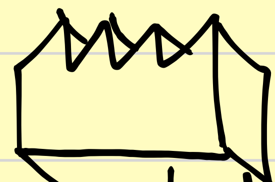
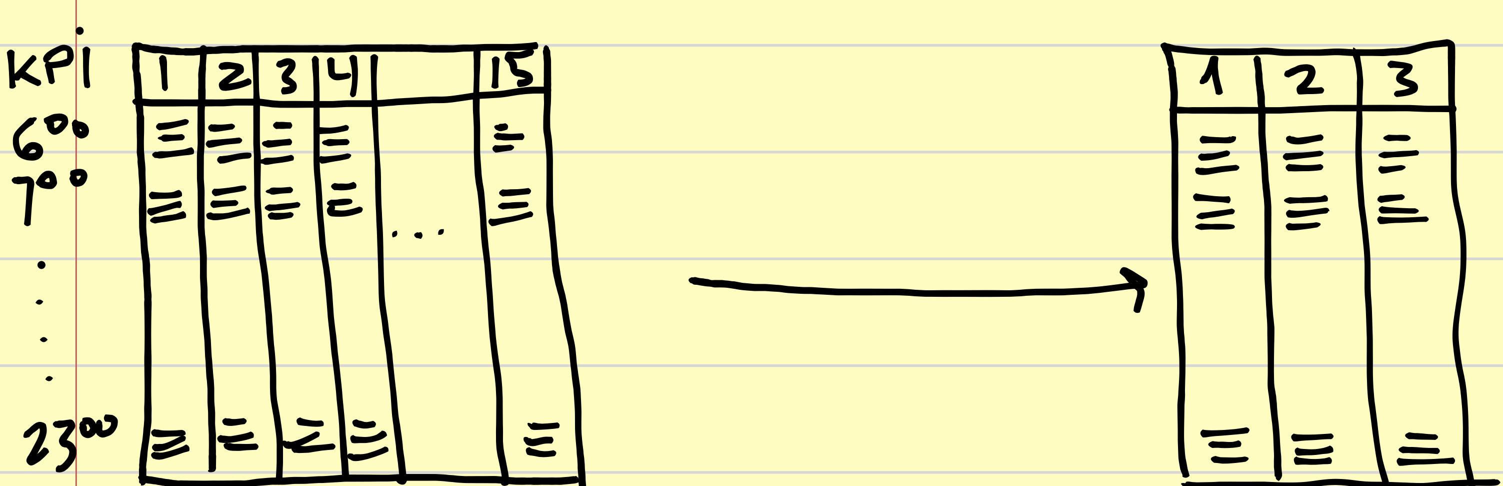
DIMENSIONALITY REDUCTION

How much variability can we explain from many-dimensional datasets and still be able to manage them?

CASE: we have a business process  that is measured with many KPIs (Key Performance Indicators), and we aim to find out a method that allows us to manage the process with less (reduction) KPIs but still keeping a high amount of the explainability intact.

Example: We have a factory that gathers 15 KPIs every day. Each KPI cost 10.000€/month to keep & update, and we ask the question if we could reduce to 3 KPIs and still represent 80% of the variability of the process.

