

24.10 Principal Components

Chapter 13 §11 proved that the k -dimensional subspace **through the origin** closest to a cloud of points is spanned by the leading left singular vectors, and then said plainly what it had not proved: the centering, the covariance matrix, and the reading of $\sigma_i^2 / \sum_l \sigma_l^2$ as a fraction of variance. This section supplies all three — and the variance fraction turns out to be a theorem, not a convention — and then marks exactly where linear algebra stops and statistics begins. The honest answer is: more is provable than one might expect, and less than the phrase “explained variance” usually suggests.

Throughout, the field is $\mathbb{R}$. The data are $N \geq 1$ points $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^n$, written as the columns of a matrix, and $\|\cdot\|$ is the Euclidean norm.

24.10.1 Centering

An **affine subspace** of $\mathbb{R}^n$ is a coset $A = \mathbf{p} + U = \{\mathbf{p} + \mathbf{z} : \mathbf{z} \in U\}$ of a subspace U , and its dimension is $\dim U$; Chapter 18 §01 calls U the direction space and shows in Proposition 18.1.4 that every affine hull has this form. The subspaces of $\mathbb{R}^n$ are the affine subspaces containing $\mathbf{0}$.

Chapter 13 §11 fitted subspaces, which all pass through $\mathbf{0}$. That is the wrong family for data: a cloud of points far from the origin is badly described by any line through the origin, and the quantity minimized is then dominated by the distance from $\mathbf{0}$ to the cloud rather than by its shape. Widening the competition to all affine subspaces fixes this, and it costs nothing: the extra freedom is used up in one place, and we can say in advance where.

Subtract the mean once, and the affine problem becomes the linear problem Chapter 13 solved.

Definition 24.10.1. Data matrix, mean, and centered data matrix

Let $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^n$. The **data matrix** is the matrix $\mathbf{X} \in M_{n \times N}(\mathbb{R})$ whose j -th **column** is $\mathbf{x}_j$. The **mean** of the data is

$$\bar{\mathbf{x}} := \frac{1}{N} \sum_{j=1}^N \mathbf{x}_j \in \mathbb{R}^n,$$

and the **centered data matrix** is

$$\mathbf{X}_c := \mathbf{X} - \bar{\mathbf{x}} \mathbf{1}^\top \in M_{n \times N}(\mathbb{R}),$$

whose j -th column is $\mathbf{c}_j := \mathbf{x}_j - \bar{\mathbf{x}}$.

Two clauses deserve comment. **Columns** are the data points, so n is the number of measured quantities and N the number of records; the opposite convention is also in use, and every formula below would be transposed under it. And $\bar{\mathbf{x}} \mathbf{1}^\top$ is the $n \times N$ matrix all of whose columns equal $\bar{\mathbf{x}}$, since $\mathbf{1} \in \mathbb{R}^N$ is the all-ones vector.

The defining property of the centered matrix is that its own mean is zero:

$$\mathbf{X}_c \mathbf{1} = \mathbf{X} \mathbf{1} - \bar{\mathbf{x}} (\mathbf{1}^\top \mathbf{1}) = N \bar{\mathbf{x}} - N \bar{\mathbf{x}} = \mathbf{0},$$

using $\mathbf{1}^\top \mathbf{1} = N$. In words, each **row** of $\mathbf{X}_c$ sums to 0.

Example. Centering three data sets

- (a) $N = 4$ points in $\mathbb{R}^2$: $(5, 4), (4, 1), (2, 3), (1, 0)$. Their sum is $(12, 8)$, so $\bar{\mathbf{x}} = (3, 2)$ and

$$\mathbf{X}_c = \begin{pmatrix} 2 & 1 & -1 & -2 \\ 2 & -1 & 1 & -2 \end{pmatrix}.$$

Both rows sum to 0.

- (b) $N = 1$. Then $\bar{\mathbf{x}} = \mathbf{x}_1$ and $\mathbf{X}_c = \mathbf{O}$: a single point carries no shape at all. This degenerate case matters, because every ratio below has $0/0$ here and the statements must exclude it or say what they mean.
- (c) Data already centered: if $\mathbf{X}\mathbf{1} = \mathbf{0}$ then $\bar{\mathbf{x}} = \mathbf{0}$ and $\mathbf{X}_c = \mathbf{X}$. Centering is idempotent, since $\mathbf{X}_c\mathbf{1} = \mathbf{0}$ always.

Change one clause and the object changes. Suppose one subtracts the mean of each *column* instead of forming the mean of the columns, replacing $\mathbf{x}_j$ by $\mathbf{x}_j - (\frac{1}{n}\mathbf{1}_n^\top \mathbf{x}_j)\mathbf{1}_n$ with $\mathbf{1}_n \in \mathbb{R}^n$. That removes the average of the measured quantities inside each record, which is a meaningful operation for some data and a meaningless one for most, and in any case is not the one below. The clause that matters is *which index is averaged over*.

24.10.2 The best-fitting affine subspace

Distances to an affine subspace reduce to distances to its direction space, and the reduction is Chapter 11's Best Approximation Theorem.

Lemma 24.10.1. Distance to a flat, and where its translate wants to be

Let $U \subseteq \mathbb{R}^n$ be a subspace, let $Q = \mathbf{I}_n - P_U$ be the orthogonal projection onto $U^\perp$, and let $\mathbf{p} \in \mathbb{R}^n$.

- (a) For every $\mathbf{x} \in \mathbb{R}^n$, $d(\mathbf{x}, \mathbf{p} + U) = \|Q(\mathbf{x} - \mathbf{p})\|$, and the nearest point of $\mathbf{p} + U$ exists.
- (b) For points $\mathbf{x}_1, \dots, \mathbf{x}_N$ with mean $\bar{\mathbf{x}}$,

$$\sum_{j=1}^N \|Q(\mathbf{x}_j - \mathbf{p})\|^2 = \sum_{j=1}^N \|Q(\mathbf{x}_j - \bar{\mathbf{x}})\|^2 + N\|Q(\bar{\mathbf{x}} - \mathbf{p})\|^2.$$

In particular the left side is smallest exactly when $\bar{\mathbf{x}} \in \mathbf{p} + U$.

Idea. Part (a) is the Best Approximation Theorem after sliding the whole picture by $-\mathbf{p}$. Part (b) is the one-dimensional identity “sum of squares about a point = sum of squares about the mean + N times the squared displacement”, proved the way it always is: expand the square and notice that the cross term is a multiple of $\sum_j (\mathbf{x}_j - \bar{\mathbf{x}})$, which is $\mathbf{0}$ by the definition of the mean.

Proof.

- (a) A point of $\mathbf{p} + U$ is $\mathbf{p} + \mathbf{z}$ with $\mathbf{z} \in U$, and $\|\mathbf{x} - (\mathbf{p} + \mathbf{z})\| = \|(\mathbf{x} - \mathbf{p}) - \mathbf{z}\|$. By Theorem 11.3.5 the quantity $\|(\mathbf{x} - \mathbf{p}) - \mathbf{z}\|$ over $\mathbf{z} \in U$ is smallest exactly at $\mathbf{z} = P_U(\mathbf{x} - \mathbf{p})$, where its

value is $\|(\mathbf{x} - \mathbf{p}) - P_U(\mathbf{x} - \mathbf{p})\| = \|Q(\mathbf{x} - \mathbf{p})\|$. So the minimum is attained and equals $\|Q(\mathbf{x} - \mathbf{p})\|$, which is $d(\mathbf{x}, \mathbf{p} + U)$ by Definition 11.3.3 applied after the translation.

