EULERPAD WORKBOOKS

Linear Algebra for Machine Learning by Hand

250 problems from matrix products and tensor shapes to the SVD, PCA, attention and LoRA, with every step worked

Free sample

Chapter 18: PCA and the covariance matrix, problems and worked solutions

First edition, November 2026 · full book: 250 problems in 25 chapters

About this sample

This is Chapter 18 of *Linear Algebra for Machine Learning by Hand*, complete: its 10 problems, then the worked solution to each one and the mistakes that most often go wrong on them. The full workbook has 250 problems in 25 chapters, 16 of them written only for the workbook, in two volumes: *Problems* and *Worked Solutions*.

About the workbook

This workbook covers the linear algebra behind machine learning, from the four ways to read a matrix product to the SVD, and then puts it to work inside models: tensor shapes and broadcasting, einsum, least squares and ridge, condition numbers, PCA, PageRank, attention, LoRA, convolutions, kernels and graph Laplacians.

Problems are small enough to do by hand and concrete enough to check: three tokens of attention, a 4×4 image compressed by the SVD, a six-node graph split by its Fiedler vector. Each asks for a number, a matrix or a shape, and each solution shows every line.

Figures are part of the solutions. Shapes are drawn as blocks sized to their dimensions, products as matrices with the rows and columns that meet shaded, maps as what they do to the unit square and circle, and every figure is computed from the problem's own numbers.

Chapters marked as workbook-only appear nowhere else. The rest also have free versions at eulerpad.app/practice; this edition collects them in teaching order, renumbers their cross-references, and adds a full answer key.

Every numerical claim in every solution is verified by a computer check (finite differences, simulation or exact arithmetic) before the book is built.

© 2026 Spencer Uresk · EulerPad · eulerpad.app
You are welcome to share this free sample unchanged.

Contents of the full workbook

VECTORS, PRODUCTS AND SHAPES

1 Vector norms, distances and cosine similarity WORKBOOK-ONLY

2 Vectors in 3-D: dot and cross products, lines and planes

3 Matrix multiplication four ways WORKBOOK-ONLY

4 Tensor shapes, batching and broadcasting WORKBOOK-ONLY

5 Index notation and einsum WORKBOOK-ONLY

SYSTEMS, SUBSPACES AND MAPS

6 Gaussian elimination, inverses and determinants by hand

7 Rank and the four fundamental subspaces WORKBOOK-ONLY

8 Linear transformations and change of basis

9 Determinants as volume: change of variables and normalizing flows WORKBOOK-ONLY

10 Orthogonal matrices, rotations and rotary position embeddings WORKBOOK-ONLY

ORTHOGONALITY AND LEAST SQUARES

11 Orthogonal bases, Gram–Schmidt and the QR decomposition

12 Least squares, projections and the SVD

13 Linear regression in matrix form WORKBOOK-ONLY

14 Matrix norms, condition numbers and numerical stability WORKBOOK-ONLY

EIGENVALUES AND THE SVD

15 Eigenvalues and eigenvectors by hand

16 Quadratic forms and positive definite matrices

17 The SVD: geometry and low-rank approximation WORKBOOK-ONLY

18 PCA and the covariance matrix IN THIS SAMPLE

19 Power iteration, Rayleigh quotients and PageRank WORKBOOK-ONLY

20 Matrix exponentials and linear systems of ODEs

LINEAR ALGEBRA INSIDE MODELS

21 Attention as matrix products WORKBOOK-ONLY

22 Low-rank adaptation, parameter counts and matmul FLOPs WORKBOOK-ONLY

23 Convolution as a matrix: Toeplitz, im2col and circulant matrices WORKBOOK-ONLY

24 Kernels and Gram matrices WORKBOOK-ONLY

25 Graph Laplacians and spectral clustering WORKBOOK-ONLY

CHAPTER 18

PCA and the covariance matrix

Principal component analysis is the eigen-decomposition of a covariance matrix, read as geometry: the eigenvectors are the directions along which the data spread most, the eigenvalues are the variances along them, and keeping the top few directions is the projection with the least squared error among all projections of that dimension. Every property of PCA that gets quoted, that the components are uncorrelated, that the explained variance ratios sum to one, that the reconstruction error is the sum of the discarded eigenvalues, that the SVD computes the same thing, is a short calculation with a symmetric matrix. These ten problems do each calculation on four points in the plane, where every number is exact, and prove the general statement alongside: the covariance and its decomposition, the variance along a direction, total and explained variance, scores and reconstructions, the error and its optimality, the SVD route, whitening, and what happens when one feature is rescaled. The five mistakes at the end are the ones that produce components that look fine: the mean left in, directions read from the left singular vectors, explained variance computed from unsquared singular values, reconstructions without the mean added back, and eigenvectors taken in the order the library returns them.

