Principal Component Analysis (PCA)

Nguyen Minh Hoang vs Ho Thi Kim Cuong

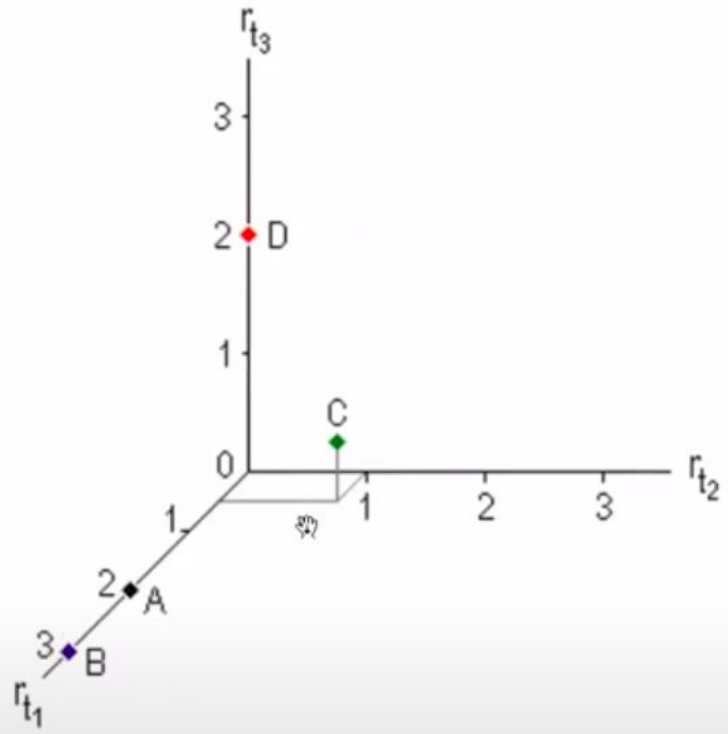
https://www.youtube.com/watch?v=jZH-TZjiwOY

Outline

- Idea of PCA and steps
- How it works through example
- Method overview
- Practice

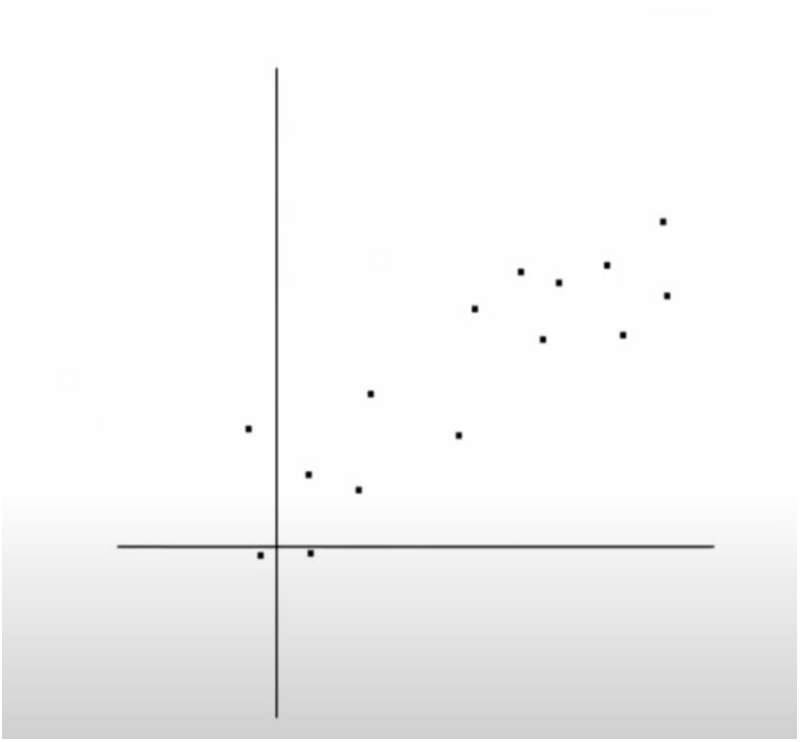
Idea of PCA

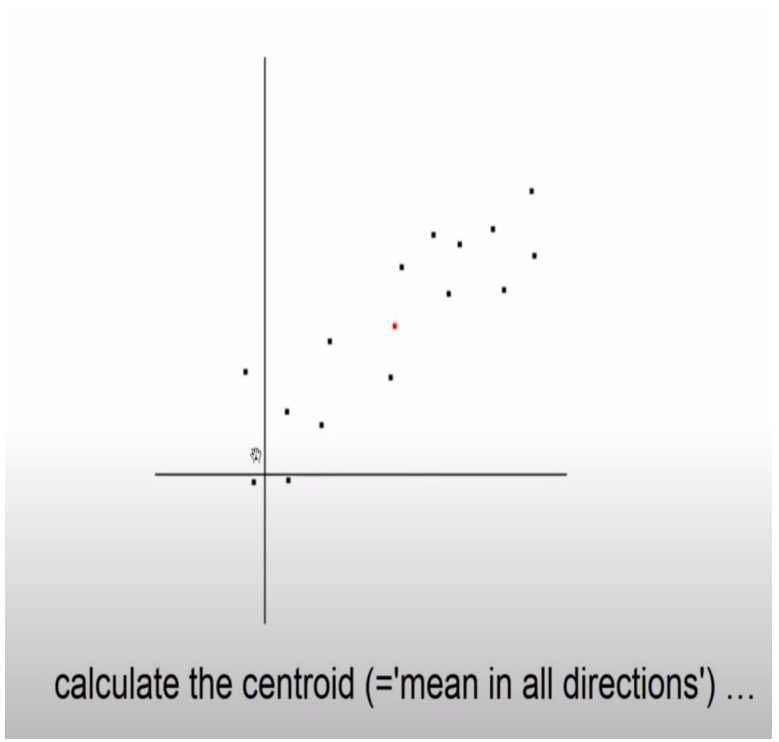
We are working with this representation:

