Scientific Machine Learning 15
Dimensionality Reduction with Principal Component Analysis (PCA)

Donghyun Ko

January 4, 2026

What you will learn. Many variables in a dataset are redundant or noisy, and a simple rotation of axes can remove linear correlations while ranking directions by their variability. You'll see two equivalent ways to get this rotation—diagonalizing the covariance (eigen route) or using the SVD of the data. You'll learn to read *scores* (uncorrelated sample coordinates) and *loadings* (principal directions), and to choose how many components to keep using explained variance (e.g., 95%). We finish with when to use extensions like Kernel/Nonlinear PCA, Functional PCA, Robust PCA, Sparse PCA, and Probabilistic PCA.

Contents

1 Introduction

1.1 Motivation and objectives

Real datasets often have redundant or highly correlated variables which usually result in high dimension; some directions are rich in signal while others are dominated by noise. PCA re-expresses the data in a new orthonormal basis that (i) removes linear correlations and (ii) orders coordinates by variance so we can discard the least informative ones to reduce the dimension containing significantly signal-embedded variables only.

Objective of PCA. Given centered data $X \in \mathbb{R}^{m\times n}$ (rows: variables, cols: samples), find an orthonormal $P \in \mathbb{R}^{m\times m}$ such that

$$Y \;=\; PX, \quad \text{where} \quad C_Y \;=\; \frac{1}{n}YY^\top \text{ is diagonal and variance-ordered.}$$

The k-th row $p_k^\top$ of P is the k-th principal component (PC); the columns of Y are the corresponding *scores*. The rest of this posting will tell you details about this.

1.2 Geometric Intuition

Think of the data as a cloud of points in $\mathbb{R}^m$. An orthonormal matrix P simply *rotates* (or reflects) the coordinate axes—no stretching, no squeezing. Distances and angles are preserved:

$$\|x\|_2 \;=\; \|Px\|_2 \quad \text{for all } x \in \mathbb{R}^m,$$

so geometry is unchanged; only our *view* of the cloud changes.

Shadows and spread. Projecting points onto a unit direction 'u' produces 1D "shadows" whose spread equals the variance such that $\mathrm{Var}(u^\top X) = u^\top C_X u$ from (**??**). If we rotate the axes so that one axis *aligns* with the direction of greatest spread, that axis captures the longest shadow (largest variance). PCA chooses axes so that:

- The first axis (PC1) points along the direction of *maximal* variance;
- PC2 is orthogonal to PC1 and captures the next-largest variance; and so on.

Geometrically, many real data clouds look like a tilted *ellipsoid*. PCA aligns the axes with the ellipsoid's principal axes and orders them by the sizes of their variances.

Coordinates in the PC basis. Let $P = [p_1^\top; \ldots; p_m^\top]$ collect the PCs as rows. Each centered sample x has PC coordinates (scores) such that:

$$y \;=\; Px, \qquad y_k \;=\; p_k^\top x \quad (k = 1, \ldots, m),$$

and we can reconstruct x exactly by

$$x \;=\; P^\top y \;=\; \sum_{k=1}^{m} y_k\, p_k.$$

Theorem (Parseval in the PC basis & variance of PC coordinates). Let $P \in \mathbb{R}^{m\times m}$ be orthonormal ($PP^\top = I$), $y = Px$ for $x \in \mathbb{R}^m$, and X be a centered data matrix with covariance $C_X = \frac{1}{n}XX^\top$. If P is chosen as the PCA rotation (rows are eigenvectors of C_X), then:

(a) (*Parseval / Energy preservation*)

$$\|x\|_2^2 \;=\; \sum_{k=1}^{m} y_k^2.$$

(b) (*Variance of PC coordinates*) Across the dataset, the variance of each coordinate y_k equals the k-th eigenvalue λ_k because

$$C_Y = \tfrac{1}{n}YY^\top = PC_XP^\top = \mathrm{diag}(\lambda_1, \ldots, \lambda_m) \quad \Rightarrow \quad \mathrm{Var}(y_k) = \lambda_k.$$

Proof.
(a) Since P is orthonormal, $P^\top P = I$. Hence,

$$\|x\|_2^2 \;=\; x^\top x \;=\; x^\top (P^\top P)x \;=\; (Px)^\top (Px) \;=\; y^\top y \;=\; \sum_{k=1}^{m} y_k^2$$

(b) Write the sample covariance in the PC basis:

$$C_Y \;=\; \frac{1}{n}YY^\top \;=\; \frac{1}{n}(PX)(PX)^\top \;=\; P\Big(\frac{1}{n}XX^\top\Big)P^\top \;=\; PC_XP^\top. \tag{1}$$

If rows of P are eigenvectors of C_X, then $PC_XP^\top = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$. The diagonal entry (k, k) of C_Y equals

$$(C_Y)_{kk} \;=\; \frac{1}{n}\,(y_k y_k^\top) \;=\; \mathrm{Var}(y_k),$$

because X (hence Y) is centered across samples. From (1), we therefore get $(C_Y)_{kk} = \lambda_k$.

