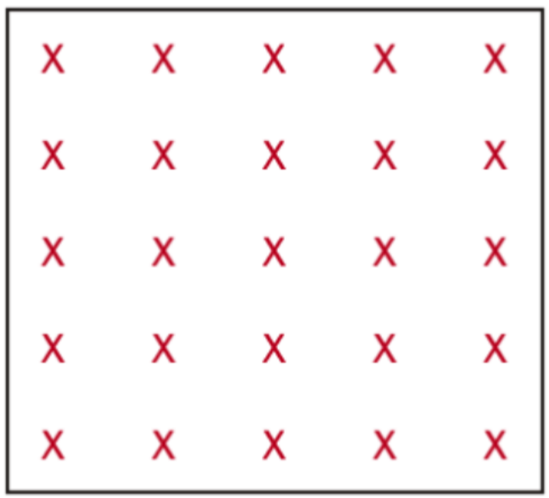
Dataset

- Features
- Observations

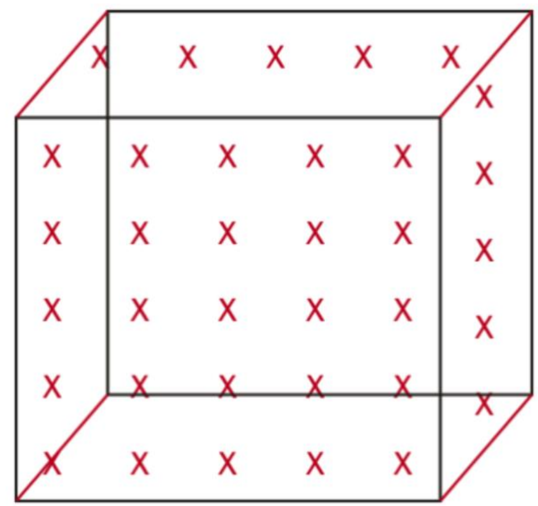
Dimension

Dimensions represent the total no. of features

1-D



2-D



3-D

Curse of Dimensionality

More dimensions don't always mean more information - sometimes they just mean more complexity.

Resulting in computational complexity and deteriorating predicting power

Solution:

Dimensionality Reduction

- the task of reducing the number of input features in a dataset, Two methods are available:

- **Feature Selection**

  - find a subset of the input features

Feature selection

e.g. LASSO

- **Feature Projection**

- find the projection of the original data into some low-dimensional space

Feature projection

e.g. PCA

Principal Component Analysis

To reduce the dimensionality, Principal Component Analysis (*PCA*) uses the projection of the

original data into the *principal components*.

The principal components describe the maximum amount of *variation* captured

`N` principal components (where `N < M, M is the number of features`), we move from the M-dimensional space to the N-dimensional space, where new features are combinations of the existing features.

Let’s consider a two dimensional dataset containing the height and weight of individual persons

Person	Height (cm)	Weight (kg)
A	150	52
B	170	55
C	155	58
D	162	62
E	191	68

Step 1: Standardize the Data

1.1 Compute the Mean and Standard Deviation for Each Feature

Height (cm):

- Mean of Height (X_1):

$$\bar{X}_1 = \frac{150 + 170 + 155 + 162 + 191}{5} = \frac{828}{5} = 165.6$$

- Standard deviation of Height (X_1):

$$\sigma_{\text{height}} = \sqrt{\frac{(150 - 165.6)^2 + (170 - 165.6)^2 + (155 - 165.6)^2 + (162 - 165.6)^2 + (191 - 165.6)^2}{5}}$$

$$\sigma_{\text{height}} = \sqrt{\frac{(-15.6)^2 + (4.4)^2 + (-10.6)^2 + (-3.6)^2 + (25.4)^2}{5}}$$

$$\sigma_{\text{height}} = \sqrt{\frac{243.36 + 19.36 + 112.36 + 12.96 + 645.16}{5}} = \sqrt{\frac{1033.2}{5}} = \sqrt{206.64} \approx 14.4$$

Weight (kg):