BEFORE YOU START

- Data are $X \in \mathbb{R}^{N\times d}$ with rows $x_n^\top$ (examples). The mean is $\bar{x} = \frac{1}{N}\sum_n x_n$, the centered data are $X_c = X - \mathbf{1}\bar{x}^\top$ with rows $(x_n - \bar{x})^\top$, and the covariance is $S = \frac{1}{N}X_c^\top X_c = \frac{1}{N}\sum_n (x_n-\bar{x})(x_n-\bar{x})^\top$, the "Variance, covariance and correlation" set's $\frac{1}{N}X^\top H X$. This chapter divides by N throughout; dividing by $N-1$ instead multiplies every eigenvalue by $\frac{N}{N-1}$ and changes nothing else.
- S is symmetric and positive semidefinite (Problem 6 of the "Variance, covariance and correlation" set), so by Chapter 15 $S = Q\Lambda Q^\top$ with Q orthogonal, columns $q_1, \dots, q_d$ (the principal directions), and $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$ with $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$. Q_k is the first k columns of Q.
- The scores of x_n on the first k components are $z_n = Q_k^\top(x_n - \bar{x}) \in \mathbb{R}^k$, collected as $Z = X_c Q_k \in \mathbb{R}^{N\times k}$; the reconstruction is $\hat{x}_n = \bar{x} + Q_k z_n$. Chapter 12's projection facts are used: for W with orthonormal columns, $P = WW^\top$ satisfies $P^2 = P = P^\top$.
- The thin SVD of the centered data is $X_c = U\Sigma V^\top$ with $U^\top U = V^\top V = I$ and singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ on the diagonal of Σ (Chapter 12).
- The four points used throughout are $(5,3)$, $(3,5)$, $(-1,1)$ and $(1,-1)$, the rows of $X \in \mathbb{R}^{4\times 2}$, so $N = 4$ and $d = 2$.
- tr is the trace, with $\operatorname{tr}(AB) = \operatorname{tr}(BA)$ and $\operatorname{tr}(vv^\top) = v^\top v = \|v\|^2$; $\mathbf{1}$ is the all-ones vector and $e_1 = (1,0)^\top$.

18.1 ●○○ Compute $\bar{x}$, the centered data X_c and the covariance S for the four points.

18.2 ●○○

Find the eigenvalues and unit eigenvectors of S and write $S = Q\Lambda Q^{\top}$.

18.3 ●●○

For a unit vector u, show that the variance of the projected data $u^{\top}(x_n - \bar{x})$ is $u^{\top} S u$, and show with a Lagrange multiplier that it is largest at $u = q_1$, where it equals λ_1. Evaluate it for the four points at $u = q_1$ and at $u = e_1$.

18.4 ●●○

Show that $\operatorname{tr} S = \sum_{j} \lambda_j = \frac{1}{N}\sum_{n} \|x_n - \bar{x}\|^2$, and compute the fraction of the total variance carried by the first component for the four points.

18.5 ●●○

Compute the scores $z_n = q_1^{\top}(x_n - \bar{x})$ and the one-component reconstructions $\hat{x}_n = \bar{x} + z_n q_1$ for the four points, and show that $\hat{x}_n - \bar{x} = q_1 q_1^{\top}(x_n - \bar{x})$ is a projection of the centered point.

18.6 ●●○

Show that the mean squared reconstruction error with k components, $\frac{1}{N}\sum_{n} \|x_n - \hat{x}_n\|^2$, equals $\sum_{j>k} \lambda_j$, and compute it for the four points with $k = 1$.

18.7 ●●●

Let $W \in \mathbb{R}^{d \times k}$ have orthonormal columns and reconstruct by $\hat{x}_n = \bar{x} + W W^{\top}(x_n - \bar{x})$. Show that the mean squared error is $\operatorname{tr} S - \operatorname{tr}(W^{\top} S W)$, and that $\operatorname{tr}(W^{\top} S W) \le \sum_{j \le k} \lambda_j$ with equality when the columns of W span $q_1, \dots, q_k$. Conclude that the PCA subspace minimizes the error.

18.8 ●●○

Let $X_c = U\Sigma V^\top$. Show that $S = \frac{1}{N} V \Sigma^2 V^\top$, so the principal directions are the right singular vectors with $\lambda_j = \sigma_j^2 / N$, and that the scores are $Z = X_c V = U\Sigma$. Find the singular values of X_c for the four points.

18.9 ●●○

Let $Z = X_c Q$ be the full $N \times d$ matrix of scores. Show that its columns have mean zero and that $\frac{1}{N} Z^\top Z = \Lambda$, so the scores are uncorrelated with variances λ_j, and that $W = Z \Lambda^{-1/2}$ satisfies $\frac{1}{N} W^\top W = I$ when every $\lambda_j > 0$.

18.10 ●●●

Scale the second feature by $c = 10$: $X' = XD$ with $D = \mathrm{diag}(1, 10)$. Show that $S' = DSD$, find its eigenvalues and top eigenvector to three decimal places, and show that the correlation matrix $R = D_S^{-1/2} S D_S^{-1/2}$, with $D_S = \mathrm{diag}(S_{11}, S_{22})$, is unchanged by the scaling. Find R and its eigen-decomposition for the four points.

CHAPTER 18 · WORKED SOLUTIONS

PCA and the covariance matrix

ANSWER KEY

18.1 $\bar{x} = (2, 2)^\top$; X_c has rows $(3, 1)$, $(1, 3)$, $(-3, -1)$, $(-1, -3)$; $S = \begin{pmatrix} 5 & 3 \\ 3 & 5 \end{pmatrix}$

18.2 $\lambda_1 = 8$, $\lambda_2 = 2$; $q_1 = \frac{1}{\sqrt{2}}(1, 1)^\top$, $q_2 = \frac{1}{\sqrt{2}}(1, -1)^\top$; $S = Q\Lambda Q^\top$ with $Q = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ and $\Lambda = \operatorname{diag}(8, 2)$

18.3 $\operatorname{Var}\left(u^\top(x - \bar{x})\right) = u^\top S u \le \lambda_1$, with equality at $u = q_1$; for the four points it is 8 along q_1 and 5 along e_1

18.4 $\operatorname{tr} S = \sum_j \lambda_j = \frac{1}{N}\sum_n \|x_n - \bar{x}\|^2$; for the four points the first component explains $8/10 = 80\%$ of the variance

