:

$$\begin{pmatrix}3\\2\end{pmatrix} = 3\begin{pmatrix}1\\0\end{pmatrix} + 2\begin{pmatrix}0\\1\end{pmatrix}, \qquad 3 = (3,2)\cdot(1,0), \quad 2 = (3,2)\cdot(0,1).$$

Each coefficient is the dot product of the vector with an axis: its shadow on that axis, exactly like a score. The vectors $\mathbf{a}$ and $\mathbf{a}'$ are perpendicular and of length one, so they are such a pair of axes, only rotated. Hence, for any vector $\mathbf{v}$:

$$\mathbf{v} = (\mathbf{a}^\top \mathbf{v})\,\mathbf{a} + (\mathbf{a}'^\top \mathbf{v})\,\mathbf{a}'.$$

Why the coefficients are dot products. The dot product with a unit vector gives the projection of $\mathbf{v}$ on that vector, so $\mathbf{a}^\top\mathbf{v}$ and $\mathbf{a}'^\top\mathbf{v}$ are the shadows of $\mathbf{v}$ on the two axes. Because $\mathbf{a}$ and $\mathbf{a}'$ are perpendicular, these shadows are exactly the amounts by which we must travel along each axis to reach $\mathbf{v}$: moving along one axis does not move us along the other. With slanted axes this would fail. For example, with the axes $(1,0)$ and $(1,1)/\sqrt{2}$, at 45° to each other, the vector $\mathbf{v} = (0,1)$ has shadows 0 and $1/\sqrt{2}$, but it is reached as $-1\cdot(1,0) + \sqrt{2}\cdot(1,1)/\sqrt{2}$: the shadow on an axis and the step along it differ, because each step along one axis also moves us along the other.

7.2 Applying the tool to Sa

Sa is simply a vector: two numbers. Nobody has "made" it a combination of $\mathbf{a}$ and $\mathbf{a}'$; we *choose* to describe it on these axes, because they are the two directions that matter at the current angle: $\mathbf{a}$ is where we are looking, $\mathbf{a}'$ is where turning takes us. With $\mathbf{v} = \mathbf{Sa}$:

$$\mathbf{Sa} = (\mathbf{a}^\top\mathbf{Sa})\,\mathbf{a} + (\mathbf{a}'^\top\mathbf{Sa})\,\mathbf{a}'.$$

(This is the general fact with $\mathbf{v}$ replaced by the particular vector **Sa**: its coefficient along $\mathbf{a}$ is $\mathbf{a}^\top(\mathbf{Sa})$ and its coefficient along $\mathbf{a}'$ is $\mathbf{a}'^\top(\mathbf{Sa})$.) Both coefficients are quantities we already know:

- the coefficient along $\mathbf{a}$, $\mathbf{a}^\top\mathbf{Sa}$, is the **variance** at this angle (Step 2): the height of the variance curve in Figure 1;
- the coefficient along $\mathbf{a}'$, $\mathbf{a}'^\top\mathbf{Sa}$, is **half the slope** (Step 4).

So

$$\mathbf{Sa} = (\text{variance})\cdot\mathbf{a} + (\tfrac{1}{2}\,\text{slope})\cdot\mathbf{a}'.$$

Figure 2 shows this split at two angles.



45°: part along $\mathbf{a}' \neq 0$
not flat

PC1: part along $\mathbf{a}' = 0$
flat: $\mathbf{Sa} = 29.7\,\mathbf{a}$

Figure 2: The vector **Sa** (orange) split into its part along $\mathbf{a}$ (dashed) and its part along the direction of turning $\mathbf{a}'$ (red). Left: at 45° the red part is not zero, so the variance curve (Figure 1) is not flat there. Right: along PC1 there is no red part, so **Sa** lies on the line of $\mathbf{a}$.

Check with our patients

At 45°: $\mathbf{Sa} = 23.3\,\mathbf{a} - 10.8\,\mathbf{a}'$. Indeed $23.3\,(0.707, 0.707) - 10.8\,(-0.707, 0.707) = (16.5 + 7.6,\ 16.5 - 7.6) = (24.1, 8.8)$. The variance there is 23.3, and the slope is $2 \times (-10.8) = -21.6$. At PC1: $\mathbf{Sa} = 29.7\,\mathbf{a} + 0\,\mathbf{a}'$: variance 29.7, slope 0.

7.3 Using the condition

At the top of the variance curve (Figure 1) the slope is zero, so the coefficient in red is zero:

$$\mathbf{Sa} = (\mathbf{a}^\top \mathbf{Sa})\,\mathbf{a} + 0 \cdot \mathbf{a}' = (\mathbf{a}^\top \mathbf{Sa})\,\mathbf{a}.$$

The direction of turning has left the equation. What remains says that $\mathbf{Sa}$ lies entirely along $\mathbf{a}$: multiplying $\mathbf{a}$ by $\mathbf{S}$ does not turn it, it only stretches it. The stretch factor $\mathbf{a}^\top \mathbf{Sa}$ is just a number; we give it a short name, λ:

$$\mathbf{Sa} = \lambda\,\mathbf{a}.$$

A non-zero vector that $\mathbf{S}$ does not turn, only stretches, is called an **eigenvector** of $\mathbf{S}$; the stretch factor λ is its **eigenvalue**.

Why rename the number λ? The logic and the mathematics

Renaming is not a new assumption: $\mathbf{Sa} = (\mathbf{a}^\top \mathbf{Sa})\,\mathbf{a}$ and $\mathbf{Sa} = \lambda\mathbf{a}$ with $\lambda = \mathbf{a}^\top \mathbf{Sa}$ say the same thing. The renaming matters because it changes what we treat as unknown. As written, $\mathbf{Sa} = (\mathbf{a}^\top \mathbf{Sa})\mathbf{a}$ is awkward: $\mathbf{a}$ appears three times on the right, and the number in front depends on the very vector we are looking for. If instead we treat λ as a second unknown and look for a number λ and a vector $\mathbf{a} \neq \mathbf{0}$ with $\mathbf{Sa} = \lambda\mathbf{a}$, then for each fixed λ the equation is *linear* in $\mathbf{a}$, and linear algebra tells us exactly when it has a solution (Step 7). Is this cheating? No: any solution automatically satisfies $\lambda = \mathbf{a}^\top \mathbf{Sa}$, as the next subsection shows.