- Mean of Weight (X_2):

$$\bar{X}_2 = \frac{52 + 55 + 58 + 62 + 68}{5} = \frac{295}{5} = 59$$

- Standard deviation of Weight (X_2):

$$\sigma_{\text{weight}} = \sqrt{\frac{(52 - 59)^2 + (55 - 59)^2 + (58 - 59)^2 + (62 - 59)^2 + (68 - 59)^2}{5}}$$

$$\sigma_{\text{weight}} = \sqrt{\frac{(-7)^2 + (-4)^2 + (-1)^2 + (3)^2 + (9)^2}{5}} = \sqrt{\frac{49 + 16 + 1 + 9 + 81}{5}} = \sqrt{\frac{156}{5}} = \sqrt{31.2} \approx 5.59$$

$$\text{Standardized Height} = \frac{\text{Height} - \bar{X}_1}{\sigma_{\text{height}}}$$

$$\text{Standardized Weight} = \frac{\text{Weight} - \bar{X}_2}{\sigma_{\text{weight}}}$$

- **For Person A:**
  - Standardized Height: $\frac{150-165.6}{14.4} = \frac{-15.6}{14.4} \approx -1.08$
  - Standardized Weight: $\frac{52-59}{5.59} = \frac{-7}{5.59} \approx -1.25$
- **For Person B:**
  - Standardized Height: $\frac{170-165.6}{14.4} = \frac{4.4}{14.4} \approx 0.31$
  - Standardized Weight: $\frac{55-59}{5.59} = \frac{-4}{5.59} \approx -0.72$
- **For Person C:**
  - Standardized Height: $\frac{155-165.6}{14.4} = \frac{-10.6}{14.4} \approx -0.74$
  - Standardized Weight: $\frac{58-59}{5.59} = \frac{-1}{5.59} \approx -0.18$
- **For Person D:**
  - Standardized Height: $\frac{162-165.6}{14.4} = \frac{-3.6}{14.4} \approx -0.25$
  - Standardized Weight: $\frac{62-59}{5.59} = \frac{3}{5.59} \approx 0.54$
- **For Person E:**
  - Standardized Height: $\frac{191-165.6}{14.4} = \frac{25.4}{14.4} \approx 1.76$
  - Standardized Weight: $\frac{68-59}{5.59} = \frac{9}{5.59}$ [illegible].61

Standardized Dataset:		
Person	**Standardized Height (X_1)**	**Standardized Weight (X_2)**
A	-1.08	-1.25
B	0.31	-0.72
C	-0.74	-0.18
D	-0.25	0.54
E	1.76	1.61

Original Dataset:					**Standardized Dataset:**		
Person	**Height (cm)**	**Weight (kg)**			**Person**	**Standardized Height (X_1)**	**Standardized Weight (X_2)**
A	150	52			A	-1.08	-1.25
B	170	55			B	0.31	-0.72
C	155	58			C	-0.74	-0.18
D	162	62			D	-0.25	0.54
E	191	68			E	1.76	1.61

Step 2: Calculate the Covariance Matrix

Covariance Matrix:

$$\mathrm{Cov}(X) = \begin{bmatrix} \mathrm{cov}(X_1, X_1) & \mathrm{cov}(X_1, X_2) \\ \mathrm{cov}(X_2, X_1) & \mathrm{cov}(X_2, X_2) \end{bmatrix}$$

A covariance matrix is a square matrix that describes the covariance between pairs of variables in a dataset. It provides a comprehensive view of how different variables change together.

Covariance

Indicates how variables change in relation to each other

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n}$$

Calculates the average of the product of deviations from means

X_i: Represents the i-th value of variable X.

Y_i: Represents the i-th value of variable Y.

n: Represents the total number of data points.

Standardized Height (X_1)	Standardized Weight (X_2)
-1.08	-1.25
0.31	-0.72
-0.74	-0.18
-0.25	0.54
1.76	1.61

	x1	x2
x1	cov(x1,x1)	cov(x1,x2)
x2	cov(x2,x1)	cov(x2,x2)

2.1 Calculate Covariance of Height with Height (Var(X_1))

$$\operatorname{cov}(X_1, X_1) = \frac{1}{5}\left[(-1.08)^2 + (0.31)^2 + (-0.74)^2 + (-0.25)^2 + (1.76)^2\right]$$

$$\operatorname{cov}(X_1, X_1) = \frac{1}{5}\left[1.1664 + 0.0961 + 0.5476 + 0.0625 + 3.0976\right] = \frac{4.9702}{5} = 0.9940$$

2.2 Calculate Covariance of Height with Weight (Cov(X_1, X_2))

$$\operatorname{cov}(X_1, X_2) = \frac{1}{5}\left[(-1.08)(-1.25) + (0.31)(-0.72) + (-0.74)(-0.18) + (-0.25)(0.54) + (1.76)(1.61)\right]$$

$$\operatorname{cov}(X_1, X_2) = \frac{1}{5}\left[1.35 + (-0.2232) + 0.1332 + (-0.135) + 2.8336\right] = \frac{4.9586}{5} = 0.9917$$