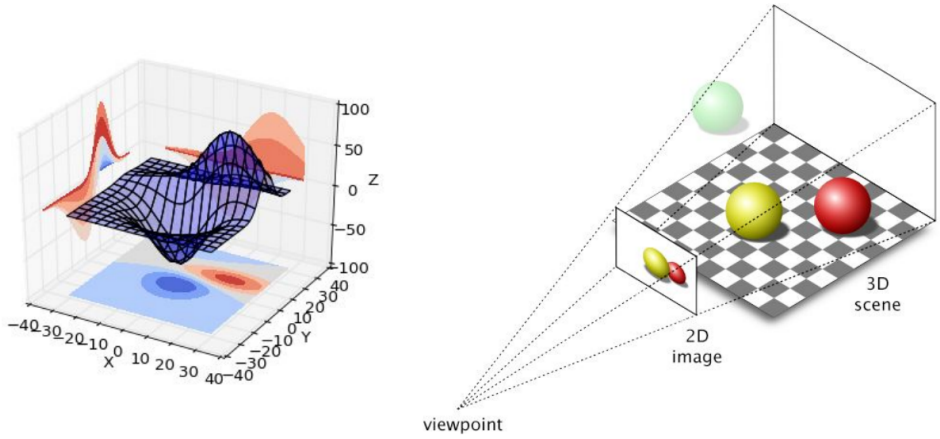
Why truncation works. If we keep only the first $m^\star$ axes (the longest radii) and drop the rest, we approximate x by

$$\hat{x} \;=\; \sum_{k=1}^{m^\star} y_k\, p_k \;=\; P_{m^\star}^\top P_{m^\star}\, x,$$

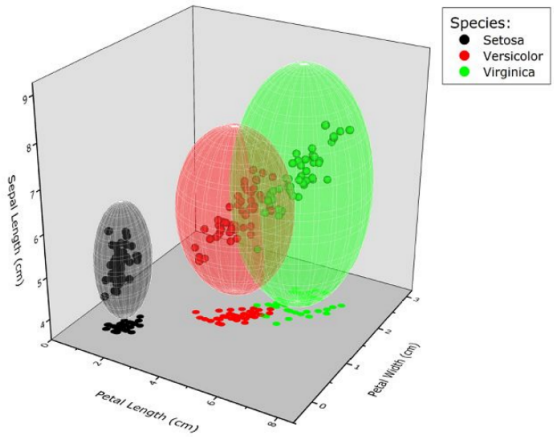
and the (squared) reconstruction error is the energy left in the discarded directions:

$$\|x - \hat{x}\|_2^2 \;=\; \sum_{k=m^\star+1}^{m} y_k^2.$$

Projecting Data - 3D to 2D



Projecting Data - 3D to 2D



2 Background

2.1 Eigendecomposition

Definition. For a square matrix $A \in \mathbb{R}^{n\times n}$ that has n linearly independent eigenvectors $\{v_1, \ldots, v_n\}$ and corresponding eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$, define

$$P = [v_1 \ \cdots \ v_n], \quad D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n).$$

Then, $AP = PD \iff A = PDP^{-1}$
If A is symmetric, its eigenvectors are orthogonal, so P can be chosen orthonormal:

$$P^{-1} = P^\top, \qquad A = PDP^\top.$$

Proof. Suppose $A \in \mathbb{R}^{n\times n}$ is symmetric ($A = A^\top$) and let $Av_i = \lambda_i v_i$, $Av_j = \lambda_j v_j$. Then, $v_i^\top A v_j = v_i^\top(\lambda_j v_j) = \lambda_j v_i^\top v_j$, and $v_i^\top A v_j = (Av_i)^\top v_j = (\lambda_i v_i)^\top v_j = \lambda_i v_i^\top v_j$. Subtracting gives $(\lambda_i - \lambda_j)v_i^\top v_j = 0$. If $\lambda_i \neq \lambda_j$, then $v_i^\top v_j = 0$, proving orthogonality of distinct eigenvectors. Normalizing each v_i gives an orthonormal basis, so $P^\top P = I$ and $A = PDP^\top$.

Example. With $A = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$, we seek the eigenvalues λ and the eigenvectors $v \neq 0$ that satisfy $Av = \lambda v$.

Step 1. Characteristic equation.

$$\det(A - \lambda I) = \begin{vmatrix} 3-\lambda & 1 \\ 1 & 3-\lambda \end{vmatrix} = (3-\lambda)^2 - 1 = 0 \Rightarrow 3 - \lambda = \pm 1.$$

Hence,

$$\lambda_1 = 4, \qquad \lambda_2 = 2.$$

Step 2. Solve $(A - \lambda I)v = 0$**.**
(a) For $\lambda_1 = 4$:

$$(A - 4I) = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}, \quad \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0.$$

From the first row: $-v_1 + v_2 = 0 \Rightarrow v_2 = v_1$. This tells us that both coordinates increase or decrease together — the direction lies along the line $x = y$. Any nonzero multiple of $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ satisfies this, so

$$v^{(4)} \propto \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

(b) For $\lambda_2 = 2$:

$$(A - 2I) = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0.$$

Reducing the system gives $v_1 + v_2 = 0$, or $v_2 = -v_1$. Now, one coordinate increases while the other decreases — the direction lies along the line $x = -y$. Hence,

$$v^{(2)} \propto \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Step 3. Normalize for an orthonormal basis. The length of each unnormalized eigenvector is

$$\|[1,1]^\top\| = \sqrt{2}, \quad \|[1,-1]^\top\| = \sqrt{2}.$$

Normalize by dividing each by $\sqrt{2}$:

$$\hat{v}^{(4)} = \tfrac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \hat{v}^{(2)} = \tfrac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Normalization ensures $\|\hat{v}_i\| = 1$, and since A is symmetric, the eigenvectors are automatically orthogonal:

$$\hat{v}^{(4)\top} \hat{v}^{(2)} = (1/2)(1 \cdot 1 + 1 \cdot (-1)) = 0.$$

Step 4. Verification.

$$A\hat{v}^{(4)} = \frac{1}{\sqrt{2}} \begin{bmatrix} 4 \\ 4 \end{bmatrix} = 4\,\hat{v}^{(4)}, \qquad A\hat{v}^{(2)} = \frac{1}{\sqrt{2}} \begin{bmatrix} 2 \\ -2 \end{bmatrix} = 2\,\hat{v}^{(2)}.$$

Thus, both pairs $(\lambda_1, \hat{v}^{(4)})$ and $(\lambda_2, \hat{v}^{(2)})$ satisfy $Av = \lambda v$.