7.4 The general case: any number of measurements

Everything above was drawn in the plane, where there is only one direction of turning. With p measurements, a unit vector $\mathbf{a}$ can be turned in $p - 1$ perpendicular directions $\mathbf{u}$ (with three measurements, $\mathbf{a}$ moves on a sphere and can be turned in two independent directions). The top is where the slope is zero in every one of these directions:

$$\mathbf{u}^\top(\mathbf{Sa}) = 0 \quad \text{for every } \mathbf{u} \perp \mathbf{a} \qquad \Longrightarrow \qquad \mathbf{Sa} = \lambda\,\mathbf{a}.$$

Splitting $\mathbf{Sa}$ into its part along $\mathbf{a}$ and its parts along the $\mathbf{u}$'s, all the latter are zero, so again all of $\mathbf{Sa}$ lies along $\mathbf{a}$. The picture changes; the conclusion does not. This is why the eigenvector equation, unlike the angle calculation, works with any number of measurements.

8 Step 6: what the eigenvalue tells us, and the choice of PC1

Where we are and where we are going

Ultimate aim: the direction of greatest spread.
This step: every flat point of the variance curve (Figure 1), that is, every angle at which its slope is zero, gives an eigenvector. We find out what the number λ means, and use it to choose among the eigenvectors.

We already know the answer from Step 5, where λ was defined as $\mathbf{a}^\top \mathbf{S}\mathbf{a}$; but it is worth seeing it directly from the equation, because from now on we will treat $\mathbf{S}\mathbf{a} = \lambda\mathbf{a}$ as the starting point. Multiply both sides of $\mathbf{S}\mathbf{a} = \lambda\mathbf{a}$ on the left by the row vector $\mathbf{a}^\top$; on the right, the number λ moves to the front, and $\mathbf{a}^\top\mathbf{a} = 1$ because $\mathbf{a}$ has length one:

$$\mathbf{a}^\top \mathbf{S}\mathbf{a} = \mathbf{a}^\top(\lambda\mathbf{a}) = \lambda\,\mathbf{a}^\top\mathbf{a} = \lambda.$$

The left-hand side is the variance of the scores along $\mathbf{a}$ (Step 2). So var $= \lambda$.

The eigenvalue is the variance of the scores along its eigenvector: the height of the variance curve (Figure 1) at that flat point. Therefore

PC1 is the eigenvector of $\mathbf{S}$ with the **largest** eigenvalue λ_1, $\qquad \mathbf{S}\mathbf{a}_1 = \lambda_1\mathbf{a}_1$.

Why the largest?

The best direction must be a flat point of the variance curve. Every flat point is an eigenvector, and its variance is its eigenvalue. So the best direction is the eigenvector with the largest eigenvalue.
In more detail: the best direction is the top of the curve, and the top is a flat point because the curve is smooth and repeats every 180°, so it does have a highest point, where its slope is zero. In two dimensions the variance curve (Figure 1) has two flat points, its top and its bottom, and both satisfy $\mathbf{S}\mathbf{a} = \lambda\mathbf{a}$. The equation above, variance $= \lambda$, tells us the height at each. We want the most spread, so we take the one with the largest λ. We no longer search over angles: we solve $\mathbf{S}\mathbf{a} = \lambda\mathbf{a}$ directly (Step 7), which gives all the flat points at once, and pick the largest.

Check with our patients

The direction found by rotating the line in Part 2 passes the test: $\mathbf{S}\,(0.970, 0.245)^\top = (28.8, 7.3)^\top = 29.7\,(0.970, 0.245)^\top$. $\mathbf{S}$ only stretches it, by 29.7, which is more than any other direction (the coughs axis alone gives 28.2). This is a *check*; the next step computes PC1 from $\mathbf{S}$ alone.

8.1 How to read $\mathbf{S}\mathbf{a} = \lambda\mathbf{a}$

Since this equation is the heart of PCA, it is worth pausing on how to read it and what to take from it.

Reading it. The left side, $\mathbf{S}\mathbf{a}$, is a matrix times a vector: a new vector which, in general, points somewhere else, because a matrix both turns and stretches the vectors it multiplies. The right side, $\lambda\mathbf{a}$, is the *same* vector $\mathbf{a}$ multiplied by a number: it points along the same line as $\mathbf{a}$, only longer or shorter. So the equation says: *for this particular* $\mathbf{a}$*, multiplying by* $\mathbf{S}$ *does not turn it; it only stretches it, by the factor* λ. Note that both $\mathbf{a}$ and λ are unknown: we are not given $\mathbf{a}$ and asked for λ; we are asking which directions $\mathbf{S}$ leaves unturned, and by how much it stretches each.

$\mathbf{a}$	$\mathbf{S}\mathbf{a}$	unturned?
$(0.707,\ 0.707)$, at 45°	$(24.1,\ 8.8)$, at about 20°	no: $\mathbf{S}$ has turned it towards PC1
$(0.970,\ 0.245)$, at 14.2°	$(28.8,\ 7.3) = 29.7\,(0.970,\ 0.245)$	yes: stretched by 29.7

What to take from it.

1. **It marks the flat points of the variance curve.** By Step 5, the part of $\mathbf{S}\mathbf{a}$ along the direction of turning is half the slope of the variance curve (Figure 1). "$\mathbf{S}\mathbf{a}$ points along $\mathbf{a}$" therefore means

"the slope is zero". Every solution is a candidate for the best direction, a top or a bottom of the curve. The equation alone does not say which candidate is best.

2. λ **is the variance along that direction.** As shown above, $\mathbf{a}^\top \mathbf{S}\mathbf{a} = \lambda$, so the stretch factor is not merely a number from linear algebra: it measures how spread out the patients are along $\mathbf{a}$. This is how we choose: PC1 is the solution with the largest λ (29.7), PC2 the next (5.2).

3. **A statistical reading: each measurement co-varies with the score in proportion to its own weight.** Entry j of $\mathbf{S}\mathbf{a}$ is the covariance between measurement j and the score z along $\mathbf{a}$:

$$(\mathbf{S}\mathbf{a})_j = \frac{1}{n-1}\sum_{i=1}^{n} d_{ij}\,(\mathbf{a}^\top \mathbf{d}_i) = \frac{1}{n-1}\sum_{i=1}^{n} d_{ij}\,z_i = \operatorname{cov}(x_j,\, z),$$

because entry j of $\mathbf{S}\mathbf{a} = \frac{1}{n-1}\sum_i \mathbf{d}_i(\mathbf{d}_i^\top \mathbf{a})$ is $\frac{1}{n-1}\sum_i d_{ij}z_i$, and both the d_{ij} and the scores have mean zero. So $\mathbf{S}\mathbf{a} = \lambda\mathbf{a}$ says that the covariances of coughs and sneezes with the PC1 score (28.8 and 7.3) are in the same proportion as the recipe of PC1 (0.970 and 0.245). The direction is *self-consistent*: the weights that define the score are reproduced by how the measurements move together with that score. For any other direction the two disagree, and turning towards the measurement that co-varies more with the score increases the variance. This is also why the loadings can be read as how strongly each measurement is associated with the component.