2.3 Calculate Covariance of Weight with Weight (Var(X_2))

$$\operatorname{cov}(X_2, X_2) = \frac{1}{5}\left[(-1.25)^2 + (-0.72)^2 + (-0.18)^2 + (0.54)^2 + (1.61)^2\right]$$

$$\operatorname{cov}(X_2, X_2) = \frac{1}{5}\left[1.5625 + 0.5184 + 0.0324 + 0.2916 + 2.5921\right] = \frac{5.9970}{5} = 1.1994$$

Covariance Matrix:

$$\mathrm{Cov}(X) = \begin{bmatrix} \mathrm{cov}(X_1, X_1) & \mathrm{cov}(X_1, X_2) \\ \mathrm{cov}(X_2, X_1) & \mathrm{cov}(X_2, X_2) \end{bmatrix}$$

$$\mathrm{Cov}(X) = \begin{bmatrix} 0.9940 & 0.9917 \\ 0.9917 & 1.1994 \end{bmatrix}$$

Step 3: Perform Eigen Decomposition

calculate the **eigenvalues** and **eigenvectors** of the covariance matrix to determine the principal components.

Step 3.2: Eigenvalue and Eigenvector Calculation

The eigenvalues and eigenvectors are calculated from the **characteristic equation.** The characteristic equation is:

$$\det(\mathrm{Cov}(X) - \lambda I) = 0$$

Where:

- λ represents the eigenvalue.
- I is the identity matrix.

So, the characteristic equation for our covariance matrix $\mathrm{Cov}(X)$ becomes:

$$\det\left(\begin{bmatrix} 0.9940 & 0.9917 \\ 0.9917 & 1.1994 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) = 0$$

Which simplifies to:

$$\det\begin{bmatrix} 0.9940 - \lambda & 0.9917 \\ 0.9917 & 1.1994 - \lambda \end{bmatrix} = 0$$

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$$

In our case, we have:

$$\det \begin{bmatrix} 0.9940 - \lambda & 0.9917 \\ 0.9917 & 1.1994 - \lambda \end{bmatrix} = (0.9940 - \lambda)(1.1994 - \lambda) - (0.9917)(0.9917)$$

We need to expand the product $(0.9940 - \lambda)(1.1994 - \lambda)$.

$$(a - b)(c - d) = ac - ad - bc + bd$$

So, applying this to our case:

$$(0.9940 - \lambda)(1.1994 - \lambda) = 0.9940 \times 1.1994 - 0.9940 \times \lambda - \lambda \times 1.1994 + \lambda^2$$

Now, we can simplify:

$$1.1910 - 2.1934\lambda + \lambda^2$$

Substitute in what we've found so far:

$$1.1910 - 2.1934\lambda + \lambda^2 - 0.9835 = 0$$

So, the equation becomes:

$$\lambda^2 - 2.1934\lambda + 0.2075 = 0$$

This is now a **quadratic equation** of the form:

$$\lambda^2 + b\lambda + c = 0$$

Where:

- $b = -2.1934$
- $c = 0.2075$

The **quadratic formula** is a well-known method for solving quadratic equations of the form $ax^2 + bx + c = 0$. The quadratic formula is:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

In our equation $\lambda^2 - 2.1934\lambda + 0.2075 = 0$, we have:

- $a = 1$ (the coefficient of λ^2)
- $b = -2.1934$
- $c = 0.2075$

Now, substitute these values into the quadratic formula:

$$\lambda = \frac{-(-2.1934) \pm \sqrt{(-2.1934)^2 - 4(1)(0.2075)}}{2(1)}$$

$$\lambda = \frac{2.1934 \pm \sqrt{4.8099 - 0.83}}{2}$$

Simplify the expression inside the square root:

$$\lambda = \frac{2.1934 \pm \sqrt{3.9799}}{2}$$

Take the square root of 3.9799:

$$\sqrt{3.9799} \approx 1.9949$$

So, we now have:

$$\lambda = \frac{2.1934 \pm 1.9949}{2}$$

This gives two possible solutions for λ:

- For the **plus** case: $\lambda_1 = \frac{2.1934+1.9949}{2} = \frac{4.1883}{2} \approx 2.09415$
- For the **minus** case: $\lambda_2 = \frac{2.1934-1.9949}{2} = \frac{0.1985}{2} \approx 0.09925$

