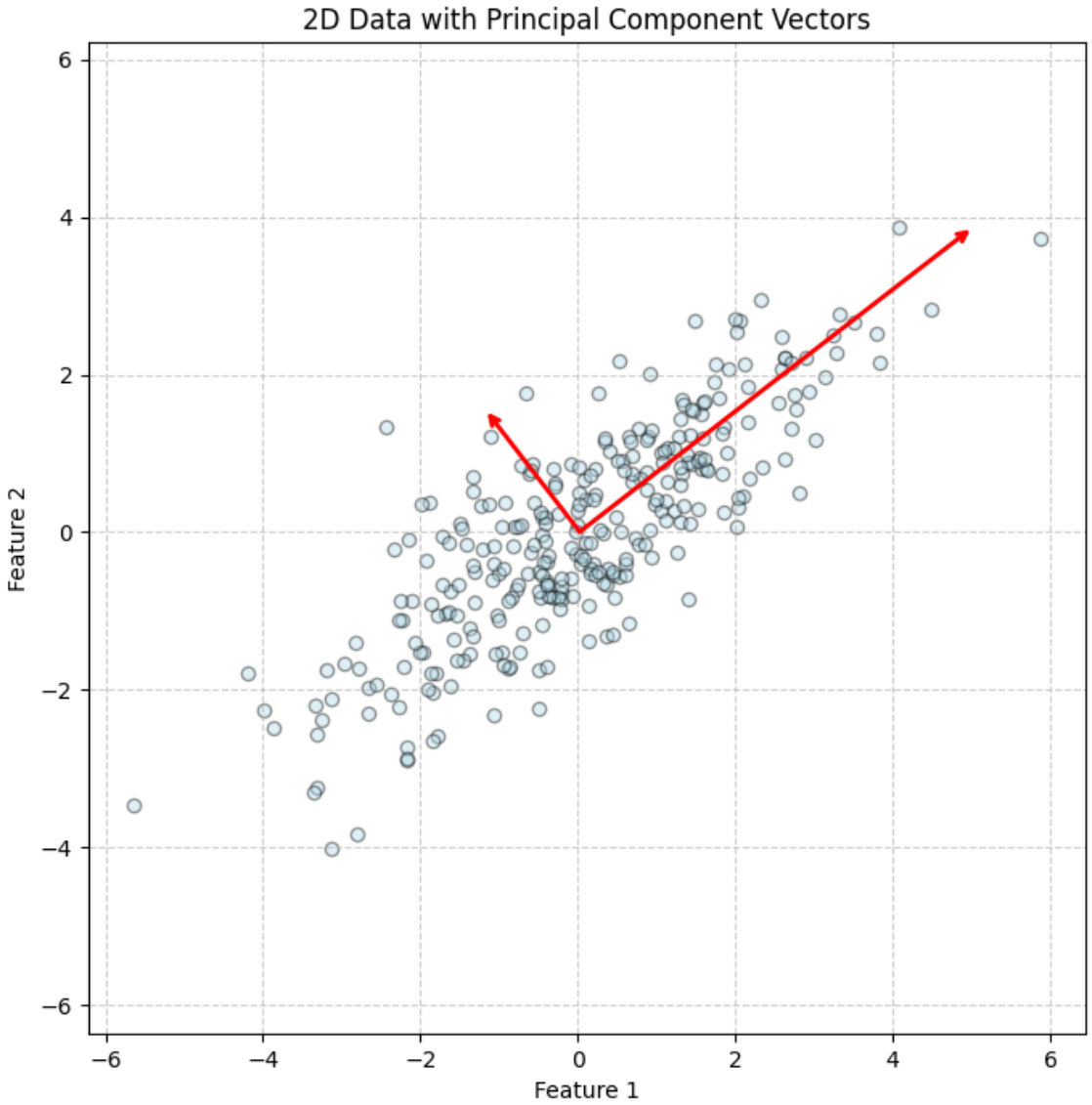
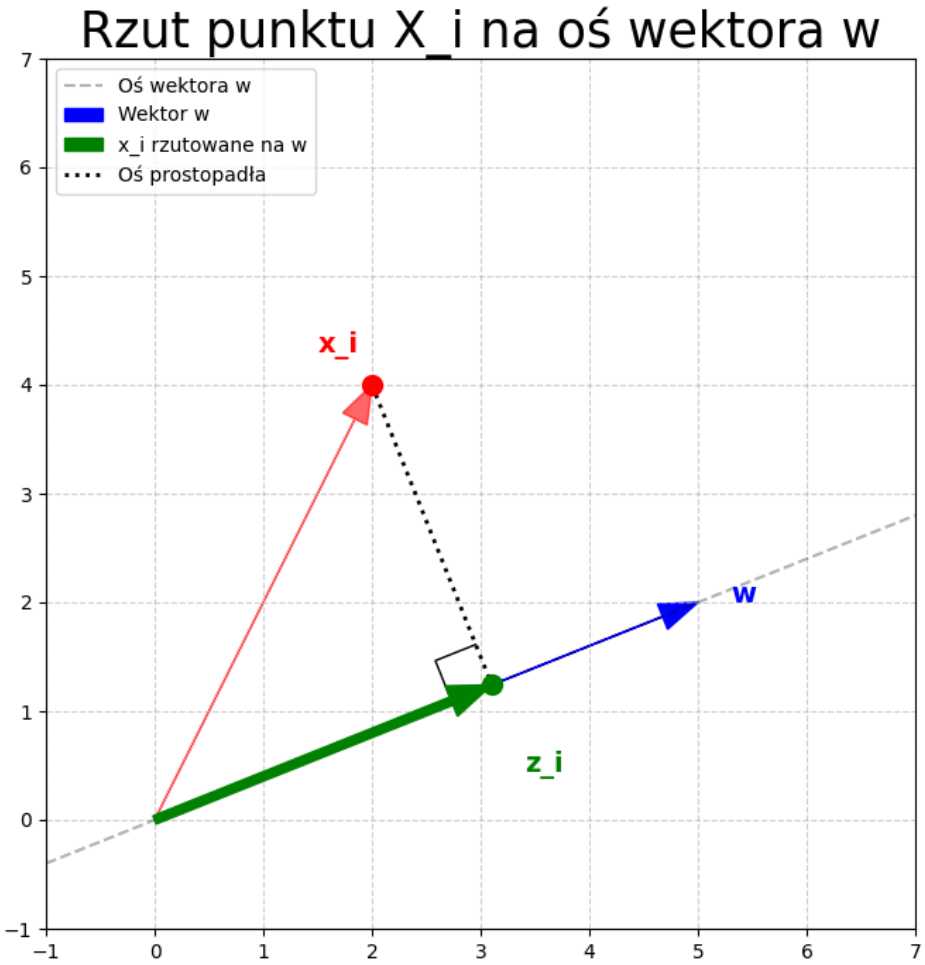
Principal Component Analysis - PCA