- (b) Put $\mathbf{y}_j = Q(\mathbf{x}_j - \bar{\mathbf{x}})$ and $\mathbf{d} = Q(\bar{\mathbf{x}} - \mathbf{p})$, so that $Q(\mathbf{x}_j - \mathbf{p}) = \mathbf{y}_j + \mathbf{d}$ by linearity of Q . Since Q is linear and $\sum_j (\mathbf{x}_j - \bar{\mathbf{x}}) = N\bar{\mathbf{x}} - N\bar{\mathbf{x}} = \mathbf{0}$,

$$\sum_j \mathbf{y}_j = Q\left(\sum_j (\mathbf{x}_j - \bar{\mathbf{x}})\right) = \mathbf{0}.$$

Therefore, expanding each square,

$$\begin{aligned} \sum_j \|\mathbf{y}_j + \mathbf{d}\|^2 &= \sum_j (\|\mathbf{y}_j\|^2 + 2\langle \mathbf{y}_j, \mathbf{d} \rangle + \|\mathbf{d}\|^2) \\ &= \sum_j \|\mathbf{y}_j\|^2 + 2\left\langle \sum_j \mathbf{y}_j, \mathbf{d} \right\rangle + N\|\mathbf{d}\|^2 \\ &= \sum_j \|\mathbf{y}_j\|^2 + N\|\mathbf{d}\|^2, \end{aligned}$$

which is the identity. The second term is ≥ 0 and vanishes exactly when $\mathbf{d} = \mathbf{0}$, that is when $\bar{\mathbf{x}} - \mathbf{p} \in \ker Q = U$ (Theorem 11.3.2), that is when $\bar{\mathbf{x}} \in \mathbf{p} + U$. This proves the lemma. ■

Part (b) is the whole of “the best affine subspace passes through the mean”. It holds for **every** direction space U , before any optimization over U has happened, which is why the affine problem costs nothing extra.

Theorem 24.10.2. The Best-Fitting Affine Subspace

Let $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^n$ have mean $\bar{\mathbf{x}}$ and centered data matrix $\mathbf{X}_c$ (Definition 24.10.1), let $\sigma_1 \geq \sigma_2 \geq \dots \geq 0$ be the singular values of $\mathbf{X}_c$ (Definition 13.9.1), and let $\mathbf{u}_1, \mathbf{u}_2, \dots$ be the left singular vectors of a fixed singular value decomposition of $\mathbf{X}_c$ (Theorem 13.9.1). Let $1 \leq k \leq \min(n, N)$. Among all affine subspaces $A \subseteq \mathbb{R}^n$ of dimension at most k , the total squared distance

$$D(A) := \sum_{j=1}^N d(\mathbf{x}_j, A)^2$$

is smallest for

$$A_0 = \bar{\mathbf{x}} + \text{span}(\mathbf{u}_1, \dots, \mathbf{u}_k),$$

and its smallest value is $\sum_{i>k} \sigma_i^2$. Moreover **every** minimizing A contains $\bar{\mathbf{x}}$.

Idea. Two optimizations in sequence, and Lemma 24.10.1 does the outer one first. ① For a *fixed* direction space U , the best translate is the one through $\bar{\mathbf{x}}$, whatever U is. ② With the translate fixed at $\bar{\mathbf{x}}$, the objective becomes the total squared distance from the **centered** points to the subspace U — which is exactly the quantity Chapter 13 §11 minimized. So nothing new has to be proved about subspaces; the work is entirely in noticing that step ① does not interact with step ②.

Proof. Write $A = \mathbf{p} + U$ with $\dim U \leq k$, and let $Q = \mathbf{I}_n - P_U$. By Lemma 24.10.1 (a),

$$D(\mathbf{p} + U) = \sum_j \|Q(\mathbf{x}_j - \mathbf{p})\|^2,$$

and by part (b) of the same lemma this is at least $\sum_j \|Q(\mathbf{x}_j - \bar{\mathbf{x}})\|^2 = D(\bar{\mathbf{x}} + U)$, with equality exactly when $\bar{\mathbf{x}} \in \mathbf{p} + U$. So it suffices to minimize $D(\bar{\mathbf{x}} + U)$ over subspaces U with $\dim U \leq k$.

Now $\|Q(\mathbf{x}_j - \bar{\mathbf{x}})\| = \|\mathbf{c}_j - P_U \mathbf{c}_j\| = d(\mathbf{c}_j, U)$ by Definition 11.3.3, where $\mathbf{c}_j = \mathbf{x}_j - \bar{\mathbf{x}}$ is the j -th column of $\mathbf{X}_c$. Hence

$$D(\bar{\mathbf{x}} + U) = \sum_{j=1}^N d(\mathbf{c}_j, U)^2,$$

which is the quantity $D(U)$ of Corollary 13.11.4 for the matrix $\mathbf{X}_c \in M_{n \times N}(\mathbb{R})$. That corollary, applicable because $1 \leq k \leq \min(n, N)$, says that the minimum over $\dim U \leq k$ is $\sum_{i>k} \sigma_i^2$ and is attained at $U_0 = \text{span}(\mathbf{u}_1, \dots, \mathbf{u}_k)$. Combining, $D(A) \geq \sum_{i>k} \sigma_i^2$ for every admissible A , with equality at $A_0 = \bar{\mathbf{x}} + U_0$.

Finally let $A = \mathbf{p} + U$ be any minimizer. Then $D(\mathbf{p} + U) = D(\bar{\mathbf{x}} + U)$, since the right side is admissible and no smaller; by the equality clause of Lemma 24.10.1 (b) this forces $\bar{\mathbf{x}} \in \mathbf{p} + U = A$. This proves the theorem. ■

Quick check. Is the minimizer unique? Take the four centered points $(1, 0)$, $(-1, 0)$, $(0, 1)$, $(0, -1)$ in $\mathbb{R}^2$ and $k = 1$.

Solution. No. Here $\bar{\mathbf{x}} = \mathbf{0}$ and $\mathbf{X}_c \mathbf{X}_c^\top = \text{diag}(2, 2)$, so $\sigma_1 = \sigma_2 = \sqrt{2}$ and the theorem's minimum is $\sigma_2^2 = 2$. Every line through the origin attains it: for $\text{span}(\mathbf{e}_1)$ the four squared distances are 0, 0, 1, 1; for $\text{span}((1, 1))$ each point is at squared distance $1 - \frac{1}{2} = \frac{1}{2}$. Total 2 in both cases. What the theorem asserts is that A_0 is a minimizer and that $\sum_{i>k} \sigma_i^2$ is the minimum value; uniqueness fails exactly when $\sigma_k = \sigma_{k+1}$, just as the truncation of Definition 13.11.1 is nonunique then.