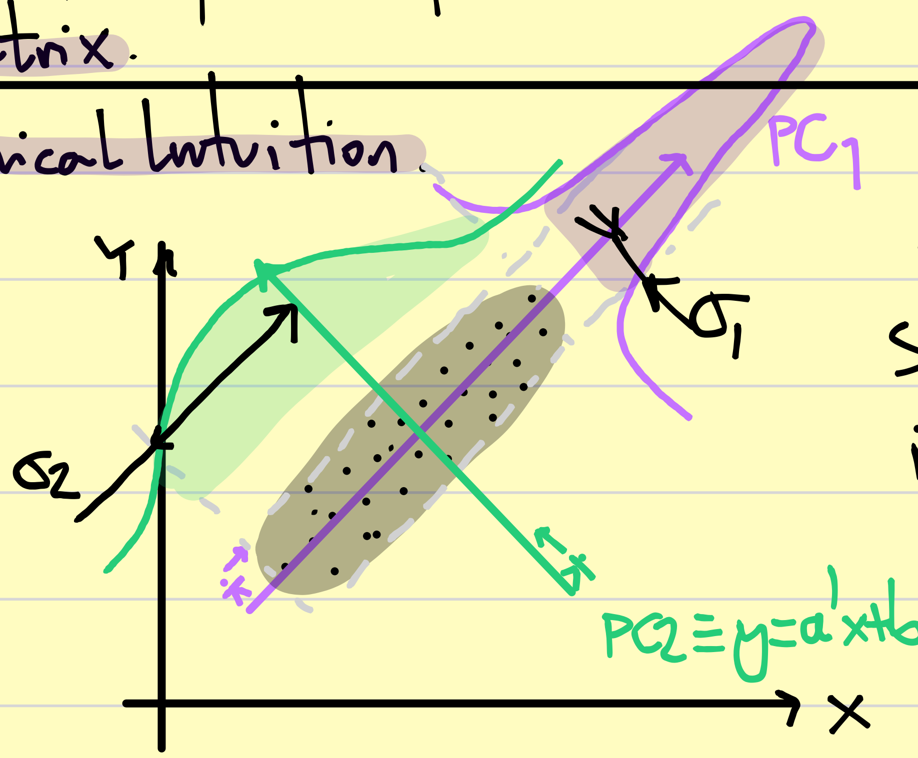


Explainability	100%	80%
Cost	150000€/month	30000€/month

PRINCIPAL COMPONENT ANALYSIS (PCA)

Def. The principal components are the eigenvectors of the covariance matrix.

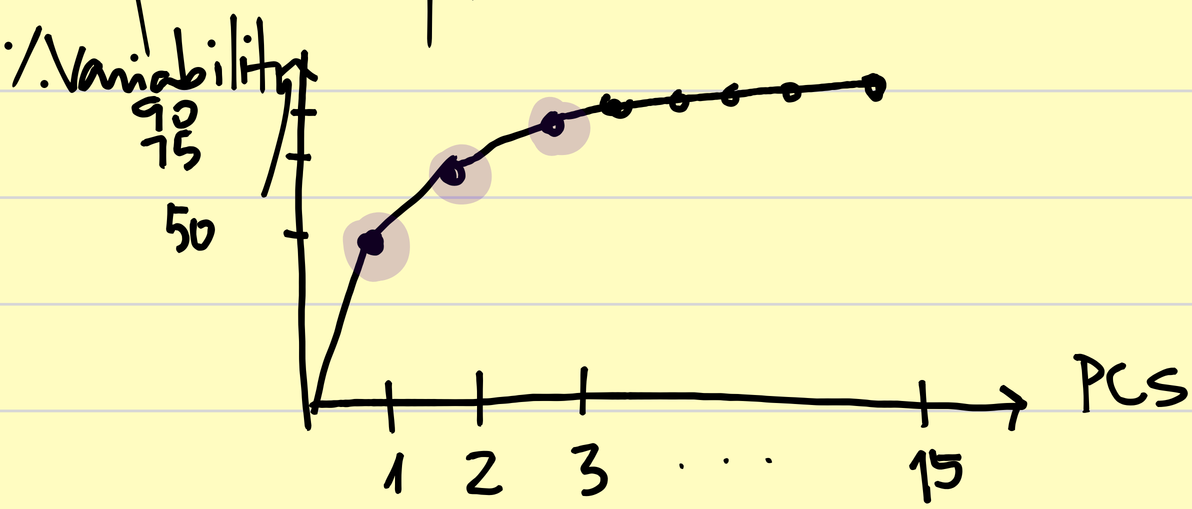
Graphical Intuition



- The direction in which the std Deviation σ_1 of the data is minimal, is the first PC.
- The perpendicular ⊥ direction to PC_1, is called PC_2.
- And because $\sigma_1 \ll \sigma_2$,

we can say that the first PC_1 explains more variability of the data that the second PC_2.

- The vector → PC_1 explains the dataset better than any other direction.
- When you have a "n" dimensional dataset, you can calculate "n" different PCs. But usually the variability explained by them is PARETO distributed:



COVARIANCE MATRIX 3×3 (3 variables)

	KPI$_1$	KPI$_2$	KPI$_3$
CW$_1$	≡	≡	≡
CW$_2$	≡	≡	≡
⋮			
CW$_n$	≡	≡	≡
	x	Y	Z

$$\boxed{\text{COV MATRIX}[X,Y,Z] = A = \begin{bmatrix} VAR(X) & COV(X,Y) & COV(X,Z) \\ COV(X,Y) & VAR(Y) & COV(Y,Z) \\ COV(X,Z) & COV(Y,Z) & VAR(Z) \end{bmatrix}}$$

$$VAR(X) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \qquad COV(X,Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Example for 2 KPIs:

KPI$_1$ ≡ €/Unit [17, 19, 23, 22] ≡ x

KPI$_2$ ≡ Revenue [34, 41, 46, 45] ≡ y

$$\text{COV MATRIX} = \begin{bmatrix} VAR(X) & COV(X,Y) \\ COV(X,Y) & VAR(Y) \end{bmatrix} = \begin{bmatrix} 7'67 & 14'5 \\ 14'5 & 29'67 \end{bmatrix}$$

$$\bar{x} = \frac{17+19+23+22}{4} = 20'25$$

$$\bar{y} = \frac{34+41+46+45}{4} = 41'5$$

$$VAR(X) = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{(17-20'5)^2 + (19-20'5)^2 + (23-20'5)^2 + (22-20'5)^2}{4-1} = 7'67$$

$$VAR(Y) = \frac{\sum(y_i - \bar{y})^2}{n-1} = \frac{(34-41'5)^2 + (41-41'5)^2 + (46-41'5)^2 + (45-41'5)^2}{4-1} = 29'67$$

$$\text{cov}(X,Y) = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1} = \frac{(17-20'5)(34-41'5)+(19-20'5)(41-41'5)+(23-20'5)(46-41'5)+(22-20'5)(45-41'5)}{4-1}$$

$$= \frac{43'5}{3} = 14'5$$

$$\boxed{A = \begin{bmatrix} 7'67 & 14'5 \\ 14'5 & 29'67 \end{bmatrix}}$$

Eigenvector calculation:

$A\vec{v} = \lambda\vec{v}$: λ [Eigenvalues] $\vec{v}$ [Eigenvectors]

$\hookrightarrow \det[A-\lambda I] = 0 \longrightarrow \lambda \longrightarrow \vec{v}$

$$\det\left[\begin{bmatrix} 7'67 & 14'5 \\ 14'5 & 29'67 \end{bmatrix} - \lambda\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right] = 0 \rightarrow \det\begin{bmatrix} 7'67-\lambda & 14'5 \\ 14'5 & 29'67-\lambda \end{bmatrix} = 0$$

$$\boxed{\det\begin{bmatrix} a & c \\ b & d \end{bmatrix} = a.d - c.b} \quad \rightarrow [7'67-\lambda][29'67-\lambda] - 14'5 \cdot 14'5 = 0$$

$$\rightarrow 7'67 \cdot 29'67 - 7'67\lambda - 29'67 \cdot \lambda + \lambda^2 - 14'5 \cdot 14'5 = 0$$

$$\rightarrow \lambda^2 - 37'34\lambda + 17'3189 = 0 \qquad ax^2 + bx + c = 0$$

$$\rightarrow \lambda = \frac{37'34 \pm \sqrt{37'34^2 - 4 \cdot 17'3189}}{2} = \frac{37'34 \pm 36'4}{2} = \begin{cases} \lambda_1 = 36'87 \\ \lambda_2 = 0'47 \end{cases}$$

$$\boxed{\lambda_1 = 36'87} \rightarrow A \cdot \vec{v_1} = \lambda_1 \vec{v_1} \rightarrow \begin{bmatrix} 7'67 & 14'5 \\ 14'5 & 29'67 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = 36'87 \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix}$$

$$\rightarrow \begin{array}{l} 7'67 \cdot v_{11} + 14'5 \cdot v_{12} = 36'87 \cdot v_{11} \\ 14'5 \cdot v_{11} + 29'67 \cdot v_{12} = 36'87 \cdot v_{12} \end{array} \Bigg| \triangleright \begin{array}{l} v_{11} = \dfrac{14'5}{29'2} v_{12} = 0'49 v_{12} \\ v_{11} = 0'49 v_{12} \end{array}$$

$$v_{12} = 1 \rightarrow v_{11} = 0'49 \rightarrow \vec{v_1} = \begin{bmatrix} 0'49 \\ 1 \end{bmatrix}$$

EIGENVECTOR NR1

$\equiv$

PRINCIPAL COMPONENT

$$\boxed{\lambda_2 = 0'47} \rightarrow A \cdot \vec{v_2} = \lambda_2 \vec{v_2} \rightarrow \begin{bmatrix} 7'67 & 14'5 \\ 14'5 & 29'67 \end{bmatrix} \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = 0'47 \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix}$$

$$\rightarrow \begin{array}{l} 7'67 v_{21} + 14'5 v_{22} = 0'47 v_{21} \\ 14'5 v_{21} + 29'67 v_{22} = 0'47 v_{22} \end{array} \Bigg| \triangleright \begin{array}{l} v_{21} = \dfrac{14'5}{-7'2} v_{22} = -2'01 v_{22} \\ v_{21} = -2'01 v_{22} \end{array}$$

$$v_{22} = 1 \rightarrow v_{21} = -2'01 \rightarrow \vec{v_2} = \begin{bmatrix} -2'01 \\ 1 \end{bmatrix}$$