15 lutego 2026

Na początku zdefiniujemy, jak wyglądają nasze dane. Wektory X_i, o wymiarach $d \times 1$, reprezentują obserwacje o d cechach. Natomiast $\mathbb{X}$ $(n \times d)$ jest tablicą składającą się z transponowanych wektorów X_i.

$$X_i = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \qquad \mathbb{X} = \begin{bmatrix} — & X_1^T & — \\ — & X_2^T & — \\ & \vdots & \\ — & X_n^T & — \end{bmatrix}$$


2D Data with Principal Component Vectors
Feature 2
Feature 1
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6



$$||w|| = 1$$

$$z_i = \langle X_i, w \rangle = X_i^T w$$

Naszym zadaniem będzie znalezienie takiego wektora w, który zmaksymalizuje wariancję, gdy rzutujemy na niego punkty X_i. Przy tym założymy, że dane są znormalizowane (wszystkie cechy mieszczą się w przedziale $[0, 1]$) i mają średnią $\mu = 0$.

$$\text{wariancja} = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i^T w)^2 = \frac{1}{n}(\mathbb{X}w)^T \mathbb{X}w$$

$$= w^T \left(\frac{1}{n}\mathbb{X}^T\mathbb{X}\right) w = w^T S w$$

S jest macierzą kowariancji. Jest ona symetryczna, dlatego

$$v_1 \perp v_2 \perp \ldots \perp v_d, \qquad ||v_i|| = 1$$

$$\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_d, \qquad \lambda_i \in \mathbb{R}$$

Gdzie v_i są wektorami własnymi S, a λ_i odpowiadającymi im wartościami własnymi.

Wektory v_i tworzą przestrzeń ortonormalną, dlatego możemy zapisać wektor w jako kombinację liniową:

$$w = \sum_{i=1}^{d} c_i v_i$$

Korzystając z warunku długości wektora w, możemy ograniczyć wartości współczynników c_i

$$||w|| = 1 \implies \left(\sum_{i=1}^{d} c_i v_i\right)^T \sum_{i=1}^{d} c_i v_i = 1 \implies \sum_{i=1}^{d} c_i^2 = 1$$

Podstawiamy nowy wzór na w, pamiętając, że wektory własne spełniają równanie $Sv = \lambda v$.

$$Sw = S\left(\sum_{i=1}^{d} c_i v_i\right) = \sum_{i=1}^{d} \lambda_i c_i v_i$$

$$w^T S w = \left(\sum_{i=1}^{d} c_i v_i\right)^T \left(\sum_{i=1}^{d} \lambda_i c_i v_i\right) = \sum_{i=1}^{d} \lambda_i c_i^2$$

Co daje nam nowy wzór na wariancję:

$$\sigma^2 = \sum_{i=1}^{d} \lambda_i c_i^2$$

Przypomnijmy sobie, że naszym celem jest znalezienie takiego wektora w, który zmaksymalizuje wariancję, dlatego

$$\sup \sigma^2 = \sup_{c_i} \sum_{i=1}^{d} \lambda_i c_i^2 = \max\{\lambda_i; \quad i \in \{1, 2, \ldots, d\}\}$$

Z tego wynika, że $c_i = 1$, gdy $\lambda_i = \max\{\lambda\}$, w przeciwny razie $c_i = 0$.

Dzięki temu, wnioskujemy, że wektor w maksymalizujący wariancję, jest wektorem własnym macierzy S, odpowiadającym największej wartości własnej.

$$w = v_1 \quad \text{oraz} \quad \sigma^2 = \lambda_1$$