24.10.3 The sample covariance matrix

Fix a direction $\mathbf{w} \in \mathbb{R}^n$ with $\mathbf{w} \neq \mathbf{0}$ and look at the data through it: each point $\mathbf{x}_j$ has the single coordinate $t_j = \mathbf{w}^\top \mathbf{x}_j$ along $\mathbf{w}$. The mean of these numbers is

$$\frac{1}{N} \sum_j \mathbf{w}^\top \mathbf{x}_j = \mathbf{w}^\top \bar{\mathbf{x}},$$

and their spread about that mean is the number

$$\begin{aligned} \frac{1}{N} \sum_{j=1}^N (\mathbf{w}^\top \mathbf{x}_j - \mathbf{w}^\top \bar{\mathbf{x}})^2 &= \frac{1}{N} \sum_{j=1}^N (\mathbf{w}^\top \mathbf{c}_j)^2 \\ &= \frac{1}{N} \mathbf{w}^\top \mathbf{X}_c \mathbf{X}_c^\top \mathbf{w}, \end{aligned}$$

the last step because $\sum_j (\mathbf{w}^\top \mathbf{c}_j)^2 = \|\mathbf{X}_c^\top \mathbf{w}\|^2 = \mathbf{w}^\top \mathbf{X}_c \mathbf{X}_c^\top \mathbf{w}$. The direction $\mathbf{w}$ enters only through a quadratic form, and the matrix of that form does not depend on $\mathbf{w}$. That is a recurring expression, so it gets a name.

The sample covariance matrix is the single matrix that answers “how spread out are the data in direction $\mathbf{w}$?” for every $\mathbf{w}$ at once.

Definition 24.10.2. Sample covariance matrix

Let $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^n$ with centered data matrix $\mathbf{X}_c$ (Definition 24.10.1). The **sample covariance matrix** of the data is

$$\mathbf{S} := \frac{1}{N} \mathbf{X}_c \mathbf{X}_c^\top \in M_n(\mathbb{R}),$$

with entries

$$s_{ab} = \frac{1}{N} \sum_{j=1}^N (x_{aj} - \bar{x}_a)(x_{bj} - \bar{x}_b) \quad (1 \leq a, b \leq n).$$

For $\mathbf{w} \in \mathbb{R}^n$ with $\mathbf{w} \neq \mathbf{0}$, the **variance of the data in the direction $\mathbf{w}$** is the number $\mathbf{w}^\top \mathbf{S} \mathbf{w} / \|\mathbf{w}\|^2$, which for a **unit $\mathbf{w}$** is $\mathbf{w}^\top \mathbf{S} \mathbf{w}$; and the **total variance** of the data is $\text{tr } \mathbf{S}$.

Clause by clause. The factor $1/N$ makes $\mathbf{S}$ an average, so it does not grow when data are duplicated. The entry s_{aa} is the ordinary variance of the a -th measured quantity, and s_{ab} for $a \neq b$ says whether quantities a and b tend to deviate from their means in the same direction (positive) or in opposite directions (negative). The centering is inside the definition, not optional: without it the same formula would measure spread about $\mathbf{0}$ rather than about the data.

Note also what $\mathbf{S}$ does **not** depend on: the order of the columns, since the defining sum runs over all of them. It does depend on the units of each coordinate, which is the subject of the warning below.

Example. Three covariance matrices

- (a) The four points of Example (Centering three data sets) (a). Then

$$\mathbf{X}_c \mathbf{X}_c^\top = \begin{pmatrix} 10 & 6 \\ 6 & 10 \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} 5/2 & 3/2 \\ 3/2 & 5/2 \end{pmatrix}.$$

Both coordinates have variance $5/2$, and the positive off-diagonal entry says they rise and fall together.

- (b) Two points $\pm \mathbf{v}$ in $\mathbb{R}^n$ with $\mathbf{v} \neq \mathbf{0}$, so $N = 2$ and $\bar{\mathbf{x}} = \mathbf{0}$. Then $\mathbf{X}_c \mathbf{X}_c^\top = 2\mathbf{v}\mathbf{v}^\top$ and $\mathbf{S} = \mathbf{v}\mathbf{v}^\top$, of rank one: all the spread is in one direction, and $\mathbf{w}^\top \mathbf{S} \mathbf{w} = (\mathbf{w}^\top \mathbf{v})^2$ vanishes for every $\mathbf{w} \perp \mathbf{v}$.

- (c) $N = 1$. Then $\mathbf{X}_c = \mathbf{0}$ and $\mathbf{S} = \mathbf{0}$: no direction has any spread.

Change the definition in one place and it stops being a covariance. Replacing $\mathbf{X}_c$ by $\mathbf{X}$ gives $\frac{1}{N} \mathbf{X} \mathbf{X}^\top$, whose (a, a) entry is the mean of x_{aj}^2 rather than the variance; for the data of (a) it is $\begin{pmatrix} 46/4 & 30/4 \\ 30/4 & 26/4 \end{pmatrix}$, which is not $\mathbf{S}$, and its top eigenvector is a multiple of $(1, 0.72\dots)$ rather than of $(1, 1)$. The clause that fails is the subtraction of the mean, and with it every statement below.

Proposition 24.10.3. The Covariance Matrix and the Singular Value Decomposition

Let $\mathbf{X}_c \in M_{n \times N}(\mathbb{R})$ be a centered data matrix with singular values $\sigma_1 \geq \sigma_2 \geq \dots$ (padded with zeros to length n), and let $\mathbf{X}_c = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be a singular value decomposition with columns $\mathbf{u}_1, \dots, \mathbf{u}_n$ of $\mathbf{U}$. Put $\mathbf{S} = \frac{1}{N}\mathbf{X}_c\mathbf{X}_c^\top$. Then:

- (a) $\mathbf{S}$ is symmetric and $\mathbf{S} \succeq \mathbf{O}$;
- (b) $\mathbf{S}\mathbf{u}_i = \frac{\sigma_i^2}{N}\mathbf{u}_i$ for every i , so the eigenvalues of $\mathbf{S}$ are $\sigma_1^2/N \geq \dots \geq \sigma_n^2/N$ and the $\mathbf{u}_i$ are an orthonormal eigenbasis;
- (c) for every $\mathbf{w} \neq \mathbf{0}$, the variance of the data in the direction $\mathbf{w}$ is the Rayleigh quotient $R_{\mathbf{S}}(\mathbf{w})$ of Definition 17.1.1;
- (d) $\text{tr } \mathbf{S} = \frac{1}{N} \sum_{j=1}^N \|\mathbf{x}_j - \bar{\mathbf{x}}\|^2 = \frac{1}{N} \sum_i \sigma_i^2$.

Idea. Every part is one line once $\mathbf{X}_c = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ is written down: (a) and (b) because $\mathbf{X}_c\mathbf{X}_c^\top = \mathbf{U}(\mathbf{\Sigma}\mathbf{\Sigma}^\top)\mathbf{U}^\top$ is already a diagonalization, (c) because the computation before Definition 24.10.2 was the Rayleigh quotient, and (d) because the trace does not notice the unitary factors and reads the same quantity off the rows and off the columns.

Proof.

- (a) $\mathbf{S}^\top = \frac{1}{N}(\mathbf{X}_c\mathbf{X}_c^\top)^\top = \frac{1}{N}\mathbf{X}_c\mathbf{X}_c^\top = \mathbf{S}$, and for every $\mathbf{w}$, $\mathbf{w}^\top\mathbf{S}\mathbf{w} = \frac{1}{N}\|\mathbf{X}_c^\top\mathbf{w}\|^2 \geq 0$; with symmetry this is Definition 13.1.1.
- (b) Since $\mathbf{V}$ is orthogonal, $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_N$, so

$$\mathbf{X}_c\mathbf{X}_c^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{V}\mathbf{\Sigma}^\top\mathbf{U}^\top = \mathbf{U}(\mathbf{\Sigma}\mathbf{\Sigma}^\top)\mathbf{U}^\top,$$

and $\mathbf{\Sigma}\mathbf{\Sigma}^\top = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$ by the shape of $\mathbf{\Sigma}$ in Theorem 13.9.1. Multiplying on the right by $\mathbf{u}_i$ and using $\mathbf{U}^\top\mathbf{u}_i = \mathbf{e}_i$ gives $\mathbf{X}_c\mathbf{X}_c^\top\mathbf{u}_i = \sigma_i^2\mathbf{u}_i$; divide by N . The $\mathbf{u}_i$ are orthonormal and there are n of them, so they are an orthonormal basis of $\mathbb{R}^n$, and the list σ_i^2/N is decreasing.