Step 5. Final decomposition.

$$P = \tfrac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad D = \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix}, \quad A = PDP^\top.$$

The matrix P rotates coordinates to align with A's principal axes: the first axis ($x = y$) corresponds to the larger eigenvalue $\lambda_1 = 4$ (greater variance or stretching), and the second axis ($x = -y$) corresponds to $\lambda_2 = 2$ (smaller variance or compression).

2.2 Singular Value Decomposition (SVD)

Eigen-decomposition decomposes a square matrix into eigenvectors and eigenvalues. SVD provides another way to factorize a general matrix, into singular vectors and singular values.

Key definitions and meanings.

- **Eigenvalue (λ):** A scalar indicating how much a square matrix A stretches or compresses a vector that points in a particular direction. If $Av = \lambda v$, then v keeps its direction and its length scales by λ. For symmetric A, all λ are real and correspond

to "principal directions" of variation.

- **Eigenvector (v):** A nonzero vector whose direction remains unchanged under the transformation A. Each eigenvector defines a line through the origin that is invariant under A; only its length changes by the factor λ.
- **Singular value (σ):** A nonnegative scalar measuring how much a general (not necessarily square) matrix A stretches vectors along certain orthogonal directions. It is the square root of an eigenvalue of $A^\top A$ or $AA^\top$. Large singular values indicate directions in which A has the greatest effect or variance.
- **Singular vector (u, v):** The unit vectors defining the input and output directions of A's action in the Singular Value Decomposition (SVD):

  $$Av_i = \sigma_i u_i, \qquad A^\top u_i = \sigma_i v_i.$$

  Here, v_i (right-singular vector) lies in the input space, and u_i (left-singular vector) lies in the output space. Together, they describe how A rotates and scales space along orthogonal axes.

Definition. Any $A \in \mathbb{R}^{m\times n}$ can be factorized as

$$A = U\Sigma V^\top, \text{ where}$$

- $U \in \mathbb{R}^{m\times m}$ is an orthogonal (orthonormal) matrix whose columns u_i satisfy $U^\top U = I_m$. The columns of U are called the *left-singular vectors* of A and form an orthonormal set of eigenvectors of $AA^\top$:

  $$AA^\top = U\,\Sigma^2\,U^\top$$

- $V \in \mathbb{R}^{n\times n}$ is an orthogonal matrix whose columns v_i satisfy $V^\top V = I_n$. The columns of V are the *right-singular vectors* of A and form an orthonormal set of eigenvectors of $A^\top A$:

  $$A^\top A = V\,\Sigma^2\,V^\top$$

- $\Sigma \in \mathbb{R}^{m\times n}$ is a (rectangular) diagonal matrix whose nonnegative diagonal entries

  $$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0 \quad \text{(arranged in descending order)}$$

  are called the *singular values* of A with $r = \text{rank}(A)$. They are the square roots of the nonzero eigenvalues of both $AA^\top$ and $A^\top A$. Large singular values indicate directions where A has the strongest effect (highest variance or energy).

- **Geometric meaning.** Each pair (u_i, v_i) satisfies

  $$Av_i = \sigma_i u_i, \qquad A^\top u_i = \sigma_i v_i,$$

  so A maps the input direction v_i to the output direction u_i, scaled by σ_i.

- **Ordering.** Singular values are ordered $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. For $r = \min(m, n)$, the *thin SVD* uses $U \in \mathbb{R}^{m\times r}$, $\Sigma \in \mathbb{R}^{r\times r}$, and $V \in \mathbb{R}^{n\times r}$ without zero padding.
- **Connection between SVD and Eigen-decomposition.** Given $A = U\Sigma V^\top$, multiply by its transpose:

$$AA^\top = (U\Sigma V^\top)(V\Sigma^\top U^\top) = U\Sigma^2 U^\top, \quad A^\top A = V\Sigma^2 V^\top.$$

Therefore, the columns of U are eigenvectors of $AA^\top$ and those of V are eigenvectors of $A^\top A$, both with eigenvalues σ_k^2. The decomposition always exists even when A is not square, extending Eigen-decomposition to general rectangular matrices.

Example. Given $A = \begin{bmatrix} 3 & 1 \\ 0 & 2 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{3\times 2}$, compute $A^\top A = \begin{bmatrix} 3 & 0 & 0 \\ 1 & 2 & 0 \end{bmatrix} \begin{bmatrix} 3 & 1 \\ 0 & 2 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 9 & 3 \\ 3 & 5 \end{bmatrix}$.

1) Eigenvalues of $A^\top A$ (then singular values). The characteristic polynomial is

$$\det \begin{bmatrix} 9-\lambda & 3 \\ 3 & 5-\lambda \end{bmatrix} = (9-\lambda)(5-\lambda) - 9 = \lambda^2 - 14\lambda + 36.$$

Hence,

$$\lambda = \frac{14 \pm \sqrt{14^2 - 4\cdot 36}}{2} = \frac{14 \pm \sqrt{52}}{2} = 7 \pm \sqrt{13}.$$

Therefore,

$$\lambda_1 = 7 + \sqrt{13} \approx 10.6055, \qquad \lambda_2 = 7 - \sqrt{13} \approx 3.3945,$$

and the singular values are

$$\sigma_1 = \sqrt{\lambda_1} = \sqrt{7 + \sqrt{13}} \approx 3.257, \qquad \sigma_2 = \sqrt{\lambda_2} = \sqrt{7 - \sqrt{13}} \approx 1.842.$$