4. **What the equation does not fix.** The length of $\mathbf{a}$: if $\mathbf{a}$ is a solution, so are $2\mathbf{a}$ and $-\mathbf{a}$. That is why we add $\|\mathbf{a}\| = 1$, and why software may report the loadings with the opposite sign. Nor does a single solution exist: with p measurements there are p solutions, mutually perpendicular (Step 8), which are the p principal components ordered by λ.

Check with our patients

Covariance of coughs with the PC1 scores: $\frac{1}{5}\,(5.917 \cdot 5.53 + 4.917 \cdot 5.18 + 2.917 \cdot 3.48 + 2.083 \cdot 2.47 + 5.083 \cdot 4.40 + 6.583 \cdot 7.33) = \frac{144.1}{5} = 28.8$, the first entry of $\mathbf{S}\mathbf{a}_1$; and $28.8/29.7 = 0.970$, the first loading.

In one sentence: $\mathbf{S}\mathbf{a} = \lambda\mathbf{a}$ says that $\mathbf{a}$ is a direction the covariance matrix only stretches, never turns. Such directions are the flat points of the variance curve, λ is the variance along each, and PCA keeps the ones with the largest λ.

9 Step 7: computing PC1 by hand

Where we are and where we are going

Ultimate aim: the direction of greatest spread.
This step: we know that PC1 is the eigenvector of $\mathbf{S}$ with the largest eigenvalue. We compute it from $\mathbf{S}$ alone: first all the eigenvalues, then the eigenvector of the largest.

9.1 The eigenvalues

Start from the eigenvector equation and move everything to one side: $\mathbf{S}\mathbf{a} - \lambda\mathbf{a} = \mathbf{0}$. We would like to factor out $\mathbf{a}$, but $\mathbf{S} - \lambda$ makes no sense: one cannot subtract a number from a matrix. The identity matrix $\mathbf{I}$ solves this, because it leaves every vector unchanged, $\mathbf{I}\mathbf{a} = \mathbf{a}$, so $\lambda\mathbf{a} = \lambda\mathbf{I}\mathbf{a}$:

$$\mathbf{S}\mathbf{a} = \lambda\mathbf{a} \iff \mathbf{S}\mathbf{a} - \lambda\mathbf{I}\mathbf{a} = \mathbf{0} \iff (\mathbf{S} - \lambda\mathbf{I})\,\mathbf{a} = \mathbf{0}.$$

In two dimensions, $\lambda\mathbf{I} = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}$, so subtracting it simply subtracts λ from the diagonal:

$$\mathbf{S} - \lambda\mathbf{I} = \begin{pmatrix} 28.24 - \lambda & 5.82 \\ 5.82 & 6.67 - \lambda \end{pmatrix}.$$

Why set the determinant to zero?

Read $(\mathbf{S} - \lambda\mathbf{I})\mathbf{a} = \mathbf{0}$ literally: multiply the matrix $\mathbf{S} - \lambda\mathbf{I}$ by the vector $\mathbf{a}$, and the result is the zero vector $\mathbf{0} = (0, 0)$. The vector $\mathbf{a} = \mathbf{0}$ always does this, for every λ, but it is not a direction. So the question is: for which λ is there a *non-zero* $\mathbf{a}$ that the matrix turns into the zero vector?
An invertible matrix never does this. Write $\mathbf{M} = \mathbf{S} - \lambda\mathbf{I}$. If $\mathbf{M}$ had an inverse $\mathbf{M}^{-1}$, we could multiply both sides of $\mathbf{Ma} = \mathbf{0}$ by it and get $\mathbf{a} = \mathbf{M}^{-1}\mathbf{0} = \mathbf{0}$, contradicting $\mathbf{a} \neq \mathbf{0}$. So $\mathbf{M}$ must have no inverse: it must be **singular**. A standard fact of linear algebra is that a square matrix is singular exactly when its determinant is zero.
Why this is the determinant, for a 2×2 matrix. Write $\mathbf{M} = \begin{pmatrix} p & q \\ r & t \end{pmatrix}$ and $\mathbf{a} = (a_1, a_2)$. Then

$$\mathbf{Ma} = a_1 \begin{pmatrix} p \\ r \end{pmatrix} + a_2 \begin{pmatrix} q \\ t \end{pmatrix},$$

a combination of the two columns. This combination can be the zero vector with $\mathbf{a} \neq \mathbf{0}$ only if the two columns point along the same line (one cancels the other), that is, when $p/r = q/t$, which is $pt - qr = 0$: the determinant.
With our numbers. For $\lambda_1 = 29.71$,

$$(\mathbf{S} - \lambda_1\mathbf{I}) \begin{pmatrix} 0.970 \\ 0.245 \end{pmatrix} = \begin{pmatrix} -1.47 \cdot 0.970 + 5.82 \cdot 0.245 \\ 5.82 \cdot 0.970 - 23.04 \cdot 0.245 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}:$$

a non-zero vector turned into zero, and indeed $\det(\mathbf{S} - \lambda_1\mathbf{I}) = 0$. For a value that is not an eigenvalue, say $\lambda = 10$,

$$\det \begin{pmatrix} 18.24 & 5.82 \\ 5.82 & -3.33 \end{pmatrix} = -60.74 - 33.87 = -94.6 \neq 0,$$

and the only solution is $\mathbf{a} = \mathbf{0}$.
Hence the condition $\det(\mathbf{S} - \lambda\mathbf{I}) = 0$, the **characteristic equation**. It involves λ alone, so we can solve it first and find $\mathbf{a}$ afterwards.