18.5 Scores $(2\sqrt{2}, 2\sqrt{2}, -2\sqrt{2}, -2\sqrt{2})$; reconstructions $(4, 4)^\top$, $(4, 4)^\top$, $(0, 0)^\top$, $(0, 0)^\top$; $\hat{x}_n - \bar{x} = q_1 q_1^\top (x_n - \bar{x})$ is the orthogonal projection of the centered point onto the line through q_1

18.6 $\frac{1}{N}\sum_n \|x_n - \hat{x}_n\|^2 = \sum_{j>k} \lambda_j$; for the four points with $k = 1$ the error is $\lambda_2 = 2$

18.7 $\text{Error}(W) = \operatorname{tr} S - \operatorname{tr}(W^\top S W) \ge \operatorname{tr} S - \sum_{j \le k} \lambda_j = \sum_{j>k} \lambda_j$, with equality when the columns of W span $q_1, \dots, q_k$

18.8 $S = \frac{1}{N} V \Sigma^2 V^\top$: $q_j = v_j$ and $\lambda_j = \sigma_j^2 / N$; the scores are $Z = U\Sigma$; for the four points $\sigma_1 = 4\sqrt{2}$ and $\sigma_2 = 2\sqrt{2}$

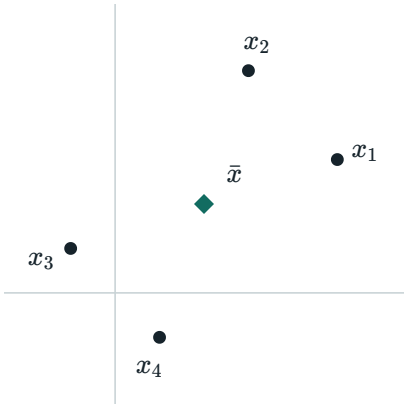
18.9 $\frac{1}{N} Z^\top Z = \Lambda$: the scores are uncorrelated, with variance λ_j along the j-th component; $W = X_c Q \Lambda^{-1/2}$ has identity covariance

18.10 $S' = DSD = \begin{pmatrix} 5 & 30 \\ 30 & 500 \end{pmatrix}$ with $\lambda \approx 501.81, 3.19$ and $q_1' \approx (0.060, 0.998)^\top$, against $(0.707, 0.707)^\top$ before scaling; $R = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}$ is unchanged by any positive rescaling of the features, with eigenvalues 1.6 and 0.4 and Problem 18.2's eigenvectors

Problem 18.1

●○○

> Compute $\bar{x}$, the centered data X_c and the covariance S for the four points.

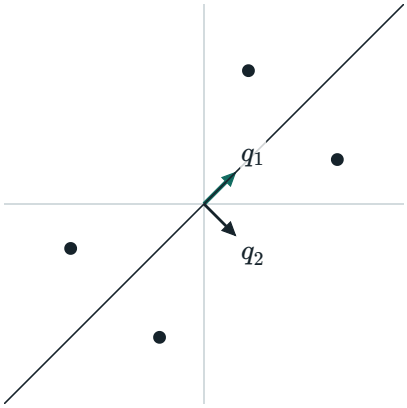
1	$\bar{x} = \frac{1}{4}\big((5,3)+(3,5)+(-1,1)+(1,-1)\big)^\top = \frac{1}{4}(8,8)^\top = (2,2)^\top$. 	Average each coordinate over the four points.
2	The centered rows are $(3,1)$, $(1,3)$, $(-3,-1)$ and $(-1,-3)$.	Subtract $\bar{x}^\top$ from each row; they sum to zero, as centered rows must.
3	$X_c^\top X_c = \begin{pmatrix} 9+1+9+1 & 3+3+3+3 \\ 12 & 1+9+1+9 \end{pmatrix} = \begin{pmatrix} 20 & 12 \\ 12 & 20 \end{pmatrix}$.	$M^\top M$ sums the outer products of the rows of M: the $(1,1)$ entry is $\sum_n (x_{n1}-\bar{x}_1)^2$ and the $(1,2)$ entry $\sum_n (x_{n1}-\bar{x}_1)(x_{n2}-\bar{x}_2)$.
4	$S = \frac{1}{4}\begin{pmatrix} 20 & 12 \\ 12 & 20 \end{pmatrix} = \begin{pmatrix} 5 & 3 \\ 3 & 5 \end{pmatrix}$.	Divide by $N = 4$.
5	$\bar{x} = (2,2)^\top$; X_c **has rows** $(3,1)$, $(1,3)$, $(-3,-1)$, $(-1,-3)$; $S = \begin{pmatrix} 5 & 3 \\ 3 & 5 \end{pmatrix}$	$S_{11} = 5$ is the variance of the first coordinate, $S_{12} = 3$ the covariance, and the correlation is $3/5 = 0.6$ (Problem 18.10). This is `np.cov(X.T, bias=True)`; without `bias=True` every entry is $\frac{4}{3}$ as large. Leaving the mean in is Mistake 1.

Problem 18.2

●○○

> Find the eigenvalues and unit eigenvectors of S and write $S = Q\Lambda Q^\top$.

1	$\det(S-\lambda I) = (5-\lambda)^2 - 9 = 0$, so $5-\lambda = \pm 3$ and $\lambda_1 = 8$, $\lambda_2 = 2$.	The characteristic polynomial of a 2×2 matrix, as in Chapter 15.
2	$\lambda_1 = 8$: $(S-8I)v = \begin{pmatrix} -3 & 3 \\ 3 & -3 \end{pmatrix} v = 0$ gives $v_1 = v_2$, so $q_1 = \frac{1}{\sqrt{2}}(1,1)^\top$.	Solve the homogeneous system and normalize.