Thus, the **eigenvalues** are:

$$\lambda_1 \approx 2.09415 \quad \text{and} \quad \lambda_2 \approx 0.09925$$

Step 3: Eigenvector Calculation for $\lambda_1 = 2.09415$

We start with the equation:

$$(\text{Cov}(X) - \lambda_1 I)v = 0$$

Substitute $\lambda_1 = \mathbf{2.09415}$ and the identity matrix I:

$$\text{Cov}(X) - \lambda_1 I = \begin{bmatrix} 0.9940 & 0.9917 \\ 0.9917 & 1.1994 \end{bmatrix} - 2.09415 \times \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Performing the matrix subtraction:

$$\begin{bmatrix} 0.9940 & 0.9917 \\ 0.9917 & 1.1994 \end{bmatrix} - \begin{bmatrix} 2.09415 & 0 \\ 0 & 2.09415 \end{bmatrix} = \begin{bmatrix} 0.9940 - 2.09415 & 0.9917 \\ 0.9917 & 1.1994 - 2.09415 \end{bmatrix}$$

$$= \begin{bmatrix} -1.10015 & 0.9917 \\ 0.9917 & -0.89475 \end{bmatrix}$$

Now, the equation becomes:

$$\begin{bmatrix} -1.10015 & 0.9917 \\ 0.9917 & -0.89475 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

This gives us a system of two linear equations:

1. $-1.10015 v_1 + 0.9917 v_2 = 0$
2. $0.9917 v_1 - 0.89475 v_2 = 0$

We will solve this system for v_1 and v_2.

Step 4: Solve the System of Equations

From equation (1):

$$-1.10015v_1 + 0.9917v_2 = 0$$

Solve for v_1:

$$v_1 = \frac{0.9917}{1.10015}v_2 \approx 0.9015v_2$$

Now substitute this into equation (2):

$$0.9917v_1 - 0.89475v_2 = 0$$

Substitute $v_1 = 0.9015v_2$ into the equation:

$$0.9917 \times 0.9015v_2 - 0.89475v_2 = 0$$

Factor out v_2:

$$v_2(0.9917 \times 0.9015 - 0.89475) = 0$$

Simplifying the terms inside the parentheses:

$$0.9917 \times 0.9015 = 0.89475$$

So the equation becomes:

$$v_2(0.89475 - 0.89475) = 0$$

$$v_2 \times 0 = 0$$

$$v_2 \times 0 = 0$$

This equation is trivially true for any v_2, so we conclude that $v_1 = 0.9015 v_2$. Thus, the eigenvector corresponding to $\lambda_1 = 2.09415$ is:

$$v_1 = \begin{bmatrix} 0.9015 \\ 1 \end{bmatrix}$$

To normalize the eigenvector (so that it has unit length), we divide by the magnitude of the vector. The magnitude of v_1 is:

$$\|v_1\| = \sqrt{0.9015^2 + 1^2} = \sqrt{0.8137 + 1} = \sqrt{1.8137} \approx 1.347$$

So the normalized eigenvector is:

$$v_1 = \begin{bmatrix} \frac{0.9015}{1.347} \\ \frac{1}{1.347} \end{bmatrix} \approx \begin{bmatrix} 0.6692 \\ 0.7421 \end{bmatrix}$$

Thus, the normalized eigenvector corresponding to $\lambda_1 = 2.09415$ is:

$$v_1 \approx \begin{bmatrix} 0.6692 \\ 0.7421 \end{bmatrix}$$

Step 5: Eigenvector Calculation for $\lambda_2 = 0.09925$

We now repeat the same process for the second eigenvalue $\lambda_2 = 0.09925$.

Start with:

$$(\text{Cov}(X) - \lambda_2 I)v = 0$$

Substitute $\lambda_2 = 0.09925$:

$$\text{Cov}(X) - \lambda_2 I = \begin{bmatrix} 0.9940 & 0.9917 \\ 0.9917 & 1.1994 \end{bmatrix} - 0.09925 \times \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.9940 - 0.09925 & 0.9917 \\ 0.9917 & 1.1994 - 0.09925 \end{bmatrix} = \begin{bmatrix} 0.89475 & 0.9917 \\ 0.9917 & 1.10015 \end{bmatrix}$$

Now, the equation becomes:

$$\begin{bmatrix} 0.89475 & 0.9917 \\ 0.9917 & 1.10015 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