For a 2×2 matrix the determinant is the product of the diagonal minus the product of the off-diagonal:

$$(28.24 - \lambda)(6.67 - \lambda) - 5.82^2 = 0.$$

Multiplying out the product of the diagonal,

$$(28.24 - \lambda)(6.67 - \lambda) = \lambda^2 - (28.24 + 6.67)\,\lambda + 28.24 \times 6.67 = \lambda^2 - 34.91\,\lambda + 188.36,$$

and subtracting $5.82^2 = 33.87$ gives a quadratic in λ. Its coefficient of λ is minus the trace (the sum of the diagonal, $28.24 + 6.67$) and its constant term is the determinant of $\mathbf{S}$ ($188.36 - 33.87 = 154.49$):

$$\lambda^2 - 34.91\,\lambda + 154.49 = 0.$$

The usual formula for a quadratic $\lambda^2 + b\lambda + c = 0$, namely $\lambda = (-b \pm \sqrt{b^2 - 4c})/2$, gives

$$\lambda = \frac{34.91 \pm \sqrt{34.91^2 - 4 \times 154.49}}{2} = \frac{34.91 \pm \sqrt{1218.71 - 617.96}}{2}$$
$$= \frac{34.91 \pm \sqrt{600.75}}{2} = \frac{34.91 \pm 24.51}{2}.$$

$$\lambda_1 = 29.7, \qquad \lambda_2 = 5.2.$$

These are exactly the top and the bottom of the variance curve in Figure 1.

Check with our patients

$\lambda_1 + \lambda_2 = 34.91 = \text{tr}(\mathbf{S}) = 28.24 + 6.67$, and $\lambda_1 \lambda_2 = 154.49 = \det(\mathbf{S})$. These two identities hold for every covariance matrix and are a quick way to catch arithmetic errors.

9.2 The eigenvector

Now put $\lambda_1 = 29.71$ back into $(\mathbf{S} - \lambda_1 \mathbf{I})\mathbf{a}_1 = \mathbf{0}$ and write out the two rows:

$$(28.24 - 29.71)\, a_{11} + 5.82\, a_{21} = 0 \quad \Longrightarrow \quad \frac{a_{21}}{a_{11}} = \frac{1.47}{5.82} = 0.25,$$
$$5.82\, a_{11} + (6.67 - 29.71)\, a_{21} = 0 \quad \Longrightarrow \quad \frac{a_{21}}{a_{11}} = \frac{5.82}{23.04} = 0.25.$$

Both rows give the same ratio. This is not a coincidence: we chose λ_1 precisely so that the matrix is singular, which means that its two rows say the same thing. So, in practice, one row is enough. This fixes only the *direction*: about 1 part sneezes for every 4 parts coughs, the "recipe" of PC1. The constraint $\|\mathbf{a}_1\| = 1$ fixes the length. Any vector with this ratio, for example $(1, 0.25)$, points the right way; dividing a vector by its own length gives a vector of length one in the same direction:

$$\|(1,\ 0.25)\| = \sqrt{1 + 0.25^2} = \sqrt{1.0625} = 1.031, \qquad \mathbf{a}_1 = \frac{(1,\ 0.25)^\top}{1.031} = (0.970,\ 0.245)^\top.$$

PC1 is found: $\mathbf{a}_1 = (0.970,\ 0.245)^\top$, with variance $\lambda_1 = 29.7$. It is exactly the direction we found geometrically in Part 2 by rotating the line.

The same steps with $\lambda_2 = 5.2$ give $\mathbf{a}_2 = (-0.245,\ 0.970)^\top$, perpendicular to $\mathbf{a}_1$. We will meet it next, as PC2.

A note on signs

If $\mathbf{a}$ is an eigenvector, so is $-\mathbf{a}$: $\mathbf{S}(-\mathbf{a}) = \lambda(-\mathbf{a})$, and it has the same length and points along the same line. So $-\mathbf{a}_1$ is an equally valid answer, and different software may report loadings with opposite signs, without any change in meaning.

10 Step 8: the next best direction

Where we are and where we are going

Ultimate aim: a few axes that keep as much spread as possible.
This step: one axis rarely keeps enough. The second axis must add *new* information, not repeat PC1. We translate "new information" into mathematics and solve.

"Not repeating PC1" means that the scores on the second axis, $z_{i2} = \mathbf{a}_2^\top \mathbf{d}_i$, are **uncorrelated** with the scores on PC1. Their covariance follows from the general definition of Section 1.3 and the same algebra as the variance in Step 2. Both sets of scores have mean zero (Step 1 holds for any direction), so each deviation is the score itself, and

$$z_{i1}z_{i2} = (\mathbf{a}_1^\top \mathbf{d}_i)(\mathbf{a}_2^\top \mathbf{d}_i) = (\mathbf{a}_1^\top \mathbf{d}_i)(\mathbf{d}_i^\top \mathbf{a}_2) = \mathbf{a}_1^\top\, \mathbf{d}_i\mathbf{d}_i^\top\, \mathbf{a}_2,$$

using once more that a number equals its transpose. Summing over patients and taking $\mathbf{a}_1^\top$ and $\mathbf{a}_2$ out of the sum, exactly as in Step 2:

$$\operatorname{cov}(z_1, z_2) = \frac{1}{n-1}\sum_{i=1}^{n} z_{i1}z_{i2} = \mathbf{a}_1^\top \left[\frac{1}{n-1}\sum_{i=1}^{n} \mathbf{d}_i\mathbf{d}_i^\top\right]\mathbf{a}_2 = \mathbf{a}_1^\top \mathbf{S}\, \mathbf{a}_2.$$

Now use the two facts we have about $\mathbf{S}$ and $\mathbf{a}_1$. By the transpose rule and the symmetry of $\mathbf{S}$, $\mathbf{a}_1^\top \mathbf{S} = (\mathbf{S}^\top \mathbf{a}_1)^\top = (\mathbf{S}\mathbf{a}_1)^\top$; and $\mathbf{a}_1$ is an eigenvector, so $\mathbf{S}\mathbf{a}_1 = \lambda_1 \mathbf{a}_1$:

$$\mathbf{a}_1^\top \mathbf{S}\mathbf{a}_2 = (\mathbf{S}\mathbf{a}_1)^\top \mathbf{a}_2 = (\lambda_1 \mathbf{a}_1)^\top \mathbf{a}_2 = \lambda_1\, \mathbf{a}_1^\top \mathbf{a}_2.$$

So, since $\lambda_1 > 0$,

$$\operatorname{cov}(z_1, z_2) = 0 \iff \mathbf{a}_1^\top \mathbf{a}_2 = 0\,:$$

uncorrelated means perpendicular. The second problem is therefore the first one with one more constraint:

$$\text{maximise } \mathbf{a}_2^\top \mathbf{S}\mathbf{a}_2 \quad \text{subject to} \quad \mathbf{a}_2^\top \mathbf{a}_2 = 1,\ \ \mathbf{a}_2^\top \mathbf{a}_1 = 0.$$