3	$\lambda_2 = 2$: $\begin{pmatrix} 3 & 3 \\ 3 & 3 \end{pmatrix} v = 0$ gives $v_2 = -v_1$, so $q_2 = \frac{1}{\sqrt{2}}(1, -1)^\top$. 	Solve and normalize.
4	$Q\Lambda Q^\top = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 8 & 0 \\ 0 & 2 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 8 & 2 \\ 8 & -2 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 10 & 6 \\ 6 & 10 \end{pmatrix} = S.$	Multiply out, with $Q = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ and the two factors of $\frac{1}{\sqrt{2}}$ combining into $\frac{1}{2}$.
5	$\lambda_1 = 8$, $\lambda_2 = 2$; $q_1 = \frac{1}{\sqrt{2}}(1, 1)^\top$, $q_2 = \frac{1}{\sqrt{2}}(1, -1)^\top$; $S = Q\Lambda Q^\top$ **with** $Q = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ **and** $\Lambda = \operatorname{diag}(8, 2)$	The eigenvectors are orthogonal because S is symmetric (Problem 15.8), and each one's sign is a free choice: $-q_1$ serves as well, which is why the check compares directions up to sign. $\lambda_1 + \lambda_2 = 10 = \operatorname{tr} S$ (Problem 18.4). The top direction is the diagonal, along which the four points visibly spread most; taking the eigenvectors in a library's ascending order is Mistake 5.

Problem 18.3

●●○

> For a unit vector u, show that the variance of the projected data $u^\top(x_n - \bar{x})$ is $u^\top S u$, and show with a Lagrange multiplier that it is largest at $u = q_1$, where it equals λ_1. Evaluate it for the four points at $u = q_1$ and at $u = e_1$.

1	The projections $t_n = u^\top(x_n - \bar{x})$ have mean 0, so their variance is $\frac{1}{N}\sum_n t_n^2 = \frac{1}{N}\sum_n u^\top(x_n - \bar{x})(x_n - \bar{x})^\top u = u^\top S u$.	$\sum_n (x_n - \bar{x}) = 0$; $t_n^2 = u^\top(x_n - \bar{x})(x_n - \bar{x})^\top u$; u comes out of the sum on both sides.
2	Maximize $u^\top S u$ subject to $u^\top u = 1$: $L = u^\top S u - \lambda(u^\top u - 1)$ and $\nabla_u L = 2Su - 2\lambda u = 0$, so $Su = \lambda u$.	The "Lagrange multipliers" set; $\nabla_u(u^\top S u) = 2Su$ for symmetric S, from the "Matrix calculus conventions" set.
3	At an eigenvector $u = q_j$, $u^\top S u = \lambda_j q_j^\top q_j = \lambda_j$.	The candidates are the unit eigenvectors.
4	The largest of these is λ_1, at q_1.	Compare the values.

5	Directly: write $u = \sum_j c_j q_j$ with $\sum_j c_j^2 = 1$; then $u^\top S u = \sum_j \lambda_j c_j^2 \le \lambda_1 \sum_j c_j^2 = \lambda_1$.	Expand in the orthonormal eigenbasis, use $S q_j = \lambda_j q_j$ and $q_i^\top q_j = 0$ for $i \neq j$; a weighted average of the λ_j is at most the largest.
6	For the four points, $q_1^\top S q_1 = 8$ and $e_1^\top S e_1 = S_{11} = 5$.	Problem 18.2; the $(1,1)$ entry of S.
7	$\operatorname{Var}\left(u^\top (x - \bar{x})\right) = u^\top S u \le \lambda_1$, **with equality at $u = q_1$; for the four points it is 8 along q_1 and 5 along e_1**	The first principal direction is the direction of greatest variance, and the variance along any unit direction is a convex combination of the eigenvalues weighted by the squared cosines with the eigenvectors. Along the diagonal the four points sit at $\pm 2\sqrt{2}$ (Problem 18.5), with variance 8, more than either raw coordinate's 5. The same argument gives the minimum λ_d at q_d. The "Variance, covariance and correlation" set's $a^\top \Sigma a$ is this quadratic form for a population covariance.

Problem 18.4

●●○

> Show that $\operatorname{tr} S = \sum_j \lambda_j = \frac{1}{N} \sum_n \|x_n - \bar{x}\|^2$, and compute the fraction of the total variance carried by the first component for the four points.

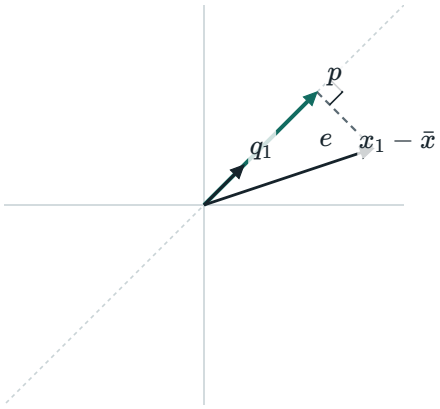
1	$\operatorname{tr} S = \operatorname{tr}(Q \Lambda Q^\top) = \operatorname{tr}(\Lambda Q^\top Q) = \operatorname{tr} \Lambda = \sum_j \lambda_j$.	The cyclic property of the trace, then $Q^\top Q = I$.
2	$\operatorname{tr} S = \frac{1}{N} \sum_n \operatorname{tr}\left((x_n - \bar{x})(x_n - \bar{x})^\top\right) = \frac{1}{N} \sum_n \|x_n - \bar{x}\|^2$.	The trace is linear and $\operatorname{tr}(v v^\top) = \|v\|^2$.
3	$\operatorname{tr} S = \sum_j S_{jj}$ is also the sum of the per-coordinate variances.	The diagonal entries of S are the variances of the coordinates.
4	For the four points, $\operatorname{tr} S = 5 + 5 = 10 = 8 + 2$, and the first component carries $\lambda_1 / (\lambda_1 + \lambda_2) = 8/10$.	Problems 18.1 and 18.2.
5	$\operatorname{tr} S = \sum_j \lambda_j = \frac{1}{N} \sum_n \|x_n - \bar{x}\|^2$; **for the four points the first component explains $8/10 = 80\%$ of the variance**	The total variance is the mean squared distance from the mean, which a rotation of the data leaves unchanged, so splitting it over the eigenvalues is meaningful: λ_j is the variance along q_j and the fractions $\lambda_j / \operatorname{tr} S$ sum to 1. In terms of the SVD they are ratios of squared singular values (Problem 18.8), not of singular values (Mistake 3).