2) Right singular vectors V (eigenvectors of $A^\top A$). Solve $(A^\top A - \lambda I)v = 0$.
For $\lambda_1 = 7 + \sqrt{13}$:

$$(A^\top A - \lambda_1 I) = \begin{bmatrix} 9-\lambda_1 & 3 \\ 3 & 5-\lambda_1 \end{bmatrix} = \begin{bmatrix} 2-\sqrt{13} & 3 \\ 3 & -2-\sqrt{13} \end{bmatrix}.$$

From the first row, $(2-\sqrt{13})\,v_1 + 3\,v_2 = 0 \Rightarrow v_1 = \dfrac{3}{\sqrt{13}-2}\,v_2$. A convenient (unnormalized) choice is

$$v_1^{(1)} \propto \begin{bmatrix} 3 \\ \sqrt{13}-2 \end{bmatrix}.$$

For $\lambda_2 = 7 - \sqrt{13}$:

$$(A^\top A - \lambda_2 I) = \begin{bmatrix} 2+\sqrt{13} & 3 \\ 3 & -2+\sqrt{13} \end{bmatrix}.$$

From the first row, $(2+\sqrt{13})\,v_1 + 3\,v_2 = 0 \Rightarrow v_1 = -\dfrac{3}{2+\sqrt{13}}\,v_2$. Take

$$v_1^{(2)} \propto \begin{bmatrix} 3 \\ -(2+\sqrt{13}) \end{bmatrix}.$$

Normalize to unit length (any overall sign is acceptable):

$$\tilde{v}_1 = \frac{1}{\sqrt{9+(\sqrt{13}-2)^2}} \begin{bmatrix} 3 \\ \sqrt{13}-2 \end{bmatrix} \approx \begin{bmatrix} 0.882 \\ 0.472 \end{bmatrix}, \qquad \tilde{v}_2 \approx \begin{bmatrix} 0.472 \\ -0.882 \end{bmatrix}.$$

Thus,

$$V = \begin{bmatrix} \tilde{v}_1 & \tilde{v}_2 \end{bmatrix} \approx \begin{bmatrix} 0.882 & 0.472 \\ 0.472 & -0.882 \end{bmatrix}, \quad V^\top V = I_2.$$

3) Left singular vectors U via $U = AV\Sigma^{-1}$. Compute AV first, then scale by $1/\sigma_i$. First column:

$$A\,\tilde{v}_1 = \begin{bmatrix} 3 & 1 \\ 0 & 2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0.882 \\ 0.472 \end{bmatrix} = \begin{bmatrix} 3 \cdot 0.882 + 1 \cdot 0.472 \\ 2 \cdot 0.472 \\ 0 \end{bmatrix} \approx \begin{bmatrix} 3.118 \\ 0.944 \\ 0 \end{bmatrix} \Rightarrow u_1 = \frac{1}{\sigma_1} A\,\tilde{v}_1 \approx \begin{bmatrix} 0.957 \\ 0.290 \\ 0 \end{bmatrix}.$$

Second column:

$$A\,\tilde{v}_2 = \begin{bmatrix} 3 & 1 \\ 0 & 2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0.472 \\ -0.882 \end{bmatrix} \approx \begin{bmatrix} 0.534 \\ -1.764 \\ 0 \end{bmatrix} \Rightarrow u_2 = \frac{1}{\sigma_2} A\,\tilde{v}_2 \approx \begin{bmatrix} 0.290 \\ -0.958 \\ 0 \end{bmatrix}.$$

These columns are (numerically) orthonormal and lie in the x–y plane. To complete $U \in \mathbb{R}^{3\times 3}$, take any unit vector orthogonal to $\{u_1, u_2\}$, e.g.

$$u_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Thus,

$$U = \begin{bmatrix} u_1 & u_2 & u_3 \end{bmatrix} \approx \begin{bmatrix} 0.957 & 0.290 & 0 \\ 0.290 & -0.958 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad U^\top U = I_3.$$

4) Assemble the (thin) SVD. With

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 3.257 & 0 \\ 0 & 1.842 \\ 0 & 0 \end{bmatrix},$$

Finally, we obtain

$$A = U\Sigma V^\top, \quad AA^\top = U\Sigma^2 U^\top, \quad A^\top A = V\Sigma^2 V^\top.$$

2.3 Variance and Covariance

Let $a = [a_1, \ldots, a_n]^\top$ be 'n' samples of a random variable A with sample mean

$$\mu_A = \frac{1}{n}\sum_{i=1}^{n} a_i,$$

and the sample variance (not an estimate)

$$\mathrm{Var}(A) = \frac{1}{n}\sum_{i=1}^{n}(a_i - \mu_A)^2 = \frac{1}{n}(a - \mu_A\mathbf{1})^\top(a - \mu_A\mathbf{1}).$$

If centered ($\mu_A = 0$):

$$\mathrm{Var}(A) = \frac{1}{n}a^\top a.$$

For two centered variables A, B with samples $a, b \in \mathbb{R}^n$:

$$\mathrm{Cov}(A, B) = \frac{1}{n}\sum_{i=1}^{n} a_i b_i = \frac{1}{n}a^\top b.$$

Stacking m centered variables as rows of $X \in \mathbb{R}^{m\times n}$ gives the covariance matrix

$$C_X = \frac{1}{n}XX^\top,$$

whose (j, k) entry is $\mathrm{Cov}(X_{j\cdot}, X_{k\cdot})$ and diagonal entries are variances.