Why is the answer again an eigenvector? Apply the argument of Step 5 to this new problem. Now $\mathbf{a}_2$ may only turn within the directions perpendicular to $\mathbf{a}_1$, so at the best $\mathbf{a}_2$ the slope is zero along every allowed direction of turning $\mathbf{u}$ (perpendicular to both $\mathbf{a}_2$ and $\mathbf{a}_1$): $\mathbf{u}^\top(\mathbf{S}\mathbf{a}_2) = 0$. What about the one direction we are *not* allowed to turn towards, $\mathbf{a}_1$? There the part of $\mathbf{S}\mathbf{a}_2$ is $\mathbf{a}_1^\top \mathbf{S}\mathbf{a}_2 = \lambda_1 \mathbf{a}_1^\top \mathbf{a}_2 = 0$ as well, by the line above. So $\mathbf{S}\mathbf{a}_2$ has no part along any direction perpendicular to $\mathbf{a}_2$, and therefore lies along $\mathbf{a}_2$: $\mathbf{S}\mathbf{a}_2 = \lambda_2 \mathbf{a}_2$. Its variance is λ_2 (Step 6), and since it cannot be $\mathbf{a}_1$, the best choice is the eigenvector with the **second largest** eigenvalue. In general,

$$\mathbf{S}\mathbf{a}_k = \lambda_k \mathbf{a}_k, \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0,$$

one principal component per eigenvector, in decreasing order of variance. (The eigenvalues of a covariance matrix are never negative, because each is a variance.)

Check with our patients

In two dimensions there is only one direction perpendicular to $\mathbf{a}_1$: $\mathbf{a}_2 = (-0.245, 0.970)^\top$, with $\lambda_2 = 5.2$. Indeed $\mathbf{a}_1^\top \mathbf{a}_2 = 0.970 \cdot (-0.245) + 0.245 \cdot 0.970 = 0$.

11 Step 9: all the components together

Where we are and where we are going

Ultimate aim: a few axes that keep as much spread as possible.
This step: we now have one direction per eigenvector. Before choosing how many to keep, we look at what the whole set does to the data and to its spread. A matrix notation makes this compact.

11.1 The matrix A: all the scores at once

A is not a new object. It is the eigenvectors we already have, placed side by side as columns:

$$\mathbf{A} = [\,\mathbf{a}_1\ \mathbf{a}_2\ \cdots\ \mathbf{a}_p\,], \qquad \mathbf{z}_i = \mathbf{A}^\top(\mathbf{x}_i - \bar{\mathbf{x}}).$$

Its only purpose is bookkeeping. The rows of $\mathbf{A}^\top$ are the columns of **A** laid flat, that is, row k of $\mathbf{A}^\top$ is $\mathbf{a}_k^\top$. A matrix times a vector multiplies each row by the vector, so entry k of $\mathbf{A}^\top\mathbf{d}_i$ is $\mathbf{a}_k^\top\mathbf{d}_i$, which is z_{ik}, the score of patient i on PC k. One multiplication gives all the scores of a patient.

Check with our patients

With $\mathbf{A} = \begin{pmatrix} 0.970 & -0.245 \\ 0.245 & 0.970 \end{pmatrix}$, patient 1 gets

$$\mathbf{z}_1 = \mathbf{A}^\top \begin{pmatrix} 5.917 \\ -0.833 \end{pmatrix} = \begin{pmatrix} 0.970 \cdot 5.917 + 0.245 \cdot (-0.833) \\ -0.245 \cdot 5.917 + 0.970 \cdot (-0.833) \end{pmatrix} = \begin{pmatrix} 5.53 \\ -2.26 \end{pmatrix}:$$

the PC1 score of Step 1, plus the PC2 score. Together they are the patient's position on the PCA plot.

11.2 A is a rotation

Entry (j,k) of $\mathbf{A}^\top\mathbf{A}$ is row j of $\mathbf{A}^\top$ times column k of **A**, that is, $\mathbf{a}_j^\top\mathbf{a}_k$. The columns of **A** have length one ($\mathbf{a}_j^\top\mathbf{a}_j = 1$) and are mutually perpendicular ($\mathbf{a}_j^\top\mathbf{a}_k = 0$ for $j \neq k$), so the diagonal entries are 1 and all the others are 0:

$$\mathbf{A}^\top\mathbf{A} = \mathbf{I}.$$

This says that $\mathbf{A}^\top$ undoes **A**: $\mathbf{A}^\top$ is the inverse of **A**. For a square matrix, an inverse works on both sides, so also $\mathbf{A}\mathbf{A}^\top = \mathbf{I}$. A matrix with this property is a **rotation** (possibly combined with a reflection). In two dimensions

$$\mathbf{A} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \qquad \theta = 14.2^\circ,$$

the angle at the top of the variance curve (Figure 1); its two columns are exactly the **a** and **a**′ of Steps 4 and 5 at that angle. The scores $\mathbf{z}_i = \mathbf{A}^\top\mathbf{d}_i$ are the coordinates of each patient on the rotated axes. **PCA rotates the axes onto the directions of greatest spread.**

11.3 The covariance of the new variables

The new variables are the scores $\mathbf{z}_i$. Their covariance matrix is built exactly as **S** was built from the $\mathbf{d}_i$ in Section 1.3 (their mean is zero, by Step 1), and the same algebra as in Step 2, now with the matrix **A** in place of the vector $\mathbf{a}_1$, gives

$$\mathrm{Cov}(\mathbf{z}) = \frac{1}{n-1}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{z}_i^\top = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{A}^\top\mathbf{d}_i)(\mathbf{A}^\top\mathbf{d}_i)^\top = \frac{1}{n-1}\sum_{i=1}^{n}\mathbf{A}^\top\mathbf{d}_i\mathbf{d}_i^\top\mathbf{A} = \mathbf{A}^\top\mathbf{S}\mathbf{A},$$

where $(\mathbf{A}^\top \mathbf{d}_i)^\top = \mathbf{d}_i^\top \mathbf{A}$ by the transpose rule. Entry (j,k) of $\mathbf{A}^\top \mathbf{S}\mathbf{A}$ is row j of $\mathbf{A}^\top$ times $\mathbf{S}$ times column k of $\mathbf{A}$, that is, $\mathbf{a}_j^\top \mathbf{S}\mathbf{a}_k$. Using $\mathbf{S}\mathbf{a}_k = \lambda_k \mathbf{a}_k$, this is $\lambda_k\, \mathbf{a}_j^\top \mathbf{a}_k$, which is λ_k when $j = k$ (unit length) and 0 otherwise (perpendicular eigenvectors). Hence

$$\mathrm{Cov}(\mathbf{z}) = \mathbf{A}^\top \mathbf{S}\mathbf{A} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p) = \Lambda,$$

where Λ (capital lambda) denotes the diagonal matrix of eigenvalues. The new variables are all **uncorrelated**, and their variances are the eigenvalues.