Problem 18.5

●●○

> Compute the scores $z_n = q_1^\top (x_n - \bar{x})$ and the one-component reconstructions $\hat{x}_n = \bar{x} + z_n q_1$ for the four

points, and show that $\hat{x}_n - \bar{x} = q_1 q_1^\top (x_n - \bar{x})$ is a projection of the centered point.

1	$z_n = \frac{1}{\sqrt{2}}(1,1) \cdot (x_n - \bar{x})$: $\frac{3+1}{\sqrt{2}} = 2\sqrt{2}$, $\frac{1+3}{\sqrt{2}} = 2\sqrt{2}$, $\frac{-3-1}{\sqrt{2}} = -2\sqrt{2}$ and $-2\sqrt{2}$. $\underset{X_c}{\begin{bmatrix} 3 & 1 \\ 1 & 3 \\ -3 & -1 \\ -1 & -3 \end{bmatrix}} \cdot \underset{q_1}{\begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}} = \underset{z}{\begin{bmatrix} 2\sqrt{2} \\ 2\sqrt{2} \\ -2\sqrt{2} \\ -2\sqrt{2} \end{bmatrix}}$	Dot product of q_1 with each centered row from Problem 18.1.
2	$z_n q_1 = \pm 2\sqrt{2} \cdot \frac{1}{\sqrt{2}}(1,1)^\top = \pm(2,2)^\top$.	The $\sqrt{2}$ factors cancel.
3	$\hat{x}_n = (2,2)^\top + (2,2)^\top = (4,4)^\top$ for $n = 1, 2$ and $(2,2)^\top - (2,2)^\top = (0,0)^\top$ for $n = 3, 4$.	Add the mean back.
4	$\hat{x}_n - \bar{x} = q_1\big(q_1^\top (x_n - \bar{x})\big) = q_1 q_1^\top (x_n - \bar{x})$, and $P = q_1 q_1^\top$ has $P^2 = q_1 (q_1^\top q_1) q_1^\top = P$ and $P^\top = P$. 	Regroup the scalar; $q_1^\top q_1 = 1$. This is Chapter 12's projection onto the line through q_1.
5	**Scores $(2\sqrt{2}, 2\sqrt{2}, -2\sqrt{2}, -2\sqrt{2})$; reconstructions $(4,4)^\top$, $(4,4)^\top$, $(0,0)^\top$, $(0,0)^\top$; $\hat{x}_n - \bar{x} = q_1 q_1^\top (x_n - \bar{x})$ is the orthogonal projection of the centered point onto the line through q_1**	Two points reconstruct to $(4,4)$ and two to $(0,0)$: the one-dimensional summary keeps where each point sits along the diagonal and forgets its ± 1 offset across it. The scores have mean 0 and variance $\frac{1}{4} \cdot 4 \cdot (2\sqrt{2})^2 = 8 = \lambda_1$ (Problem 18.9). In matrix form $\hat{X} = \mathbf{1}\bar{x}^\top + X_c q_1 q_1^\top$; dropping the first term is Mistake 4.

Problem 18.6

●●○

Show that the mean squared reconstruction error with k components, $\frac{1}{N}\sum_n \|x_n - \hat{x}_n\|^2$, equals $\sum_{j>k} \lambda_j$, and compute it for the four points with $k = 1$.

1	$x_n - \hat{x}_n = (x_n - \bar{x}) - Q_k Q_k^\top (x_n - \bar{x}) = (I - Q_k Q_k^\top)(x_n - \bar{x})$.	Problem 18.5's form with k directions; the $\bar{x}$ cancels.

2	$I - Q_k Q_k^\top = \sum_{j>k} q_j q_j^\top$.	$I = QQ^\top = \sum_{j=1}^{d} q_j q_j^\top$ because the q_j are an orthonormal basis; subtract the first k terms.
3	$\|x_n - \hat{x}_n\|^2 = \sum_{j>k} \left(q_j^\top (x_n - \bar{x})\right)^2$.	A vector in the span of the orthonormal q_j, $j > k$, has squared length equal to the sum of its squared coefficients, and the coefficients are the dot products.
4	$\frac{1}{N}\sum_n \|x_n - \hat{x}_n\|^2 = \sum_{j>k} \frac{1}{N}\sum_n \left(q_j^\top (x_n - \bar{x})\right)^2 = \sum_{j>k} q_j^\top S q_j = \sum_{j>k} \lambda_j$.	Swap the sums; Problem 18.3, step 1 with $u = q_j$; then $q_j^\top S q_j = \lambda_j$.
5	For the four points with $k = 1$ the error is $\lambda_2 = 2$; directly, $\|(5,3) - (4,4)\|^2 = 2$ and likewise for each point, mean 2.	Problem 18.5's reconstructions.
6	**$\frac{1}{N}\sum_n \|x_n - \hat{x}_n\|^2 = \sum_{j>k} \lambda_j$; for the four points with $k = 1$ the error is $\lambda_2 = 2$**	The variance the kept components carry and the error the dropped ones leave are the two halves of $\operatorname{tr} S$: 8 kept and 2 lost out of 10. Choosing k by a threshold on the explained variance is choosing it by this error. Problem 18.7 shows that no other k-dimensional subspace does better.