- (c) By the computation preceding Definition 24.10.2, the variance in direction $\mathbf{w}$ is $\mathbf{w}^\top\mathbf{S}\mathbf{w}/\|\mathbf{w}\|^2$. Since $\mathbf{S}$ is symmetric, $R_{\mathbf{S}}(\mathbf{w}) = \langle \mathbf{S}\mathbf{w}, \mathbf{w} \rangle / \langle \mathbf{w}, \mathbf{w} \rangle = \mathbf{w}^\top\mathbf{S}\mathbf{w}/\|\mathbf{w}\|^2$ by Definition 17.1.1, which is the same number.
- (d) The j -th diagonal entry of $\mathbf{X}_c^\top\mathbf{X}_c$ is $\|\mathbf{c}_j\|^2$, so $\text{tr}(\mathbf{X}_c^\top\mathbf{X}_c) = \sum_j \|\mathbf{c}_j\|^2$; and $\text{tr}(\mathbf{X}_c\mathbf{X}_c^\top) = \text{tr}(\mathbf{X}_c^\top\mathbf{X}_c)$ by Theorem 0.9.6 (3). On the other hand the display in (b) and the same trace identity give

$$\begin{aligned} \text{tr}(\mathbf{X}_c\mathbf{X}_c^\top) &= \text{tr}(\mathbf{U}(\mathbf{\Sigma}\mathbf{\Sigma}^\top)\mathbf{U}^\top) \\ &= \text{tr}((\mathbf{\Sigma}\mathbf{\Sigma}^\top)\mathbf{U}^\top\mathbf{U}) = \sum_i \sigma_i^2, \end{aligned}$$

using $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_n$. Dividing by N gives both equalities. This proves the proposition. ■

Part (c) connects this section to Chapter 17. The question “which direction has the most spread?” is the question “where is $R_{\mathbf{S}}$ largest?”, and Proposition 17.1.1 answers it: the maximum

is $\lambda_1(\mathbf{S}) = \sigma_1^2/N$, attained at $\mathbf{u}_1$. “Which direction has the most spread among those orthogonal to $\mathbf{u}_1, \dots, \mathbf{u}_{k-1}$?” is answered the same way by Proposition 17.1.3, with maximum σ_k^2/N at $\mathbf{u}_k$. So the leading left singular vectors are not only the best-fitting directions of Theorem 24.10.2; they are also, one at a time, the directions of greatest remaining spread. The two descriptions agree because they are the two halves of one Pythagorean identity, which is the next theorem.

Definition 24.10.3. Principal directions and components

In the situation of Proposition 24.10.3, the unit vectors $\mathbf{u}_1, \mathbf{u}_2, \dots$ are the **principal directions** of the data, and the numbers $\mathbf{u}_i^\top \mathbf{c}_1, \dots, \mathbf{u}_i^\top \mathbf{c}_N$ are the i -th **principal components**: the coordinates of the centered data along $\mathbf{u}_i$.

24.10.4 How much variance a subspace explains

Theorem 24.10.4. Variance Captured by the Principal Directions

In the situation of Proposition 24.10.3, let $1 \leq k \leq \min(n, N)$, let $U_0 = \text{span}(\mathbf{u}_1, \dots, \mathbf{u}_k)$, and let P be the orthogonal projection onto a subspace U with $\dim U \leq k$. Write

$$\text{var}(U) := \frac{1}{N} \sum_{j=1}^N \|P(\mathbf{x}_j - \bar{\mathbf{x}})\|^2.$$

Then:

- (a) $\text{var}(U)$ is the total variance of the projected points $P\mathbf{x}_1, \dots, P\mathbf{x}_N$, whose mean is $P\bar{\mathbf{x}}$;
- (b) $\text{var}(U) + \frac{1}{N} D(\bar{\mathbf{x}} + U) = \text{tr } \mathbf{S}$, where D is as in Theorem 24.10.2;
- (c) $\text{var}(U) \leq \text{var}(U_0) = \frac{1}{N} \sum_{i \leq k} \sigma_i^2$;
- (d) if $\mathbf{X}_c \neq \mathbf{O}$, the fraction of the total variance captured by U_0 is

$$\frac{\text{var}(U_0)}{\text{tr } \mathbf{S}} = \frac{\sigma_1^2 + \dots + \sigma_k^2}{\sigma_1^2 + \sigma_2^2 + \dots}.$$

Idea. Everything is Pythagoras applied N times and then added up. For each centered point, $\|\mathbf{c}_j\|^2$ splits into the part P keeps and the part it throws away; averaging the first over j is $\text{var}(U)$, averaging the second is the fitting error of Theorem 24.10.2, and averaging the total is $\text{tr } \mathbf{S}$ by Proposition 24.10.3 (d). Once (b) is available, (c) and (d) are just the theorem already proved, divided by N .

Proof.

- (a) The mean of $P\mathbf{x}_1, \dots, P\mathbf{x}_N$ is $\frac{1}{N} \sum_j P\mathbf{x}_j = P\bar{\mathbf{x}}$ by linearity of P , so the j -th centered projected point is $P\mathbf{x}_j - P\bar{\mathbf{x}} = P\mathbf{c}_j$. By Proposition 24.10.3 (d) applied to these N points, their total variance is $\frac{1}{N} \sum_j \|P\mathbf{c}_j\|^2 = \text{var}(U)$.
- (b) Fix j . The vectors $P\mathbf{c}_j$ and $\mathbf{c}_j - P\mathbf{c}_j$ are orthogonal, the first lying in U and the second

in $U^\perp$ (Theorem 11.3.2), so Theorem 11.1.3 gives

$$\|\mathbf{c}_j\|^2 = \|P\mathbf{c}_j\|^2 + \|\mathbf{c}_j - P\mathbf{c}_j\|^2.$$

Summing over j and dividing by N : the left side is $\text{tr } \mathbf{S}$ by Proposition 24.10.3 (d), the first term on the right is $\text{var}(U)$, and the second is $\frac{1}{N} \sum_j d(\mathbf{c}_j, U)^2 = \frac{1}{N} D(\bar{\mathbf{x}} + U)$, as computed in the proof of Theorem 24.10.2.

- (c) By (b), maximizing $\text{var}(U)$ over $\dim U \leq k$ is the same as minimizing $D(\bar{\mathbf{x}} + U)$, and Theorem 24.10.2 says the minimum is $\sum_{i>k} \sigma_i^2$, attained at U_0 . Hence $\text{var}(U) \leq \text{var}(U_0)$, and by (b) again,

$$\begin{aligned} \text{var}(U_0) &= \text{tr } \mathbf{S} - \frac{1}{N} \sum_{i>k} \sigma_i^2 \\ &= \frac{1}{N} \sum_i \sigma_i^2 - \frac{1}{N} \sum_{i>k} \sigma_i^2 = \frac{1}{N} \sum_{i \leq k} \sigma_i^2. \end{aligned}$$

- (d) If $\mathbf{X}_c \neq \mathbf{O}$ then some $\sigma_i > 0$, so $\text{tr } \mathbf{S} = \frac{1}{N} \sum_i \sigma_i^2 > 0$ and the quotient is defined; dividing the value in (c) by it cancels the $1/N$. This proves the theorem. ■

So “the first k principal directions explain $(\sum_{i \leq k} \sigma_i^2) / (\sum_i \sigma_i^2)$ of the variance” is a theorem about the N given points, not a convention: the numerator is the variance of the projected cloud, the denominator is the variance of the original cloud, and part (b) is what makes the two add up with nothing left over.