11.4 The eigendecomposition $\mathbf{S} = \mathbf{A}\Lambda\mathbf{A}^\top$

Undo the rotation: multiply $\mathbf{A}^\top \mathbf{S}\mathbf{A} = \Lambda$ by $\mathbf{A}$ on the left and by $\mathbf{A}^\top$ on the right, and use $\mathbf{A}\mathbf{A}^\top = \mathbf{I}$:

$$\mathbf{A}\mathbf{A}^\top \mathbf{S}\, \mathbf{A}\mathbf{A}^\top = \mathbf{A}\Lambda\mathbf{A}^\top \quad \Longrightarrow \quad \boxed{\mathbf{S} = \mathbf{A}\,\Lambda\,\mathbf{A}^\top}.$$

A second route gives the same result directly from the eigenvector equations. Column j of $\mathbf{S}\mathbf{A}$ is $\mathbf{S}\mathbf{a}_j$, and column j of $\mathbf{A}\Lambda$ is $\lambda_j \mathbf{a}_j$ (multiplying by a diagonal matrix on the right scales each column). So writing $\mathbf{S}\mathbf{a}_j = \lambda_j \mathbf{a}_j$ for all j at once, column by column, gives $\mathbf{S}\mathbf{A} = \mathbf{A}\Lambda$; multiplying on the right by $\mathbf{A}^\top$ gives $\mathbf{S} = \mathbf{A}\Lambda\mathbf{A}^\top$.

Read from right to left, the formula says: rotate into the PC axes ($\mathbf{A}^\top$), stretch each axis by its eigenvalue (Λ), rotate back ($\mathbf{A}$). The covariance matrix is completely described by its directions and its spreads.

Check with our patients

$$\mathbf{A}\Lambda = \begin{pmatrix} 0.970 & -0.245 \\ 0.245 & 0.970 \end{pmatrix} \begin{pmatrix} 29.7 & 0 \\ 0 & 5.2 \end{pmatrix} = \begin{pmatrix} 28.81 & -1.27 \\ 7.28 & 5.04 \end{pmatrix},$$

$$(\mathbf{A}\Lambda)\mathbf{A}^\top = \begin{pmatrix} 28.26 & 5.82 \\ 5.82 & 6.68 \end{pmatrix} \approx \mathbf{S}.$$

11.5 The total spread, and the budget we will split

The trace (sum of the diagonal) has a useful property: $\mathrm{tr}(\mathbf{B}\mathbf{C}) = \mathrm{tr}(\mathbf{C}\mathbf{B})$, because both equal $\sum_j \sum_k B_{jk} C_{kj}$. Applying it with $\mathbf{B} = \mathbf{A}$ and $\mathbf{C} = \Lambda\mathbf{A}^\top$, and then using $\mathbf{A}^\top\mathbf{A} = \mathbf{I}$:

$$\mathrm{tr}(\mathbf{S}) = \mathrm{tr}(\mathbf{A}\Lambda\mathbf{A}^\top) = \mathrm{tr}(\Lambda\mathbf{A}^\top\mathbf{A}) = \mathrm{tr}(\Lambda) = \lambda_1 + \cdots + \lambda_p.$$

The left-hand side is the sum of the variances of the original measurements; the right-hand side is the sum of the variances of the principal components. The rotation **keeps** all the spread and only redistributes it, piling as much as possible onto the first axes. If we now keep only the first k components, this total splits into two parts:

$$\mathrm{tr}(\mathbf{S}) = \underbrace{\sum_{j \le k} \lambda_j}_{\text{kept}} + \underbrace{\sum_{j > k} \lambda_j}_{\text{discarded}}.$$

The blue part will reappear as the variation retained (Step 10), the red part as the reconstruction error (Step 11).

Check with our patients

$28.24 + 6.67 = 34.91 = 29.71 + 5.20$. Keeping PC1 ($k = 1$) keeps 29.7 and discards 5.2.

12 Step 10: how many components to keep

Where we are and where we are going

Ultimate aim: a *few* axes that keep as much spread as possible.
This step: back to the objective. All p components together are only a rotation. Reduction happens when we keep only the first k; we need to know how much spread they hold.

With all p components, $\mathbf{z} = \mathbf{A}^\top(\mathbf{x} - \bar{\mathbf{x}})$. Multiplying both sides by $\mathbf{A}$ and using $\mathbf{A}\mathbf{A}^\top = \mathbf{I}$ gives $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{A}^\top(\mathbf{x} - \bar{\mathbf{x}}) = \mathbf{x} - \bar{\mathbf{x}}$, so $\mathbf{x} = \bar{\mathbf{x}} + \mathbf{A}\mathbf{z}$ exactly: nothing is lost, so PCA by itself is *not* dimensionality reduction. It becomes dimensionality reduction when we keep only the first $k < p$ columns, $\mathbf{A}_k = [\mathbf{a}_1 \cdots \mathbf{a}_k]$:

$$\mathbf{z}_{(k)} = \mathbf{A}_k^\top(\mathbf{x} - \bar{\mathbf{x}}) \qquad (k \text{ numbers per sample instead of } p).$$

The price is the variation discarded, and we can measure it exactly. The share of the total spread that we keep is

$$\text{variation retained} = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j}.$$

Check with our patients

Six patients, $2 \to 1$ dimension: $29.7/34.9 = 85\%$ of the variation is kept. For the 18 patients of Part 1, going from 6 measurements to 2 components keeps 92%.

Feature extraction, not feature selection

Each principal component combines *all* the measurements (its loadings are usually all non-zero). Reducing six measurements to two components does not mean we can stop measuring four of them: every new patient still needs all six to compute the scores.

13 Step 11: most spread is least loss

Where we are and where we are going

Ultimate aim: keep as much *information* as possible.
This step: we chose the axes by maximising spread. Is that really the same as losing the least information? We measure the loss directly and find that it is exactly the discarded (red) eigenvalues.