Problem 18.7

●●●

> Let $W \in \mathbb{R}^{d \times k}$ have orthonormal columns and reconstruct by $\hat{x}_n = \bar{x} + WW^\top (x_n - \bar{x})$. Show that the mean squared error is $\operatorname{tr} S - \operatorname{tr}(W^\top S W)$, and that $\operatorname{tr}(W^\top S W) \le \sum_{j \le k} \lambda_j$ with equality when the columns of W span $q_1, \dots, q_k$. Conclude that the PCA subspace minimizes the error.

1	With $v = x_n - \bar{x}$: $\|(I - WW^\top)v\|^2 = v^\top (I - WW^\top) v = \|v\|^2 - \|W^\top v\|^2$.	$I - WW^\top$ is a symmetric projection because $W^\top W = I_k$, and for a symmetric projection $\|Pv\|^2 = v^\top P^\top P v = v^\top P v$; then $v^\top W W^\top v = \|W^\top v\|^2$.
2	Averaging over n: error $= \frac{1}{N}\sum_n \|x_n - \bar{x}\|^2 - \frac{1}{N}\sum_n \|W^\top (x_n - \bar{x})\|^2 = \operatorname{tr} S - \operatorname{tr}(W^\top S W)$.	Problem 18.4 for the first term; for the second, $\|W^\top v\|^2 = \operatorname{tr}(W^\top v v^\top W)$ and the average of $vv^\top$ is S.
3	$\operatorname{tr}(W^\top S W) = \sum_j \lambda_j \|W^\top q_j\|^2 =: \sum_j \lambda_j w_j$.	Insert $S = \sum_j \lambda_j q_j q_j^\top$ and use $\operatorname{tr}(W^\top q_j q_j^\top W) = \|W^\top q_j\|^2$.
4	$0 \le w_j \le 1$ and $\sum_j w_j = k$.	$\|W^\top q_j\|^2 = q_j^\top W W^\top q_j = \|WW^\top q_j\|^2 \le \|q_j\|^2 = 1$ because $WW^\top$ is a projection; and $\sum_j w_j = \operatorname{tr}\left(W^\top (\sum_j q_j q_j^\top) W\right) = \operatorname{tr}(W^\top W) = \operatorname{tr} I_k = k$.

5	$\sum_j \lambda_j w_j \le \sum_{j \le k} \lambda_j$, with equality when $w_j = 1$ for $j \le k$ and 0 otherwise.	With the λ_j sorted and weights in $[0, 1]$ summing to k, moving weight from a smaller eigenvalue to a larger one that still has room never decreases the sum, so the maximum puts full weight on the k largest; $w_j = 1$ for $j \le k$ means each top eigenvector lies in the span of W's columns, which, there being k of them, is then exactly that span.
6	**Error$(W) = \operatorname{tr} S - \operatorname{tr}(W^\top S W) \ge \operatorname{tr} S - \sum_{j \le k} \lambda_j = \sum_{j > k} \lambda_j$, with equality when the columns of W span $q_1, \dots, q_k$**	PCA is the k-dimensional affine subspace of least squared perpendicular distance to the data, the flat-fitting version of least squares, and for $k = 1$ it is Problem 18.3 again. Only the span matters: WR for any orthogonal $k \times k$ matrix R gives the same projection and the same error, so the components are defined up to a rotation within the subspace, and when $\lambda_k = \lambda_{k+1}$ even the subspace is not unique.

Problem 18.8

●●○

> Let $X_c = U \Sigma V^\top$. Show that $S = \frac{1}{N} V \Sigma^2 V^\top$, so the principal directions are the right singular vectors with $\lambda_j = \sigma_j^2 / N$, and that the scores are $Z = X_c V = U \Sigma$. Find the singular values of X_c for the four points.

1	$X_c^\top X_c = V \Sigma U^\top U \Sigma V^\top = V \Sigma^2 V^\top$.	Transpose the SVD and multiply; $U^\top U = I$.
2	$S = \frac{1}{N} X_c^\top X_c = V \left(\frac{1}{N} \Sigma^2 \right) V^\top$, an eigendecomposition with $Q = V$ and $\Lambda = \Sigma^2 / N$.	V is orthogonal and Σ^2 / N is diagonal with nonincreasing entries, which is the form $Q \Lambda Q^\top$; the decomposition is unique up to the signs of the columns when the eigenvalues are distinct.
3	$Z = X_c V = U \Sigma V^\top V = U \Sigma$.	$V^\top V = I$.
4	For the four points, $\sigma_j = \sqrt{N \lambda_j}$: $\sigma_1 = \sqrt{32} = 4\sqrt{2}$ and $\sigma_2 = \sqrt{8} = 2\sqrt{2}$.	Step 2 with $N = 4$ and Problem 18.2's eigenvalues.
5	**$S = \frac{1}{N} V \Sigma^2 V^\top$: $q_j = v_j$ and $\lambda_j = \sigma_j^2 / N$; the scores are $Z = U \Sigma$; for the four points $\sigma_1 = 4\sqrt{2}$ and $\sigma_2 = 2\sqrt{2}$**	The SVD computes PCA without forming S, which matters numerically because squaring the singular values squares the condition number; it is what `sklearn.decomposition.PCA` does. The columns of U are the score vectors scaled to unit length, patterns over the examples rather than directions in feature space, which is Mistake 2. Chapter 12's singular values as square roots of the eigenvalues of $A^\top A$ is the same statement without the $\frac{1}{N}$.