Example. Suppose $X = \begin{bmatrix} 2 & 0 & -2 \\ 0 & 1 & -1 \end{bmatrix}$ (each row centered). Then,

$$C_X = \tfrac{1}{3}XX^\top = \tfrac{1}{3}\begin{bmatrix} 8 & 0 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 8/3 & 0 \\ 0 & 2/3 \end{bmatrix}.$$

Hence,

$$\mathrm{Var}(X_1) = \tfrac{8}{3}, \quad \mathrm{Var}(X_2) = \tfrac{2}{3}, \quad \mathrm{Cov}(X_1, X_2) = 0.$$

Diagonal C_X indicates uncorrelated variables that is the exact situation PCA seeks after rotation.

2.4 Orthogonal changes of basis

Let $P \in \mathbb{R}^{m\times m}$ be orthonormal ($PP^\top = I$). Then,

$$Y \;=\; PX, \qquad C_Y \;=\; \frac{1}{n}YY^\top \;=\; P\Big(\frac{1}{n}XX^\top\Big)P^\top \;=\; PC_XP^\top.$$

Hence, C_Y is similar to C_X; choosing P that diagonalizes C_X produces uncorrelated coordinates directly.

3 Principal Component Analysis

3.1 What PCA is and why we need it

PCA is a statistical procedure that applies an *orthogonal transformation* to convert possibly correlated variables into a set of *linearly uncorrelated* variables called *principal components (PCs)*. The PCs are mutually orthogonal. By keeping only the leading PCs, one achieves *dimension reduction* while preserving as much of the data variability as possible. Let $X \in \mathbb{R}^{m\times n}$ be the (centered) data matrix with

- m variables (rows), each row containing its n samples,
- n observations (columns), each column listing all m variables for one measurement.

We will transform X by a square matrix $P \in \mathbb{R}^{m\times m}$ such that:

$$Y \;=\; PX$$

and choose P so that Y has the *desirable* covariance structure described in (3) below.

Change of basis. Define the transformed data $Y \in \mathbb{R}^{m\times n}$ from $X \in \mathbb{R}^{m\times n}$ by $P \in \mathbb{R}^{m\times m}$:

$$Y \;=\; PX.$$

Write $P = [p_1, \ldots, p_m]^\top$ where $p_k^\top$ are the *rows* of P, and $X = [x_1, \ldots, x_n]$ with x_i the *columns* (one sample per column). Then, $Y = [y_1, \ldots, y_n]$ with

$$y_i \;=\; \begin{bmatrix} p_1 \cdot x_i \\ p_2 \cdot x_i \\ \vdots \\ p_m \cdot x_i \end{bmatrix} \quad\Longleftrightarrow\quad PX = \begin{bmatrix} — \; p_1 \; — \\ — \; p_2 \; — \\ \vdots \\ — \; p_m \; — \end{bmatrix} \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} p_1 \cdot x_1 & p_1 \cdot x_2 & \cdots & p_1 \cdot x_n \\ p_2 \cdot x_1 & p_2 \cdot x_2 & \cdots & p_2 \cdot x_n \\ \vdots & \vdots & \ddots & \vdots \\ p_m \cdot x_1 & p_m \cdot x_2 & \cdots & p_m \cdot x_n \end{bmatrix}$$

Geometrically, P is a rotation and a stretch that transform X into Y. In other words, P re-expresses each sample x_i in a new coordinate system whose axes are the row vectors p_k:

- If P is **orthonormal** ($PP^\top = I$), the transform is a rotation/reflection while preserving length and angle.
- For a general invertible P, it can include rotations, axis scalings (stretches), and shears.

The operation above is essentially a change of original basis to the new basis defined by P. In PCA, we will *enforce* P to be orthonormal so that the covariance transforms by similarity:

$$C_Y \;=\; \tfrac{1}{n} YY^\top \;=\; P\Big(\tfrac{1}{n} XX^\top\Big)P^\top \;=\; PC_XP^\top,$$

enabling us to choose P that *diagonalizes* C_X.

Quick navigation!

(a) *Why* perform the linear transformation (change of basis)?
To reduce redundancy/noise and reveal dominant directions (by variance)

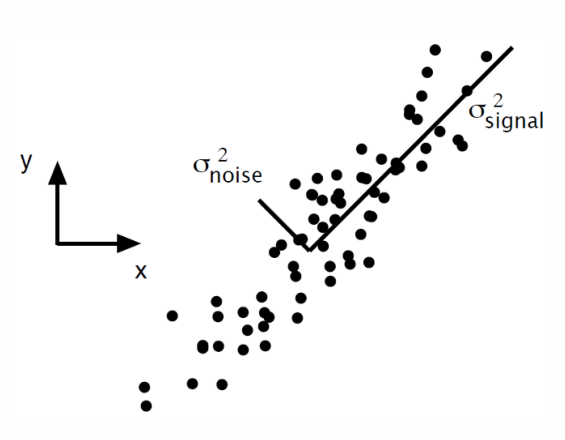
(b) *What* property do we want in Y?
A diagonal, ordered covariance $C_Y = \text{diag}(\lambda_1 \geq \cdots \geq \lambda_m)$.

(c) *How* to choose P to get that property?
Take $P = Q^\top$ from the Eigen-decomposition $C_X = QDQ^\top$, or equivalently $P = U^\top$ from the SVD with $X = U\Sigma V^\top$.

3.2 Noise, redundancy, and the covariance goal

Real measurement data are rarely perfect. The data matrix X may contain *redundancy* (high correlations between variables) and *noise* (random fluctuations). Assuming reasonably good measurements, we typically claim:

- Directions with **large variance** in measurement space carry the dominant dynamics or signal of interest.
- Directions with **small variance** are often dominated by measurement noise.