Example. Four points, done exactly

Take the four points $(5, 4), (4, 1), (2, 3), (1, 0)$ of Example (Centering three data sets) (a). Find the mean, the sample covariance matrix, the principal directions, the best-fitting line, the total squared distance to it, and the fraction of variance it explains.

Solution. Mean and centering. $\bar{\mathbf{x}} = (3, 2)$, and $\mathbf{X}_c$ has columns $(2, 2), (1, -1), (-1, 1), (-2, -2)$, as computed above.

Covariance. $\mathbf{X}_c \mathbf{X}_c^\top = \begin{pmatrix} 10 & 6 \\ 6 & 10 \end{pmatrix}$, so $\mathbf{S} = \begin{pmatrix} 5/2 & 3/2 \\ 3/2 & 5/2 \end{pmatrix}$. Since $\mathbf{X}_c \mathbf{X}_c^\top$ has the form $a\mathbf{I} + b\mathbf{J}$ it sends $(1, 1)$ to $16(1, 1)$ and $(1, -1)$ to $4(1, -1)$, so its eigenvalues are 16 and 4 with

$$\mathbf{u}_1 = \frac{1}{\sqrt{2}}(1, 1), \quad \mathbf{u}_2 = \frac{1}{\sqrt{2}}(1, -1).$$

By Proposition 24.10.3 (b), $\sigma_1^2 = 16$ and $\sigma_2^2 = 4$, that is $\sigma_1 = 4$ and $\sigma_2 = 2$, and the eigenvalues of $\mathbf{S}$ are 4 and 1. Check: $\text{tr } \mathbf{S} = 5$ and $4 + 1 = 5$.

Total variance, two ways. The squared distances of the centered points from $\mathbf{0}$ are 8, 2, 2, 8, of average 5. And $\frac{1}{N} \sum_i \sigma_i^2 = \frac{1}{4}(16 + 4) = 5$, as Proposition 24.10.3 (d) requires.

Best-fitting line. By Theorem 24.10.2 with $k = 1$ it is

$$A_0 = (3, 2) + \text{span}((1, 1)),$$

the line $y = x - 1$, and the total squared distance is $\sigma_2^2 = 4$. Directly: $(5, 4)$ and $(1, 0)$ lie

on the line, contributing 0; for $(4, 1)$ the centered point is $(1, -1)$, whose projection onto $\text{span}((1, 1))$ is $\mathbf{0}$, so its squared distance is 2; likewise 2 for $(2, 3)$. Total 4, as predicted.

Variance explained. $\text{var}(U_0) = \frac{1}{4}\sigma_1^2 = 4$, out of $\text{tr } \mathbf{S} = 5$: the first principal direction explains $16/20 = 80\%$ of the variance, and the remaining 20% is the average squared distance $4/4 = 1$ to the line. The two add to 5, which is [Theorem 24.10.4](#) (b).

Warning. Principal directions are not “the important directions”: they change when you change the units. Take the four centered points $(1, 0), (-1, 0), (0, 2), (0, -2)$. Then $\mathbf{X}_c \mathbf{X}_c^\top = \text{diag}(2, 8)$, so the first principal direction is $\mathbf{e}_2$. Now suppose the first coordinate was measured in meters and we re-express it in units three times smaller, which multiplies row 1 by 3. The data are the same data; the new centered matrix is $\mathbf{D}\mathbf{X}_c$ with $\mathbf{D} = \text{diag}(3, 1)$, and

$$(\mathbf{D}\mathbf{X}_c)(\mathbf{D}\mathbf{X}_c)^\top = \mathbf{D} \text{diag}(2, 8) \mathbf{D} = \text{diag}(18, 8),$$

so the first principal direction is now $\mathbf{e}_1$. It has turned by a right angle because of a choice of units. Nothing is wrong with the theorem — $\mathbf{D}\mathbf{X}_c$ really is a different matrix — but “the direction of greatest variance” is a statement about a coordinate system, not about the data alone. Rescaling each coordinate to variance 1 before starting is one common repair; it is a modeling decision, and it is not forced by anything proved here.

24.10.5 Computing it, and what is not proved

Two closing matters, one computational and one about scope.

Never form $\mathbf{X}_c \mathbf{X}_c^\top$. Everything above is stated in terms of $\mathbf{S}$, and $\mathbf{S}$ is the natural object to state it with; but forming $\mathbf{S}$ and then finding its eigenvalues is the worst available way to compute the principal directions. The reason is Theorem 16.8.4: passing from $\mathbf{A}$ to $\mathbf{A}^* \mathbf{A}$ squares the condition number, $\kappa_2(\mathbf{A}^* \mathbf{A}) = \kappa_2(\mathbf{A})^2$, and Section 3 of this chapter turns that into a count of lost digits — Theorem 24.3.2 says the product route loses about $2 \log_{10} \kappa_2$ decimal digits where a backward stable factorization of $\mathbf{A}$ itself loses about $\log_{10} \kappa_2$. The same arithmetic applies with $\mathbf{A} = \mathbf{X}_c^\top$ — but only when that matrix has full column rank, which is the hypothesis of Theorem 16.8.4. That asks for $\text{rank } \mathbf{X}_c = n$, and $\mathbf{X}_c \mathbf{1} = \mathbf{0}$ already forces $\text{rank } \mathbf{X}_c \leq N - 1$, so it fails whenever $N \leq n$: for three records of ten measured quantities there is nothing to apply. The count that follows needs no hypothesis on the rank, and it is the one to remember.

Here is the count in this setting. Each entry of $\mathbf{X}_c \mathbf{X}_c^\top$ is an inner product of two rows, each of length N , so by Theorem 24.1.2 the computed product is $\mathbf{X}_c \mathbf{X}_c^\top + \mathbf{\Delta}$ with $\|\mathbf{\Delta}\|_2$ of order $c(n, N) u \sigma_1^2$, where u is the unit roundoff of Section 1 and c is a modest dimensional factor. It is a factor and not a constant because the entrywise bound has to be aggregated over the whole $n \times n$ product, by the same count as in Section 3, where the corresponding step reads $\|\mathbf{C} - \mathbf{A}^* \mathbf{A}\| \leq \gamma_m \|\mathbf{A}\|_F^2 \leq n \gamma_m \|\mathbf{A}\|^2$: the length of the inner products sits inside γ , and passing from the Frobenius norm to the spectral norm puts their number in front. Here that makes $c(n, N)$ of size about nN .

If the implementation forms only the upper triangle and reflects it — an assumption about the computation, not something proved here — then $\mathbf{\Delta}$ is symmetric, and Corollary 17.3.6 says a symmetric perturbation of that size moves every eigenvalue by at most $\|\mathbf{\Delta}\|_2$. So every σ_i^2 below $c(n, N) u \sigma_1^2$ — that is, up to the factor $\sqrt{c}$, every σ_i below $\sqrt{u} \sigma_1$ — is indistinguishable from 0

in the computed product, while a backward stable factorization of $\mathbf{X}_c$ itself keeps it. Those are exactly the singular values a low-rank fit has to decide about. The right computation therefore works on $\mathbf{X}_c$ directly and never builds the product.