Problem 18.9

●●○

Let $Z = X_c Q$ be the full $N \times d$ matrix of scores. Show that its columns have mean zero and that $\frac{1}{N} Z^\top Z = \Lambda$, so the scores are uncorrelated with variances λ_j, and that $W = Z\Lambda^{-1/2}$ satisfies $\frac{1}{N} W^\top W = I$ when every $\lambda_j > 0$.

1	$\mathbf{1}^\top Z = \mathbf{1}^\top X_c Q = 0$.	The columns of X_c sum to zero, so $\mathbf{1}^\top X_c = 0$.
2	$\frac{1}{N} Z^\top Z = \frac{1}{N} Q^\top X_c^\top X_c Q = Q^\top S Q = Q^\top Q \Lambda Q^\top Q = \Lambda$.	$S = Q\Lambda Q^\top$ and $Q^\top Q = I$.
3	$\frac{1}{N} W^\top W = \Lambda^{-1/2}\big(\frac{1}{N} Z^\top Z\big)\Lambda^{-1/2} = \Lambda^{-1/2}\Lambda\Lambda^{-1/2} = I$.	Diagonal matrices multiply entry by entry.
4	**$\frac{1}{N} Z^\top Z = \Lambda$: the scores are uncorrelated, with variance λ_j along the j-th component; $W = X_c Q \Lambda^{-1/2}$ has identity covariance**	PCA rotates the data into coordinates whose covariance is diagonal, and dividing each by its standard deviation $\sqrt{\lambda_j}$ is the "Variance, covariance and correlation" set's whitening with Q and Λ taken from the data. In SVD terms $Z = U\Sigma$ and $W = \sqrt{N}\,U$. Whitening a direction with a tiny λ_j amplifies whatever noise lives there by $1/\sqrt{\lambda_j}$, so in practice it is done after truncating to k components or with a small constant added to each λ_j.

Problem 18.10

●●●

Scale the second feature by $c = 10$: $X' = XD$ with $D = \mathrm{diag}(1, 10)$. Show that $S' = DSD$, find its eigenvalues and top eigenvector to three decimal places, and show that the correlation matrix $R = D_S^{-1/2} S D_S^{-1/2}$, with $D_S = \mathrm{diag}(S_{11}, S_{22})$, is unchanged by the scaling. Find R and its eigen-decomposition for the four points.

1	$\bar{x}' = D\bar{x}$ and $X'_c = X_c D$, so $S' = \frac{1}{N} D X_c^\top X_c D = DSD = \begin{pmatrix} 5 & 30 \\ 30 & 500 \end{pmatrix}$.	Scaling a column of X scales the same column of X_c; D is symmetric; $S'_{12} = 10 \cdot 3$ and $S'_{22} = 100 \cdot 5$.
2	The eigenvalues of S' are $\lambda = \dfrac{505 \pm \sqrt{505^2 - 4 \cdot 1600}}{2} = \dfrac{505 \pm \sqrt{248625}}{2} \approx 501.812$ and 3.188.	Trace 505, determinant $2500 - 900 = 1600$, and $\sqrt{248625} \approx 498.623$.
3	$(5 - 501.812)v_1 + 30 v_2 = 0$ gives $v_1 \approx 0.0604\, v_2$, so $q'_1 \approx (0.060, 0.998)^\top$, and the first component now carries $501.81/505 \approx 99.4\%$ of the variance.	The first row of $(S' - \lambda_1 I)v = 0$; normalize.
4	$D_{S'} = \mathrm{diag}(5, 500) = D D_S D$, so $D_{S'}^{-1/2} = D^{-1} D_S^{-1/2}$ and $R' = D^{-1} D_S^{-1/2}\, DSD\, D_S^{-1/2} D^{-1} = D_S^{-1/2} S D_S^{-1/2} = R$.	Diagonal matrices commute, and $D^{-1}D = I$ on each side.

5	For the four points, $D_S = \operatorname{diag}(5, 5)$ and $R = S/5 = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}$, with eigenvalues 1.6 and 0.4 and the eigenvectors q_1, q_2 of Problem 18.2.	A positive multiple of a matrix has the same eigenvectors and scaled eigenvalues; here the two variances happen to be equal, so R is a multiple of S.
6	**$S' = DSD = \begin{pmatrix} 5 & 30 \\ 30 & 500 \end{pmatrix}$ with $\lambda \approx 501.81, 3.19$ and $q_1' \approx (0.060, 0.998)^\top$, against $(0.707, 0.707)^\top$ before scaling; $R = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}$ is unchanged by any positive rescaling of the features, with eigenvalues 1.6 and 0.4 and Problem 18.2's eigenvectors**	PCA on the covariance is not scale-invariant: measuring the second feature in different units swings the first direction from the diagonal to almost the second axis, and the 99.4% is an artifact of the units. When features are in incommensurable units, standardize first, which is PCA on R, that is on $X_c D_S^{-1/2}$; when they share a unit, as pixels do, the covariance is the right choice, because the scale carries information.

Where this goes wrong

1. PCA on uncentered data

$X^\top X$ is the matrix in every normal equation, and the covariance differs from it only by the centering step.