There is no absolute reference scale for "noise." Instead, it is defined relative to the signal strength, commonly through the *signal-to-noise ratio (SNR)*:

$$\text{SNR} = \frac{\sigma^2_{\text{signal}}}{\sigma^2_{\text{noise}}},$$

where σ^2_{signal} and σ^2_{noise} are variances of the signal and noise components, respectively. Hence, identifying directions with high variance is equivalent to finding a transformation (change of basis) that maximizes SNR and isolates meaningful dynamics.

Redundancy and correlation. Redundancy arises when two or more measured variables are linearly dependent or nearly so. In extreme cases—such as measuring the same length in meters and inches—the second variable does not contribute new information.

Let two variables A and B have centered sample vectors s.t:

$$a = [a_1, \ldots, a_n], \qquad b = [b_1, \ldots, b_n].$$

Their variances and covariance are

$$\mathrm{Var}(a) = \tfrac{1}{n} a^\top a, \quad \mathrm{Var}(b) = \tfrac{1}{n} b^\top b, \quad \mathrm{Cov}(a, b) = \tfrac{1}{n} a^\top b. \tag{2}$$

The degree of linear relationship between A and B is measured by covariance. The absolute magnitude of the correlation coefficient measures the degree of redundancy. The absolute correlation

$$|\rho_{ab}| = \Big| \frac{\mathrm{Cov}(a, b)}{\sqrt{\mathrm{Var}(a)\mathrm{Var}(b)}} \Big|$$

measures the degree of redundancy: if $|\rho_{ab}| = 1$, the two variables are perfectly correlated and one is redundant.

From two variables to many. Generalizing (2) to all m variables arranged in centered $X \in \mathbb{R}^{m \times n}$, we obtain the sample covariance matrix:

$$Cov(X) = C_X \;=\; \frac{1}{n} X X^\top \;\in\; \mathbb{R}^{m \times m}.$$

Each element has clear meaning:

- $(C_X)_{ij} = \frac{1}{n} x_i^\top x_j$ is the dot product between the ith and jth variable vectors.
- Diagonal entries $(C_X)_{ii}$ represent variances of individual variables.
- Off-diagonal entries $(C_X)_{ij}$ represent covariances (redundancies) between distinct variables.

Thus, C_X encodes both noise and redundancy information in X:

- Large diagonal entries $\Rightarrow$ large variances $\Rightarrow$ informative directions. (Keep them!)
- Large off-diagonal entries $\Rightarrow$ strong correlations $\Rightarrow$ redundant info. (Remove them!)

Covariance after transformation. For the transformed data $Y = PX$, its covariance is

$$C_Y \;=\; \frac{1}{n} YY^\top \;=\; P\Big(\tfrac{1}{n} XX^\top\Big)P^\top \;=\; P\,C_X\,P^\top,$$

i.e., C_Y is *similar* to C_X under the (orthonormal) change of basis P. We seek an orthonormal P such that C_Y has the following desirable properties:

- **Diagonal (no redundancy):** all off–diagonal entries of C_Y vanish (they are all zero), so different coordinates (rows of Y) are uncorrelated.
- **Ordered (variance ranking):** the diagonal entries are sorted in descending order,

  $$C_Y \;=\; \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m), \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq 0. \tag{3}$$

  so small–variance coordinates can be treated as noise and safely truncated for dimensionality reduction.

How to choose P (principal components). We want an orthonormal change of basis $P \in \mathbb{R}^{m\times m}$ whose *rows*

$$P = \begin{bmatrix} p_1^\top \\ p_2^\top \\ \vdots \\ p_m^\top \end{bmatrix}$$

define new axes (principal components, PCs). The variance associated with each direction p_k quantifies how "principal" that direction is; ranking these variances orders the coordinates by importance. There are two equivalent ways to obtain such an orthonormal P:

(a) **Eigen-decomposition of the covariance.** Compute the covariance $C_X = \frac{1}{n} XX^\top$ and diagonalize

$$C_X \;=\; QDQ^\top, \quad \text{where} \quad D = \mathrm{diag}(\lambda_1 \geq \cdots \geq \lambda_m \geq 0).$$

Choose

$$P \;=\; Q^\top \quad\Longrightarrow\quad C_Y \;=\; PC_XP^\top \;=\; D,$$

so the rows of P (the eigenvectors of C_X) are the PCs, and the diagonal entries of D are their variances.

(b) **Singular Value Decomposition (SVD) of the data.** Compute the SVD

$$X \;=\; U\Sigma V^\top,$$

then set

$$P \;=\; U^\top, \qquad Y = PX = \Sigma V^\top, \quad\Longrightarrow\quad C_Y \;=\; \tfrac{1}{n} YY^\top \;=\; \tfrac{1}{n}\Sigma^2.$$

The left-singular vectors U give the PC directions, and the singular values satisfy $\lambda_k = \frac{\sigma_k^2}{n}$.

In both ways, P is orthonormal, its rows are the PCs, and the variances (either λ_k or σ_k^2/n) *rank-order* directions such that small-variance coordinates can be truncated for dimensionality reduction. The rigourous derivations for these two ways will be covered in the following subsections.

3.3 PCA via Eigen-decomposition

For a *centered* data $X \in \mathbb{R}^{m\times n}$, find an orthonormal matrix $P \in \mathbb{R}^{m\times m}$ such that the transformed data $Y = PX$ has a *diagonal* covariance matrix C_Y. The rows of P form the *principal components* (PCs) of X — orthogonal directions along which the coordinates in Y are uncorrelated and ordered by variance.

Step 1: Covariance under an orthonormal change of basis. Recall the similarity relation for covariance:

$$C_Y \;=\; \frac{1}{n}YY^\top \;=\; P\Big(\frac{1}{n}XX^\top\Big)P^\top \;=\; P\,C_X\,P^\top, \tag{4}$$

where $C_X = \frac{1}{n}XX^\top$ is the sample covariance of X.