Keeping k components, each patient is summarised by the scores $z_{i1}, \ldots, z_{ik}$, where $z_{ij} = \mathbf{a}_j^\top(\mathbf{x}_i - \bar{\mathbf{x}})$ is the score of patient i on PC j (entry j of $\mathbf{z}_i$). Going back to the original measurements, start from the exact formula of Step 10, $\mathbf{x}_i = \bar{\mathbf{x}} + \mathbf{A}\mathbf{z}_i$. Written column by column ($\mathbf{A}\mathbf{z}_i$ is the sum of the columns of $\mathbf{A}$, each multiplied by the corresponding entry of $\mathbf{z}_i$), it reads

$$\mathbf{x}_i = \bar{\mathbf{x}} + \sum_{j=1}^{p} z_{ij}\,\mathbf{a}_j :$$

start at the centre, then move z_{ij} steps along each direction $\mathbf{a}_j$. If we only kept the first k scores, the natural reconstruction uses only those:

$$\hat{\mathbf{x}}_i = \bar{\mathbf{x}} + \sum_{j=1}^{k} z_{ij}\,\mathbf{a}_j,$$

which lands on the projected point. Subtracting the two formulas, what is lost is the discarded part,

$$\mathbf{x}_i - \hat{\mathbf{x}}_i = \sum_{j>k} z_{ij}\,\mathbf{a}_j.$$

Its squared length is the dot product with itself. Multiplying out the sum, every cross term contains $\mathbf{a}_j^\top \mathbf{a}_l$ with $j \neq l$, which is zero, and every square term contains $\mathbf{a}_j^\top \mathbf{a}_j = 1$. This is Pythagoras for perpendicular directions:

$$\|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2 = \Big(\sum_{j>k} z_{ij}\mathbf{a}_j\Big)^\top \Big(\sum_{l>k} z_{il}\mathbf{a}_l\Big) = \sum_{j>k} z_{ij}^2.$$

Summing over patients, and using $\sum_i z_{ij}^2 = (n-1)\lambda_j$ (the definition of the variance of the scores on PC j, Steps 2 and 6):

$$\sum_{i=1}^{n} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2 = (n-1)\sum_{j>k} \lambda_j.$$

The reconstruction error is exactly the discarded eigenvalues. By Step 9, the discarded part equals $\mathrm{tr}(\mathbf{S}) - \sum_{j\le k}\lambda_j$, and $\mathrm{tr}(\mathbf{S})$ is fixed by the data, so making the kept part as large as possible (keeping the most spread) is the *same* as making the error as small as possible (losing the least). This is the "minimise b" view of Part 2, and the idea behind compressing and reconstructing data, which reappears in autoencoders.

Check with our patients

Keeping PC1 for patient 1: $\hat{\mathbf{x}}_1 = (7.08, 4.83) + 5.53\,(0.970, 0.245) = (12.45, 6.19)$, against the original $(13, 4)$. The lost part is $z_{12}\mathbf{a}_2 = -2.26\,(-0.245, 0.970) = (0.55, -2.19)$, with squared length $2.26^2 = 5.1$. Over all six patients the squared errors add up to $26.0 = 5 \times 5.2 = (n-1)\lambda_2$.

14 Step 12: the algorithm, and the SVD

Where we are and where we are going

Ultimate aim: reached. **This step:** turn the chain into a procedure, and see how software actually computes it.

1. $\mathbf{X} \leftarrow$ the $n \times p$ data matrix, one row $\mathbf{x}_i^\top$ per sample.
2. Centre: subtract $\bar{\mathbf{x}}^\top$ from every row of $\mathbf{X}$.
3. $\mathbf{S} \leftarrow \frac{1}{n-1}\mathbf{X}^\top\mathbf{X}$, the $p \times p$ covariance matrix. (After centring, the rows of $\mathbf{X}$ are the $\mathbf{d}_i^\top$, and $\mathbf{X}^\top\mathbf{X} = \sum_i \mathbf{d}_i\mathbf{d}_i^\top$: entry (j,k) of $\mathbf{X}^\top\mathbf{X}$ is $\sum_i d_{ij}d_{ik}$, exactly the sum of Section 1.3.)
4. Solve $\mathbf{S}\mathbf{a}_j = \lambda_j\mathbf{a}_j$, with $\lambda_1 \ge \cdots \ge \lambda_p$.
5. $\mathbf{A}_k \leftarrow [\mathbf{a}_1 \cdots \mathbf{a}_k]$ and $\mathbf{Z} = \mathbf{X}\mathbf{A}_k$, the $n \times k$ scores.
6. Choose k from $\sum_{j\le k}\lambda_j / \sum_j \lambda_j$ (the scree plot).

The SVD. Libraries usually do not form $\mathbf{S}$. They compute the singular value decomposition of the centred data matrix itself, $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, which is more numerically stable. Here $\mathbf{U}$ $(n \times n)$ and $\mathbf{V}$ $(p \times p)$ have orthonormal columns, so $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$ (exactly as $\mathbf{A}^\top\mathbf{A} = \mathbf{I}$ in Step 9),

and $\boldsymbol{\Sigma}$ ($n \times p$) has the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ on its diagonal and zeros elsewhere. Substituting into $\mathbf{S}$, using the transpose rule and $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$,

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}^\top\mathbf{X} = \frac{1}{n-1}\mathbf{V}\boldsymbol{\Sigma}^\top\mathbf{U}^\top\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top = \mathbf{V}\,\frac{\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}}{n-1}\,\mathbf{V}^\top,$$

where $\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. Comparing with $\mathbf{S} = \mathbf{A}\Lambda\mathbf{A}^\top$: the right singular vectors are the principal directions, $\mathbf{A} = \mathbf{V}$, the eigenvalues are the squared singular values divided by $n-1$, $\lambda_j = \sigma_j^2/(n-1)$, and the scores are $\mathbf{Z} = \mathbf{X}\mathbf{V} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top\mathbf{V} = \mathbf{U}\boldsymbol{\Sigma}$.

Check with our patients

$\sigma_1 = \sqrt{\text{SS(distances)}} = \sqrt{5 \times 29.7} = 12.2$ and $\sigma_2 = \sqrt{5 \times 5.2} = 5.1$; conversely $\lambda_1 = 12.2^2/5 = 29.7$.