This is another system of equations:

1. $0.89475v_1 + 0.9917v_2 = 0$
2. $0.9917v_1 + 1.10015v_2 = 0$

By solving this system in the same way as we did for λ_1, we obtain the second eigenvector:

$$v_2 = \begin{bmatrix} -0.7421 \\ 0.6692 \end{bmatrix}$$

For $\lambda_1 = 0.09925$:

$$\text{Cov}(X) - \lambda_1 I = \begin{bmatrix} 0.9940 & 0.9917 \\ 0.9917 & 1.1994 \end{bmatrix} - 0.09925 \times \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.9940 - 0.09925 & 0.9917 \\ 0.9917 & 1.1994 - 0.09925 \end{bmatrix} = \begin{bmatrix} 0.89475 & 0.9917 \\ 0.9917 & 1.10015 \end{bmatrix}$$

Now, we will solve the system of equations:

$$\begin{bmatrix} 0.89475 & 0.9917 \\ 0.9917 & 1.10015 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

This gives us the following system of equations:

1. $0.89475 v_1 + 0.9917 v_2 = 0$
2. $0.9917 v_1 + 1.10015 v_2 = 0$

We can solve this system for v_1 and v_2. To simplify, let's first solve equation 1 for v_1:

$$v_1 = -\frac{0.9917}{0.89475} v_2 = -1.1072 v_2$$

Substitute this into equation 2:

$$0.9917(-1.1072 v_2) + 1.10015 v_2 = 0$$

$$-1.0964 v_2 + 1.10015 v_2 = 0$$

$$0.00375 v_2 = 0$$

$$0.00375v_2 = 0$$

So, we conclude that v_2 can be any non-zero value, and v_1 is proportional to v_2.

Eigenvector for λ_1:

Choosing $v_2 = 1$, we get:

$$v_1 = -1.1072$$

Thus, the eigenvector corresponding to $\lambda_1 = 0.09925$ is approximately:

$$v_1 \approx \begin{bmatrix} -1.1072 \\ 1 \end{bmatrix}$$

This vector can be normalized to ensure it has unit length. The magnitude is:

$$\|v_1\| = \sqrt{(-1.1072)^2 + (1)^2} = \sqrt{1.2268 + 1} = \sqrt{2.2268} \approx 1.4945$$

So, the normalized eigenvector is:

$$v_1 = \frac{1}{1.4945} \begin{bmatrix} -1.1072 \\ 1 \end{bmatrix} \approx \begin{bmatrix} -0.7403 \\ 0.6692 \end{bmatrix}$$

4. Projection onto Principal Component 1 (PC1):

Now, we can use the standardized data and project it onto the first principal component.

Recall that the first principal component (**PC1**) eigenvector is:

$$\mathbf{v_1} = [0.6692, 0.7421]$$

Projection onto PC1 for each person:

The formula for the projection is:

$$\text{Projection on PC1} = (\text{Standardized Height}) \times 0.6692 + (\text{Standardized Weight}) \times 0.7421$$

For Person A:

$$\text{Projection on PC1} = (-1.09 \times 0.6692) + (-1.25 \times 0.7421) = -0.7296 - 0.9276 = -1.6572$$

For Person B:

$$\text{Projection on PC1} = (0.31 \times 0.6692) + (-0.72 \times 0.7421) = 0.2074 - 0.5341 = -0.3267$$

For Person C:

$$\text{Projection on PC1} = (-0.74 \times 0.6692) + (-0.18 \times 0.7421) = -0.4954 - 0.1336 = -0.6290$$

For Person D:

$$\text{Projection on PC1} = (-0.25 \times 0.6692) + (0.54 \times 0.7421) = -0.1673 + 0.4005 = 0.2332$$

For Person E:

$$\text{Projection on PC1} = (1.77 \times 0.6692) + (1.61 \times 0.7421) = 1.1833 + 1.1947 = 2.3780$$

5. **Final Transformed Data** (1D Projection onto PC1):

Person	Projection on PC1
A	-1.6572
B	-0.3267
C	-0.6290
D	0.2332
E	2.3780

Original Dataset:			Transformed:	
Person	Height (cm)	Weight (kg)	Person	Principal Component
A	150	52	A	-1.6572
B	170	55	B	-0.3267
C	155	58	C	-0.6290
D	162	62	D	0.2332
E	191	68	E	2.3780

Original vs Projected