Lemma (Projected covariances in the P-basis). Let $P = [p_1^\top; \ldots; p_m^\top]$ with orthonormal rows ($PP^\top = I$) and $Y = PX$. Then, for $i \neq j$,

$$\mathrm{Cov}(y_i, y_j) = \frac{1}{n}(p_i^\top X)(p_j^\top X)^\top = p_i^\top C_X\, p_j, \quad \text{and} \quad \mathrm{Var}(y_k) = p_k^\top C_X\, p_k$$

Hence, C_Y is diagonal $\iff$ $p_i^\top C_X\, p_j = 0$ for all $i \neq j$.

Step 2: Spectral theorem for C_X. Since C_X is real symmetric positive semidefinite, the spectral theorem guarantees an eigen-decomposition):

$$C_X \;=\; QDQ^\top, \qquad Q^\top Q = I, \quad D = \mathrm{diag}(\lambda_1, \ldots, \lambda_m), \quad \lambda_1 \geq \cdots \geq \lambda_m \geq 0. \tag{5}$$

The columns of Q are orthonormal eigenvectors of C_X, and the diagonal entries of D are the corresponding eigenvalues.

Step 3: Choose P to diagonalize C_Y. Set

$$P \;=\; Q^\top. \tag{6}$$

Substitute (6) and (5) into (4):

$$C_Y \;=\; PC_XP^\top = Q^\top(QDQ^\top)Q = D, \tag{7}$$

which is diagonal with entries $\lambda_1, \ldots, \lambda_m$ in nonincreasing order. Therefore, *the principal directions are precisely the eigenvectors of* C_X (rows of P), and their variances in Y are the corresponding eigenvalues.

Step 4: What exactly becomes uncorrelated and why it matters. Let $Y = PX$ be with $P = Q^\top$. The k-th row of Y is $y_k = p_k^\top X$. From (7) and the lemma above,

$$\mathrm{Var}(y_k) = \lambda_k, \qquad \mathrm{Cov}(y_i, y_j) = 0 \;\; (i \neq j), \tag{8}$$

so each score coordinate y_k measures independent variation (no linear redundancy). Keeping the first $m^\star$ rows of P thus retains the directions of largest variance and discards low-variance directions that are typically noise-dominated.

3.4 PCA via SVD

PCA is closely related to the Singular Value Decomposition (SVD). Let the SVD of the (centered) data matrix $X \in \mathbb{R}^{m\times n}$ be

$$X \;=\; U\Sigma V^\top, \quad \text{where} \quad U \in \mathbb{R}^{m\times m},\; \Sigma \in \mathbb{R}^{m\times n},\; V \in \mathbb{R}^{n\times n}, \tag{9}$$

where U, V are orthonormal and Σ is (rectangular) diagonal with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ on its diagonal.

Key identities. Multiply (9) by $X^\top$ and X to obtain

$$XX^\top \;=\; U\Sigma^2 U^\top, \qquad X^\top X \;=\; V\Sigma^2 V^\top. \tag{10}$$

Hence, the columns of U are eigenvectors of $XX^\top$ (and of $C_X = \frac{1}{n}XX^\top$), with eigenvalues σ_k^2 (and $\lambda_k = \sigma_k^2/n$ for C_X).

Pick P from the SVD. Left-multiply (9) by $U^\top$:

$$U^\top X \;=\; U^\top(U\Sigma V^\top) \;=\; \Sigma V^\top. \tag{11}$$

Define the PCA transform by $P \;=\; U^\top$. Then,

$$Y \;=\; PX \;=\; U^\top X \;=\; \Sigma V^\top. \tag{12}$$

Thus, P uses the *left-singular vectors* of X as rows (PC directions), and Y are the *PC scores*.

Covariance of the scores is diagonal. Since X is centered, so is $Y = \Sigma V^\top$. Therefore,

$$C_Y \;=\; \frac{1}{n}YY^\top \;=\; \frac{1}{n}(\Sigma V^\top)(V\Sigma^\top) \;=\; \frac{1}{n}\,\Sigma^2, \tag{13}$$

which is diagonal and ordered when $\sigma_1 \geq \sigma_2 \geq \cdots$. Consequently, the score variances are

$$\mathrm{Var}(y_k) \;=\; \lambda_k \;=\; \frac{\sigma_k^2}{n},$$

matching the eigenvalues of C_X.

Remarks and practical notes.

- **The $1/n$ factor is harmless.** If you form the SVD of $\frac{1}{\sqrt{n}}X$ instead, the diagonal of C_Y becomes exactly the squared singular values of $\frac{1}{\sqrt{n}}X$ (no prefactor).
- **Thin SVD.** When $r = \mathrm{rank}(X) \leq \min(m, n)$, one can use the thin SVD $X = U_r\Sigma_r V_r^\top$; then $P = U_r^\top$ maps to r-dimensional scores $Y = \Sigma_r V_r^\top$ directly.
- **Computational tip.** For $m \gg n$ (many variables, few samples), computing V from $X^\top X$ can be cheaper, then recover U via $U = XV\Sigma^{-1}$.
- **Loadings and scores.** The rows of $P = U^\top$ (or columns of U) are unit-scaled *loadings*; variance-scaled loadings are $p_k\sqrt{\lambda_k}$. The columns of Y are the *PC scores*.

3.5 Explained variance, scores, loadings, and truncation

Let the PCs be the rows $p_k^\top$ of P and the transformed samples be the columns of $Y = PX$.