A The same result with a Lagrange multiplier

Most textbooks reach $\mathbf{S}\mathbf{a} = \lambda\mathbf{a}$ in three lines with a Lagrange multiplier, a general tool for maximising a function under a constraint (Bishop, 2006, Section 12.1.1 and Appendix E). Form

$$g(\mathbf{a}) = \mathbf{a}^\top\mathbf{S}\mathbf{a} - \lambda\,(\mathbf{a}^\top\mathbf{a} - 1),$$

the objective plus λ times the constraint. We need two gradients (vectors of partial derivatives). Writing $\mathbf{a}^\top\mathbf{S}\mathbf{a} = \sum_j \sum_l a_j s_{jl} a_l$ and differentiating with respect to one entry a_m, the terms with $j = m$ give $\sum_l s_{ml} a_l$ and those with $l = m$ give $\sum_j a_j s_{jm}$; by symmetry both are entry m of $\mathbf{S}\mathbf{a}$, so $\partial(\mathbf{a}^\top\mathbf{S}\mathbf{a})/\partial\mathbf{a} = 2\mathbf{S}\mathbf{a}$. In the same way ($\mathbf{S} = \mathbf{I}$), $\partial(\mathbf{a}^\top\mathbf{a})/\partial\mathbf{a} = 2\mathbf{a}$. Setting the gradient to zero:

$$\frac{\partial g}{\partial \mathbf{a}} = 2\mathbf{S}\mathbf{a} - 2\lambda\mathbf{a} = \mathbf{0} \quad \Longrightarrow \quad \mathbf{S}\mathbf{a} = \lambda\mathbf{a}.$$

Why it is the same argument. The geometric content of the Lagrange condition is this: at a constrained maximum, the gradient of the objective, here $2\mathbf{S}\mathbf{a}$, must be perpendicular to the constraint surface (the circle or sphere of unit vectors); otherwise we could move along the surface and increase the objective. Moving along the surface is turning $\mathbf{a}$, so "perpendicular to the surface" means "no part along any direction of turning", which is exactly our condition $\mathbf{a}'^\top(\mathbf{S}\mathbf{a}) = 0$. The multiplier λ is the factor between the two parallel vectors, and, as we saw, it equals the variance.

B All the equations on one page

Equation	What it says, and why it is there
$z_{i1} = \mathbf{a}_1^\top(\mathbf{x}_i - \bar{\mathbf{x}})$, $\|\mathbf{a}_1\| = 1$	score: one number per patient along a direction
$\bar{z}_1 = 0$	centred data, so spread is measured around 0
$\mathrm{var}(z_1) = \mathbf{a}_1^\top \mathbf{S}\mathbf{a}_1$	spread along any direction as one formula
$(c\mathbf{a})^\top \mathbf{S}(c\mathbf{a}) = c^2 \mathbf{a}^\top \mathbf{S}\mathbf{a}$	the length must be fixed: maximise subject to $\mathbf{a}^\top\mathbf{a} = 1$
$\mathbf{a} = (\cos\theta, \sin\theta)^\top$	only the angle is free; the variance curve (Fig. 1)
$d\,\mathrm{var}\,/d\theta = 2\mathbf{a}'^\top(\mathbf{S}\mathbf{a})$	slope of the variance curve; zero at the top
$\mathbf{S}\mathbf{a} = (\mathbf{a}^\top\mathbf{S}\mathbf{a})\mathbf{a} + (\mathbf{a}'^\top\mathbf{S}\mathbf{a})\mathbf{a}'$	split along $\mathbf{a}$ (variance) and $\mathbf{a}'$ (half the slope)
$\mathbf{S}\mathbf{a} = \lambda\mathbf{a}$	flat: eigenvector; holds in any dimension
$\mathbf{a}^\top\mathbf{S}\mathbf{a} = \lambda$	the eigenvalue is the variance; PC1 has the largest
$\det(\mathbf{S} - \lambda\mathbf{I}) = 0$	how to find the eigenvalues
$(\mathbf{S} - \lambda_1\mathbf{I})\mathbf{a}_1 = \mathbf{0}$, $\|\mathbf{a}_1\| = 1$	how to find the eigenvector
$\mathrm{cov}(z_1, z_2) = \lambda_1 \mathbf{a}_1^\top \mathbf{a}_2$	uncorrelated means perpendicular
$\mathbf{z} = \mathbf{A}^\top(\mathbf{x} - \bar{\mathbf{x}})$, $\mathbf{A}^\top\mathbf{A} = \mathbf{I}$	all scores at once; $\mathbf{A}$ is a rotation
$\mathbf{A}^\top\mathbf{S}\mathbf{A} = \Lambda$, $\mathbf{S} = \mathbf{A}\Lambda\mathbf{A}^\top$	new variables uncorrelated; eigendecomposition
$\mathrm{tr}(\mathbf{S}) = \sum_{j\le k}\lambda_j + \sum_{j>k}\lambda_j$	total spread = kept + discarded
$\sum_{j\le k}\lambda_j / \sum_j \lambda_j$	variation retained
$\sum_i \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2 = (n-1)\sum_{j>k}\lambda_j$	reconstruction error = discarded
$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$: $\mathbf{A} = \mathbf{V}$, $\lambda_j = \sigma_j^2/(n-1)$	how software computes PCA

Further reading

- C. M. Bishop (2006), *Pattern Recognition and Machine Learning*, Springer. Section 12.1.1, Maximum variance formulation, and Section 12.1.2, Minimum-error formulation (from p. 561): Steps 2–6 and Step 11 of these notes. Appendix E, Lagrange multipliers (p. 707): Appendix A.
- M. P. Deisenroth, A. A. Faisal and C. S. Ong (2020), *Mathematics for Machine Learning*, Cambridge University Press, free at https://mml-book.com. Section 4.2, Eigenvalues and eigenvectors; Chapter 10, Dimensionality reduction with PCA.
- G. James, D. Witten, T. Hastie, R. Tibshirani and J. Taylor (2023), *An Introduction to Statistical Learning with Applications in Python*, Springer. Section 12.2.
- T. Hastie, R. Tibshirani and J. Friedman (2009), *The Elements of Statistical Learning*, 2nd edition, Springer. Section 14.5.1 (principal components through the SVD).
- L. Wiskott, *Lecture notes on principal component analysis*: an alternative argument for "why the largest eigenvalue" and for the reconstruction error.
- I. T. Jolliffe and J. Cadima (2016), Principal component analysis: a review and recent developments, *Philosophical Transactions of the Royal Society A* 374: 20150202.