What this chapter does **not** do is assemble that computation. The ingredients are here — the Householder reductions of Section 3, the shifted **QR** iteration of Section 5 — but no algorithm for the singular values of a given matrix is written down or analyzed anywhere in this book. So the paragraph above pays the parenthesis of the last clause of Chapter 13 §11’s promise, that a singular value decomposition is “never” computed by forming $\mathbf{A}^* \mathbf{A}$ and finding its eigenvalues; the clause itself, how one is actually computed for a large matrix, is left unpaid.

The finite population, and the boundary. Chapter 13 §11 said that the statistics needs “a probabilistic model, a reason to prefer squared error, and an argument that the sample answer says something about the population”. Here is exactly what has been delivered.

A probabilistic model, for the finite population of the data themselves. Regard the N columns as a population of N equally likely outcomes, each of weight $1/N$. Then “expectation” means the average $\frac{1}{N} \sum_j$, which needs no measure theory and no limit, and the mean of the population is $\bar{\mathbf{x}}$. For two real-valued quantities f, g on it, with averages $\bar{f} = \frac{1}{N} \sum_j f(\mathbf{x}_j)$ and $\bar{g} = \frac{1}{N} \sum_j g(\mathbf{x}_j)$, the **covariance** of the population is

$$\text{Cov}(f, g) := \frac{1}{N} \sum_{j=1}^N (f(\mathbf{x}_j) - \bar{f})(g(\mathbf{x}_j) - \bar{g}),$$

and the variance of f is $\text{Cov}(f, f)$. With that reading, the variance of $\mathbf{x} \mapsto \mathbf{w}^\top \mathbf{x}$ is $\mathbf{w}^\top \mathbf{S} \mathbf{w}$, which for a **unit** $\mathbf{w}$ is the variance in the direction $\mathbf{w}$ of Definition 24.10.2, hence the Rayleigh quotient of Proposition 24.10.3 (c); and the covariance of $\mathbf{x} \mapsto x_a$ with $\mathbf{x} \mapsto x_b$ is s_{ab} , which is the displayed formula of Definition 24.10.2 read backwards. So $\mathbf{S}$ is the covariance matrix of that population, and Theorem 24.10.4 is a statement about it. This much is a theorem, and it is what was proved above.

A reason to prefer squared error. Three, none of them a proof that squared error is the right criterion, which is a modeling question. (i) It does not depend on the coordinate frame: an orthogonal change of coordinates is an isometry, so it leaves every distance $d(\mathbf{x}_j, A)$, hence $D(A)$, unchanged. The largest entry of the error has no such property — the matrix $\text{diag}(1, 0)$ has largest entry 1, while the same matrix read in the frame $\frac{1}{\sqrt{2}}(1, \pm 1)$ is $\frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, with largest entry $\frac{1}{2}$ — and Chapter 13 §11’s warning shows that it really does prefer a different answer. (ii) It splits: Theorem 24.10.4 (b) says captured plus lost equals total, exactly, which is what makes a *fraction* of variance meaningful at all. (iii) It has a closed-form minimizer, computable by a single factorization (Theorem 13.11.3); for the sum of **unsquared** distances $\sum_j d(\mathbf{x}_j, A)$, nothing proved in this book produces a minimizer at all.

An argument that the sample says something about a larger population. This is **not** proved here, and it cannot be with the tools of this book. Such an argument requires a probability model in which the N points are drawn from a distribution on $\mathbb{R}^n$, an expectation defined by an integral rather than a finite sum, and theorems saying that $\mathbf{S}$ converges to the distribution’s covariance and that its eigenvectors converge to the distribution’s principal directions. Every one of those belongs to probability and statistics, which this book does not develop. Chapter 13 §11’s sentence promised it; the honest correction is that linear algebra supplies the finite-population model and the reasons for squared error, and hands the inference to a subject this book does not enter.

Remark. The divisor N versus $N - 1$. Statistics usually defines the sample covariance with $1/(N - 1)$, which presupposes $N \geq 2$, because that makes it an unbiased estimator of a population covariance — a statement inside the probability model just excluded, and one we therefore cannot prove. With the finite-population reading the divisor is N , and that is the convention of [Definition 24.10.2](#), which makes sense for every $N \geq 1$. For $N \geq 2$ the choice changes nothing here: it multiplies $\mathbf{S}$ by $N/(N - 1)$, which scales every eigenvalue by the same factor, changes no eigenvector, and leaves every ratio such as [Theorem 24.10.4](#) (d) exactly as it was.

24.10.6 Exercises

24.10.6.1 A. Check your understanding

Exercise. A1

- Define the centered data matrix and the sample covariance matrix, and say what $\mathbf{w}^\top \mathbf{S} \mathbf{w}$ means for a unit vector $\mathbf{w}$.
- State [Theorem 24.10.2](#), including the value of the minimum and the claim about the mean.
- True or false: the principal directions of a data set are the eigenvectors of $\mathbf{X} \mathbf{X}^\top$. Justify your answer.
- True or false: if $N = 3$ points in $\mathbb{R}^{10}$ are given, then $\mathbf{S}$ has rank at most 2. Justify your answer.
- Why should a computation of the principal directions avoid forming $\mathbf{X}_c \mathbf{X}_c^\top$?

Solution.

- $\mathbf{X}_c = \mathbf{X} - \bar{\mathbf{x}} \mathbf{1}^\top$, the data matrix with the mean column subtracted from every column, and $\mathbf{S} = \frac{1}{N} \mathbf{X}_c \mathbf{X}_c^\top$ ([Definition 24.10.1](#), [Definition 24.10.2](#)). For a unit $\mathbf{w}$, $\mathbf{w}^\top \mathbf{S} \mathbf{w}$ is the variance of the N numbers $\mathbf{w}^\top \mathbf{x}_j$, that is the spread of the data in the direction $\mathbf{w}$.
- Among all affine subspaces of $\mathbb{R}^n$ of dimension at most $k \leq \min(n, N)$, the total squared distance $\sum_j d(\mathbf{x}_j, A)^2$ is smallest at $A_0 = \bar{\mathbf{x}} + \text{span}(\mathbf{u}_1, \dots, \mathbf{u}_k)$, the minimum value is $\sum_{i>k} \sigma_i^2$ with σ_i the singular values of $\mathbf{X}_c$, and every minimizer contains $\bar{\mathbf{x}}$.
- False in general; they are the eigenvectors of $\mathbf{X}_c \mathbf{X}_c^\top$. For the four points of [Example \(Four points, done exactly\)](#), $\mathbf{X} \mathbf{X}^\top = \begin{pmatrix} 46 & 30 \\ 30 & 26 \end{pmatrix}$, whose eigenvalues are $36 \pm 10\sqrt{10}$ and whose top eigenvector is not $(1, 1)$, while the top principal direction is. The two **matrices** agree exactly when $\bar{\mathbf{x}} = \mathbf{0}$, since $\mathbf{X} \mathbf{X}^\top = \mathbf{X}_c \mathbf{X}_c^\top + N \bar{\mathbf{x}} \bar{\mathbf{x}}^\top$. Their **eigenvectors** can agree without that: for the two points $(0, 0)$ and $(2, 0)$ the mean is $(1, 0) \neq \mathbf{0}$, yet $\mathbf{X} \mathbf{X}^\top = \text{diag}(4, 0)$ and $\mathbf{X}_c \mathbf{X}_c^\top = \text{diag}(2, 0)$ have the same eigenvectors.
- True. $\text{rank } \mathbf{S} = \text{rank } \mathbf{X}_c$ — since $\text{nul}(\mathbf{X}_c^\top) = \text{nul}(\mathbf{X}_c \mathbf{X}_c^\top)$, because $\mathbf{X}_c \mathbf{X}_c^\top \mathbf{w} = \mathbf{0}$ gives $\|\mathbf{X}_c^\top \mathbf{w}\|^2 = \mathbf{w}^\top \mathbf{X}_c \mathbf{X}_c^\top \mathbf{w} = 0$ — and $\mathbf{X}_c \mathbf{1} = \mathbf{0}$ with $\mathbf{X}_c$ having 3 columns forces $\text{rank } \mathbf{X}_c \leq 2$. Centering costs at most one from the rank and sometimes nothing:

$$\text{rank } \mathbf{X} - 1 \leq \text{rank } \mathbf{X}_c \leq \text{rank } \mathbf{X},$$

the right inequality because $\mathbf{X}_c = \mathbf{X}(\mathbf{I}_N - \frac{1}{N} \mathbf{1} \mathbf{1}^\top)$ is a product ([Theorem 3.5.9](#)), the left because $\mathbf{X} = \mathbf{X}_c + \bar{\mathbf{x}} \mathbf{1}^\top$ adds to it a matrix of rank at most 1 ([Exercise \(C1: Rank](#)

of a sum)). The drop need not happen: for $\mathbf{x}_1 = (1, 0)$, $\mathbf{x}_2 = (2, 0)$, $\mathbf{x}_3 = (0, 1)$ the mean is $(1, \frac{1}{3}) \neq \mathbf{0}$, yet $\text{rank } \mathbf{X} = \text{rank } \mathbf{X}_c = 2$.

- (e) Because forming the product destroys the small singular values. Each entry of $\mathbf{X}_c \mathbf{X}_c^\top$ is an inner product, so the computed product is $\mathbf{X}_c \mathbf{X}_c^\top + \Delta$ with $\|\Delta\|_2$ of order $c(n, N) u \sigma_1^2$ (Theorem 24.1.2); if Δ is symmetric, Corollary 17.3.6 moves every eigenvalue by at most that, so every σ_i below about $\sqrt{u} \sigma_1$ becomes indistinguishable from 0, while a backward stable factorization of $\mathbf{X}_c$ itself keeps it — and those are exactly the singular values a low-rank fit has to decide about. The conditioning statement $\kappa_2(\mathbf{X}_c \mathbf{X}_c^\top) = \kappa_2(\mathbf{X}_c)^2$ says the same thing in digits (Theorem 16.8.4, Theorem 24.3.2), but it needs $\mathbf{X}_c^\top$ to have full column rank, which fails in part (d) above; the Weyl count needs no hypothesis on the rank.

24.10.6.2 B. Practice

Exercise. B1: A four-point cloud

The points $(1, 1)$, $(3, 5)$, $(5, 3)$, $(7, 7)$ are given in $\mathbb{R}^2$.

- Compute $\bar{\mathbf{x}}$, $\mathbf{X}_c$, $\mathbf{X}_c \mathbf{X}_c^\top$, and $\mathbf{S}$.
- Find the singular values of $\mathbf{X}_c$ and the principal directions.
- Find the best-fitting line, the total squared distance to it, and the fraction of variance it explains. Verify the total squared distance directly.

Solution.

- (a) The sum of the four points is $(16, 16)$, so $\bar{\mathbf{x}} = (4, 4)$ and

$$\mathbf{X}_c = \begin{pmatrix} -3 & -1 & 1 & 3 \\ -3 & 1 & -1 & 3 \end{pmatrix}, \quad \mathbf{X}_c \mathbf{X}_c^\top = \begin{pmatrix} 20 & 16 \\ 16 & 20 \end{pmatrix},$$

so $\mathbf{S} = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}$.

- (b) $\mathbf{X}_c \mathbf{X}_c^\top$ sends $(1, 1)$ to $36(1, 1)$ and $(1, -1)$ to $4(1, -1)$, so $\sigma_1^2 = 36$, $\sigma_2^2 = 4$, that is $\sigma_1 = 6$, $\sigma_2 = 2$, with $\mathbf{u}_1 = \frac{1}{\sqrt{2}}(1, 1)$ and $\mathbf{u}_2 = \frac{1}{\sqrt{2}}(1, -1)$. The eigenvalues of $\mathbf{S}$ are 9 and 1, which are $\sigma_i^2/4$ (Proposition 24.10.3 (b)).
- (c) By Theorem 24.10.2 the best line is $(4, 4) + \text{span}((1, 1))$, that is $y = x$, with total squared distance $\sigma_2^2 = 4$. By Theorem 24.10.4 (d) it explains $36/40 = 90\%$ of the variance. Directly: $(1, 1)$ and $(7, 7)$ lie on $y = x$; the centered points $(-1, 1)$ and $(1, -1)$ both project to $\mathbf{0}$ on $\text{span}((1, 1))$, so each contributes 2. Total 4.

Exercise. B2: Variance direction by direction

For the data of Exercise B1, compute the variance in the directions $\mathbf{e}_1$, $\mathbf{e}_2$, $\frac{1}{\sqrt{2}}(1, 1)$, and $\frac{1}{\sqrt{2}}(1, -1)$, and say which of them is largest and why that was predictable.

Solution. With $\mathbf{S} = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}$ and [Definition 24.10.2](#), the variance in a unit direction $\mathbf{w}$ is $\mathbf{w}^\top \mathbf{S} \mathbf{w}$:

$$\mathbf{e}_1 \mapsto 5, \quad \mathbf{e}_2 \mapsto 5, \quad \frac{(1,1)}{\sqrt{2}} \mapsto 9, \quad \frac{(1,-1)}{\sqrt{2}} \mapsto 1.$$

The largest is 9, in the direction $\mathbf{u}_1$. This was predictable from [Proposition 24.10.3](#) (c) and Proposition 17.1.1: the variance in a unit direction is the Rayleigh quotient of $\mathbf{S}$, whose maximum is $\lambda_1(\mathbf{S}) = 9$ and is attained at a unit eigenvector for λ_1 . Note also $5 + 5 = 9 + 1 = \text{tr } \mathbf{S}$: the total variance is the same however the orthonormal frame is chosen.

Exercise. B3: Changing the units

The centered data matrix of a two-coordinate data set is $\mathbf{X}_c = \begin{pmatrix} 3 & -3 & 0 & 0 \\ 0 & 0 & 4 & -4 \end{pmatrix}$. The second coordinate is then re-expressed in units c times smaller, which multiplies row 2 by $c > 0$.

- Find the principal directions before the change.
- Find the value of c at which the first principal direction switches from one coordinate axis to the other, and say what happens exactly at that value.

Solution.

- $\mathbf{X}_c \mathbf{X}_c^\top = \text{diag}(18, 32)$, so the principal directions are $\mathbf{u}_1 = \mathbf{e}_2$ (variance $32/4 = 8$) and $\mathbf{u}_2 = \mathbf{e}_1$ (variance $18/4 = 4.5$).
- After the change the centered matrix is $\mathbf{D}\mathbf{X}_c$ with $\mathbf{D} = \text{diag}(1, c)$, and $(\mathbf{D}\mathbf{X}_c)(\mathbf{D}\mathbf{X}_c)^\top = \text{diag}(18, 32c^2)$. The first principal direction is $\mathbf{e}_2$ when $32c^2 > 18$, that is $c > 3/4$, and $\mathbf{e}_1$ when $c < 3/4$. At $c = 3/4$ the two eigenvalues are equal, $\sigma_1 = \sigma_2$, and **every** direction is a direction of greatest variance: the principal directions are not determined at all, and the best-fitting line of [Theorem 24.10.2](#) is not unique, exactly as in the Quick check above.