1	$S = \frac{1}{N} X_c^\top X_c$	Right so far: Problem 18.1.
2	"The mean is a constant offset, so the directions of spread are the same with or without it."	The shortcut that causes the mistake: a constant offset is invisible to a variance but not to a second moment; $\frac{1}{N} X^\top X = S + \bar{x}\bar{x}^\top$, the "Variance, covariance and correlation" set's last mistake.
3	$S = \frac{1}{N} X^\top X$ ✗	With the four points shifted so that $\bar{x} = (2, 12)^\top$ (add 10 to every second coordinate, which changes no variance), $\frac{1}{N} X^\top X = S + \bar{x}\bar{x}^\top = \begin{pmatrix} 9 & 27 \\ 27 & 149 \end{pmatrix}$, whose top eigenvector makes an angle of $79.5°$ with the first axis, within about $1°$ of the direction of $\bar{x}$, while the data's spread is along the $45°$ diagonal. The first component points at the mean, not along the data. On the unshifted points the error hides, because $\bar{x} = (2, 2)^\top$ happens to lie along q_1.

2. Principal directions read from U instead of V

The SVD has two sets of singular vectors, and the data matrix can be written either way round.

1	$X_c = U \Sigma V^\top$ with $U \in \mathbb{R}^{N \times d}$ and $V \in \mathbb{R}^{d \times d}$	Right so far: Problem 18.8.

2	"The first singular vector is the first principal component."	The shortcut that causes the mistake: u_1 has one entry per example and v_1 one per feature, and a direction in the data space is a vector over the features.
3	$q_1 = u_1$ ✗	$u_1 = X_c v_1/\sigma_1$ is the first column of scores scaled to unit length (Problem 18.8, $Z = U\Sigma$): a pattern over the N examples, not a direction in $\mathbb{R}^d$. The shapes only agree when $N = d$, which is when the error produces no exception. With examples as columns, as some texts write, the roles swap and the directions are the left singular vectors, so what has to be checked is which side the examples are on; Chapter 12's mistake of building U from the eigenvectors of $A^\top A$ is the same confusion from the other side.

3. Explained variance from singular values instead of their squares

Singular values are what the SVD prints, and they decrease like the eigenvalues do.

1	$\lambda_j = \sigma_j^2/N$	Right so far: Problem 18.8.
2	"The share of component 1 is $\sigma_1/\sum_j \sigma_j$."	The shortcut that causes the mistake: the share is a ratio of variances, and a variance is a squared singular value.
3	Share of component $1 = \dfrac{\sigma_1}{\sigma_1+\sigma_2} = \dfrac{4\sqrt{2}}{6\sqrt{2}} = \dfrac{2}{3}$ ✗	The variance along q_1 is σ_1^2/N, so the share is $\sigma_1^2/\sum_j \sigma_j^2 = 32/40 = 0.8$ (Problem 18.4). The unsquared ratio understates dominant components and overstates weak ones, so a threshold such as 95% applied to it keeps too many components. `explained_variance_ratio_` in scikit-learn uses the squares.

4. Reconstruction without the mean added back

The scores are computed from centered data, and decoding looks like the transpose of encoding.

1	$z_n = Q_k^\top (x_n - \bar{x})$	Right so far: Problem 18.5.
2	"Decode by applying Q_k: $\hat{x}_n = Q_k z_n$."	The shortcut that causes the mistake: the encoder subtracted $\bar{x}$, so $Q_k z_n$ reconstructs the centered point and the decoder has to add $\bar{x}$ back.

3	$\hat{x}_n = Q_k Q_k^\top (x_n - \bar{x})$ ✗	For the four points this gives $(2,2)$, $(2,2)$, $(-2,-2)$, $(-2,-2)$ in place of $(4,4)$, $(4,4)$, $(0,0)$, $(0,0)$: every reconstruction is off by $\bar{x}$. The mean squared error becomes $\sum_{j>k} \lambda_j + \|\bar{x}\|^2 = 2 + 8 = 10$ instead of 2, since the cross term $\bar{x}^\top (I - Q_k Q_k^\top)(x_n - \bar{x})$ averages to zero over the centered points. In matrix form the reconstruction is $\mathbf{1}\bar{x}^\top + X_c Q_k Q_k^\top$, and `inverse_transform` adds the mean for exactly this reason.

5. Taking the first column of the eigenvector matrix as the top component

`np.linalg.eigh` returns eigenvalues in ascending order, and the columns of the eigenvector matrix follow the same order.

1	$S = Q \Lambda Q^\top$, computed as `lam, Q = np.linalg.eigh(S)`	Right so far: the eigen-decomposition of Problem 18.2.
2	"The first column of Q is the first principal component."	The habit that causes the mistake: the notation q_1 with $\lambda_1 \geq \lambda_2$, while the routine sorts the other way.
3	q_1 = `Q[:, 0]` ✗	For the four points `eigh` returns $\lambda = (2, 8)$ and `Q[:, 0]` $= \pm\frac{1}{\sqrt{2}}(1, -1)^\top$, the direction of least variance: the chosen component explains 20% of the variance and the reconstruction error is 8 rather than 2 (Problem 18.6). The columns have to be reversed, `Q[:, ::-1]` with `lam[::-1]`, or sorted by λ descending; `np.linalg.svd` returns singular values in descending order, which is one more reason to take Problem 18.8's route.

THE FULL WORKBOOK

Linear Algebra for Machine Learning by Hand

250 problems from matrix products and tensor shapes to the SVD, PCA, attention and LoRA, with every step worked

- 250 problems in 25 chapters: vectors, products and shapes, systems, subspaces and maps, orthogonality and least squares, eigenvalues and the svd and linear algebra inside models
- 16 chapters written only for the workbook
- A worked solution for every problem, one step per line with the reason beside it
- An answer key per chapter, and the common mistakes for every topic
- Two PDFs, *Problems* and *Worked Solutions*, US Letter, ready to print

Get it at eulerpad.app/workbooks/