EIGENVECTOR NR2

$\equiv$

PC_2

Example for 2 KPIs:

$KPI_1 \equiv$ €/Unit $[17, 19, 23, 22] \equiv x$

$KPI_2 \equiv$ Revenue $[34, 41, 46, 45] \equiv y$

$$\vec{v_1} = PC_1 = \begin{bmatrix} 0'49 \\ 1 \end{bmatrix} \qquad \vec{v_2} = PC_2 = \begin{bmatrix} -2'01 \\ 1 \end{bmatrix}$$

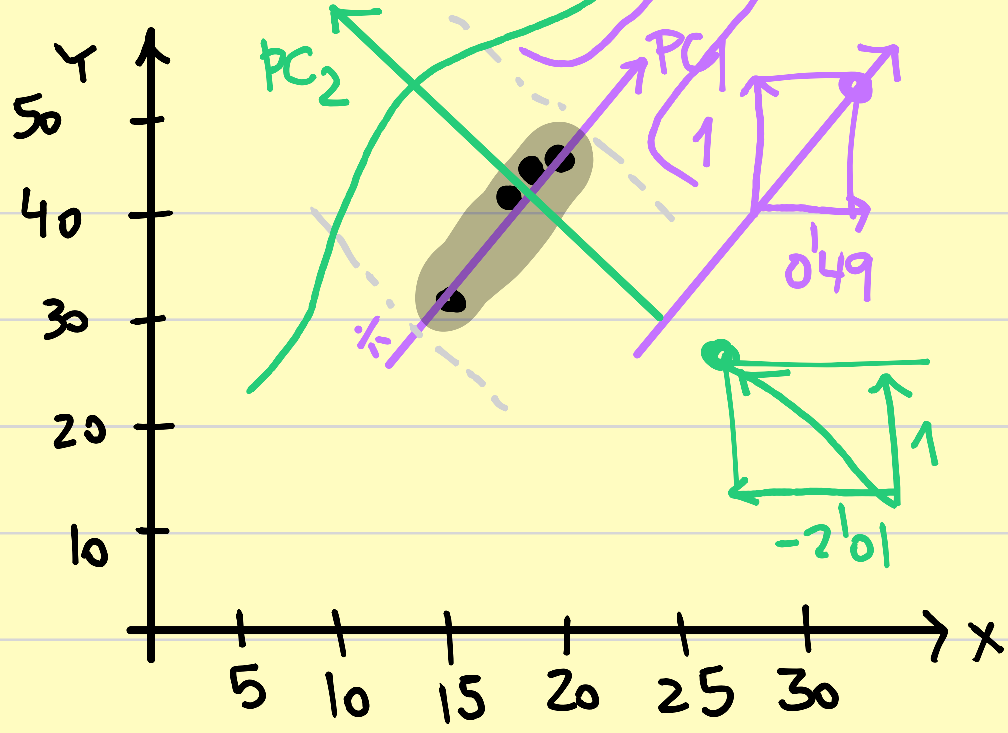
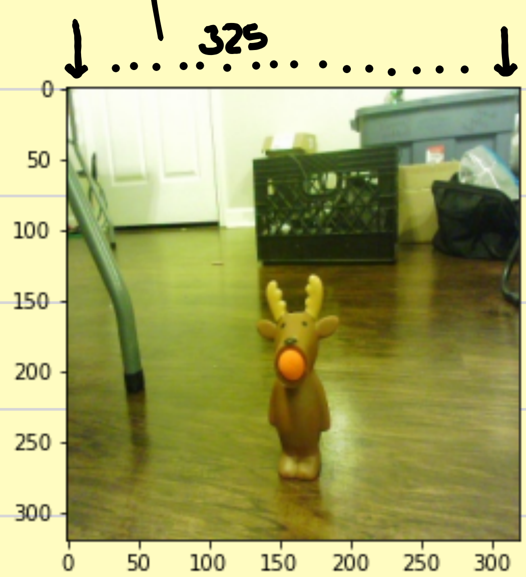


Image Compression



325x325x3 → 325 columns → 325 variables

→ PCA_1 . . . PCA_{325}