24.10.6.3 C. Going deeper

Exercise. C1: The mean is forced, and the projection is not a shortcut

Let $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^n$ with mean $\bar{\mathbf{x}}$.

- Prove that the point $\mathbf{p} \in \mathbb{R}^n$ minimizing $\sum_j \|\mathbf{x}_j - \mathbf{p}\|^2$ is unique and equals $\bar{\mathbf{x}}$. (This is the case $k = 0$, which [Theorem 24.10.2](#) excludes, so it has to be proved rather than quoted.)
- Hence deduce that for every subspace U , the affine subspace $\mathbf{p} + U$ minimizing D over translates of U is $\bar{\mathbf{x}} + U$.
- Give an example of data and a subspace U for which $\bar{\mathbf{x}} + U$ is **not** the best-fitting affine subspace of its dimension, so that the choice of U still matters.

Solution.

- (a) Apply [Lemma 24.10.1](#) (b) with $U = \{\mathbf{0}\}$, so that $Q = \mathbf{I}_n$:

$$\sum_j \|\mathbf{x}_j - \mathbf{p}\|^2 = \sum_j \|\mathbf{x}_j - \bar{\mathbf{x}}\|^2 + N\|\bar{\mathbf{x}} - \mathbf{p}\|^2.$$

The first term does not depend on $\mathbf{p}$ and the second is ≥ 0 , vanishing exactly when $\mathbf{p} = \bar{\mathbf{x}}$. So the minimum is attained at $\bar{\mathbf{x}}$ and nowhere else.

- (b) By [Lemma 24.10.1](#) (a), $D(\mathbf{p} + U) = \sum_j \|Q(\mathbf{x}_j - \mathbf{p})\|^2$, and by part (b) of that lemma this exceeds $D(\bar{\mathbf{x}} + U)$ unless $\bar{\mathbf{x}} \in \mathbf{p} + U$ — in which case $\mathbf{p} + U = \bar{\mathbf{x}} + U$, the two cosets having a common point and the same direction space. So the minimizing translate is $\bar{\mathbf{x}} + U$, and as a set it is unique.
- (c) Take the four points of [Example \(Four points, done exactly\)](#) and $U = \text{span}(\mathbf{e}_1)$. Then $\bar{\mathbf{x}} + U$ is the horizontal line $y = 2$, and the centered points $(2, 2), (1, -1), (-1, 1), (-2, -2)$ are at squared distances 4, 1, 1, 4 from U , total 10. The best line is $\bar{\mathbf{x}} + \text{span}((1, 1))$, with total 4. Both pass through the mean; only one has the right direction. Choosing the translate is free, choosing the direction space is the whole problem.

Exercise. C2: Total variance is a trace, and what that buys

Let $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^n$ with sample covariance $\mathbf{S}$, and let $\mathbf{Q} \in M_n(\mathbb{R})$ be orthogonal.

- (a) Prove that the data $\mathbf{Q}\mathbf{x}_1, \dots, \mathbf{Q}\mathbf{x}_N$ have sample covariance $\mathbf{Q}\mathbf{S}\mathbf{Q}^\top$, and deduce that the total variance $\text{tr } \mathbf{S}$ is unchanged.
- (b) Prove that for any orthonormal basis $(\mathbf{q}_1, \dots, \mathbf{q}_n)$ of $\mathbb{R}^n$, the sum of the variances of the data in the n directions $\mathbf{q}_i$ equals $\text{tr } \mathbf{S}$.
- (c) Deduce that no orthonormal frame can make every coordinate variance small: some $\mathbf{q}_i$ always has variance at least $\text{tr } \mathbf{S}/n$. For the data of [Example \(Four points, done exactly\)](#), which frames attain that bound?

Solution.

- (a) The mean of the transformed data is $\frac{1}{N} \sum_j \mathbf{Q}\mathbf{x}_j = \mathbf{Q}\bar{\mathbf{x}}$, so the transformed centered matrix is $\mathbf{Q}\mathbf{X} - \mathbf{Q}\bar{\mathbf{x}}\mathbf{1}^\top = \mathbf{Q}\mathbf{X}_c$. Hence its sample covariance is $\frac{1}{N}(\mathbf{Q}\mathbf{X}_c)(\mathbf{Q}\mathbf{X}_c)^\top = \mathbf{Q}\mathbf{S}\mathbf{Q}^\top$. By Theorem 0.9.6, $\text{tr}(\mathbf{Q}\mathbf{S}\mathbf{Q}^\top) = \text{tr}(\mathbf{S}\mathbf{Q}^\top\mathbf{Q}) = \text{tr } \mathbf{S}$, since $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_n$.
- (b) By [Definition 24.10.2](#) the variance in the unit direction $\mathbf{q}_i$ is $\mathbf{q}_i^\top \mathbf{S} \mathbf{q}_i$. Let $\mathbf{Q}$ be the orthogonal matrix with columns $\mathbf{q}_1, \dots, \mathbf{q}_n$. Then $\mathbf{q}_i^\top \mathbf{S} \mathbf{q}_i$ is the (i, i) entry of $\mathbf{Q}^\top \mathbf{S} \mathbf{Q}$, so the sum over i is $\text{tr}(\mathbf{Q}^\top \mathbf{S} \mathbf{Q}) = \text{tr } \mathbf{S}$ by part (a) applied to $\mathbf{Q}^\top$.
- (c) A sum of n numbers equal to $\text{tr } \mathbf{S}$ has a term at least the average $\text{tr } \mathbf{S}/n$, since otherwise the sum would be $< \text{tr } \mathbf{S}$. For [Example \(Four points, done exactly\)](#), $\text{tr } \mathbf{S} = 5$ and the bound is $5/2$; the variances in a frame $(\mathbf{q}_1, \mathbf{q}_2)$ are $\mathbf{q}_1^\top \mathbf{S} \mathbf{q}_1$ and $\mathbf{q}_2^\top \mathbf{S} \mathbf{q}_2$, which sum to 5, so both equal $5/2$ exactly when $\mathbf{q}_1^\top \mathbf{S} \mathbf{q}_1 = 5/2$. Writing $\mathbf{q}_1 = a\mathbf{u}_1 + b\mathbf{u}_2$ with $a^2 + b^2 = 1$ and using $\mathbf{S}\mathbf{u}_1 = 4\mathbf{u}_1$, $\mathbf{S}\mathbf{u}_2 = \mathbf{u}_2$, this reads $4a^2 + b^2 = 5/2$, and subtracting $a^2 + b^2 = 1$ gives $3a^2 = 3/2$, that is $a^2 = b^2 = 1/2$. So the bound is attained exactly by the frames whose vectors split their weight equally between $\mathbf{u}_1$ and $\mathbf{u}_2$ — for instance the standard frame $(\mathbf{e}_1, \mathbf{e}_2)$, since $\mathbf{e}_1 = \frac{1}{\sqrt{2}}(\mathbf{u}_1 + \mathbf{u}_2)$; its two variances are indeed $5/2$ each. The extreme case of the spread is visible only in the principal frame; an arbitrary frame hides it.