- **Scores.** The columns of $Y = PX$ are the *PC scores* (the coordinates of each sample in the PC basis). Because P is chosen from the eigenvectors (or left singular vectors), the *score variables* (rows of Y) are mutually uncorrelated: $\mathrm{Cov}(Y) = \frac{1}{n} YY^\top = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$.
- **Loadings.** How strongly each original variable contributes to a PC. Let $P = \begin{bmatrix} p_1^\top \\ \cdots \\ p_m^\top \end{bmatrix}$ with p_k the k-th PC (unit vector), and $C_Y = \frac{1}{n} YY^\top = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$.
  - *Unit-scaled* (p_k): shows direction (signs, relative weights) of variables in PC k.
  - *Variance-scaled* ($p_k\sqrt{\lambda_k}$): weights variables by variance explained by PC k.
  - *Correlation loadings:* $\mathrm{corr}(X_j, y_k) = \dfrac{p_{k,j}\sqrt{\lambda_k}}{\sqrt{\mathrm{Var}(X_j)}}$; if variables are standardized, this is $p_{k,j}\sqrt{\lambda_k}$.

Cumulative explained variance (CEV).

$$\mathrm{CEV}(m^\star) \;=\; \frac{\sum_{k=1}^{m^\star} \lambda_k}{\sum_{k=1}^{m} \lambda_k},$$

where λ_k are the diagonal entries of C_Y (variances along PCs, ordered $\lambda_1 \geq \cdots \geq \lambda_m$). Reduce dimensionality from m to $m^\star$ by keeping the first $m^\star$ *PCs* whose combined variance reaches a target (e.g., 95% or 99%). When variables have different units, standardize first (correlation-PCA), otherwise a few high-variance variables may dominate the CEV.

Orthogonal projection and reconstruction. Let $P_{m^\star}$ keep the first $m^\star$ rows of P. The orthogonal projector is $P_{m^\star}^\top P_{m^\star}$ and the best rank-$m^\star$ reconstruction is

$$\hat{X} \;=\; P_{m^\star}^\top P_{m^\star} X,$$

with optimal Frobenius error of

$$\|X - \hat{X}\|_F^2 \;=\; \sum_{k=m^\star+1}^{m} \lambda_k.$$

This equality indicates that the discarded variance corresponds exactly to the reconstruction error after PCA.

4 PCA Variations

4.1 Kernel, Functional, and Nonlinear PCA

Kernel PCA. Replace inner products by a positive-definite kernel $k(x, x')$. Form the Gram matrix $K \in \mathbb{R}^{n\times n}$ with $K_{ij} = k(x_i, x_j)$ and center it via

$$K_c = HKH, \qquad H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top.$$

Solve the eigenproblem $K_c\alpha = \tilde{\lambda}\alpha$, and normalize eigenvectors so that projected coordinates have unit variance in the implicit feature space (e.g., scale α by $1/\sqrt{\tilde{\lambda}}$ as needed). Out-of-sample mapping for a new x uses kernel evaluations $k(x, x_i)$ to obtain scores. Kernel PCA captures nonlinear manifolds through implicit high-dimensional feature mappings.

Functional PCA (FPCA). Extends PCA to *functional data*, where each observation $x_i(t)$ is a continuous curve or trajectory. Instead of the discrete covariance matrix, FPCA diagonalizes the covariance operator

$$C(s, t) = \mathrm{Cov}[x(s), x(t)],$$

yielding eigenfunctions $\phi_k(t)$ satisfying

$$\int C(s, t)\, \phi_k(s)\, ds = \lambda_k\, \phi_k(t).$$

Each observation can be represented using *functional scores*

$$y_{ik} = \int (x_i(t) - \mu(t))\, \phi_k(t)\, dt.$$

FPCA is powerful for time-series, spectra, or spatial-temporal data where smoothness and functional continuity are important.

Nonlinear PCA (Autoencoder). Train an encoder–decoder pair (f_θ, g_ϕ) to minimize the reconstruction loss

$$\min_{\theta,\phi}\ \frac{1}{n}\sum_{i=1}^{n} \|x_i - g_\phi(f_\theta(x_i))\|_2^2.$$

Around the data manifold, the encoder's local Jacobian often aligns with PCA directions. Unlike linear PCA, autoencoders learn nonlinear transformations, making them suitable for data that lie on curved manifolds.

4.2 Probabilistic, Robust, and Sparse PCA

Probabilistic PCA (PPCA). Model data as $x = Wz+\mu+\varepsilon$ with latent variable $z \sim \mathcal{N}(0, I)$ and noise $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. The maximum-likelihood subspace equals the PCA subspace. An EM algorithm can estimate W, σ^2 and provide uncertainty quantification.

Robust PCA. Decompose data as $X = L + S$ where L is low-rank (principal subspace) and S captures sparse outliers. This approach, often called *Principal Component Pursuit*, preserves the underlying structure under heavy-tailed noise or gross corruption.

Sparse PCA. Encourage sparsity in loading vectors (e.g., using ℓ_1 regularization) to enhance interpretability. Sparse PCA approximates the best low-rank reconstruction while retaining only the most influential variables.

When to prefer a variant.

- **Robust PCA:** Suitable for datasets with heavy-tailed noise or outliers.
- **Sparse PCA:** Useful when many correlated variables exist and interpretability is desired.
- **Kernel / Nonlinear PCA:** Best for data lying on curved or nonlinear manifolds.
- **Functional PCA:** Designed for functional, temporal, or continuous data such as signals or trajectories.
- **PPCA:** Appropriate for handling missing data or quantifying uncertainty in latent representations.

This is a posting that I summarized with study-purpose and is adapted from lecture notes of NE-795 (Scientific Machine Learning), given by Prof. Xu Wu, NC State University.