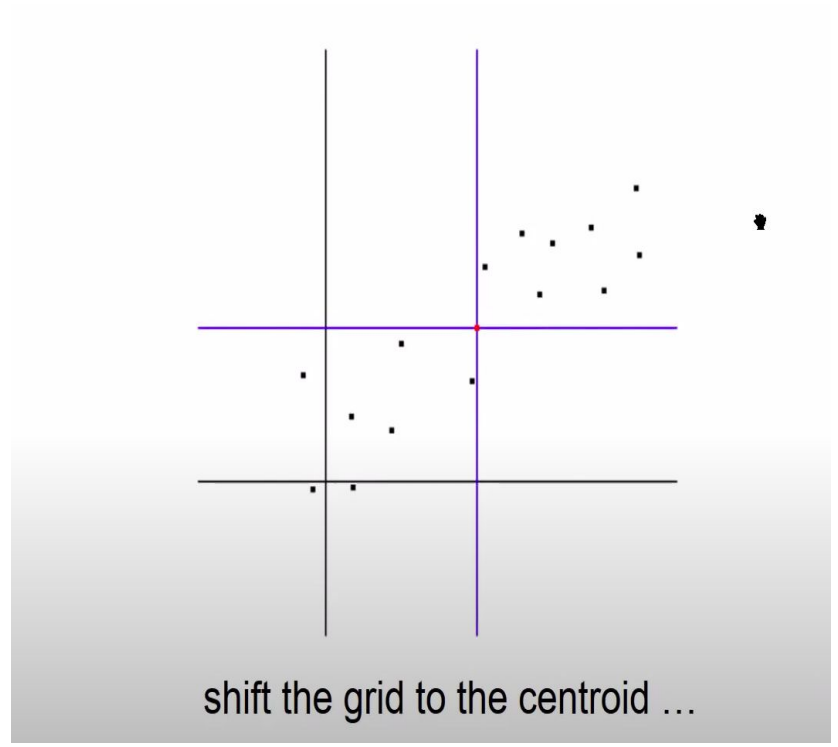


gene	t_1	t_2	t_3
A	2	0	0
B	3	0	0
C	0.5	1	0.5
D	0	0	2

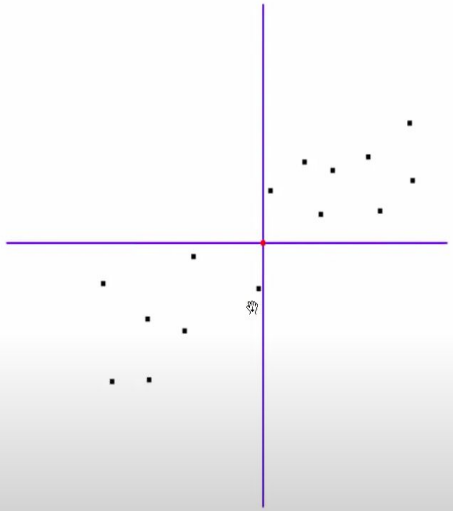
STEPS FOR PCA



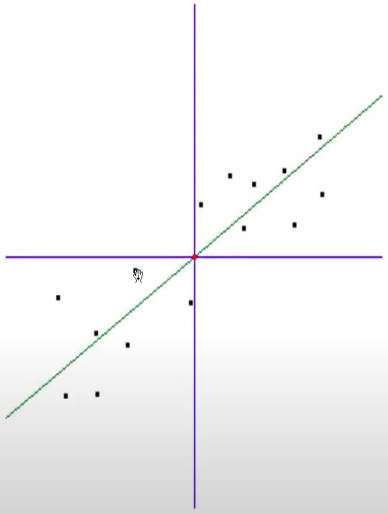




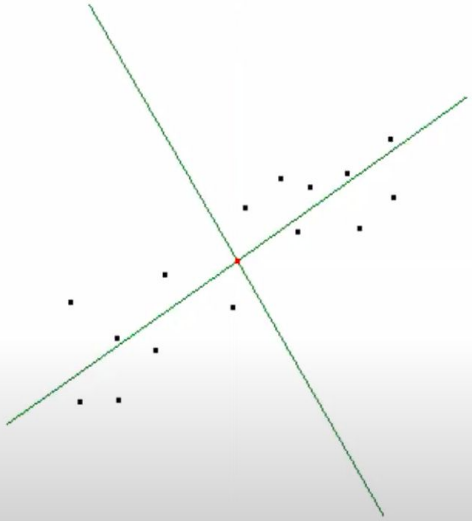
STEPS FOR PCA



take this as our new coordinate system …

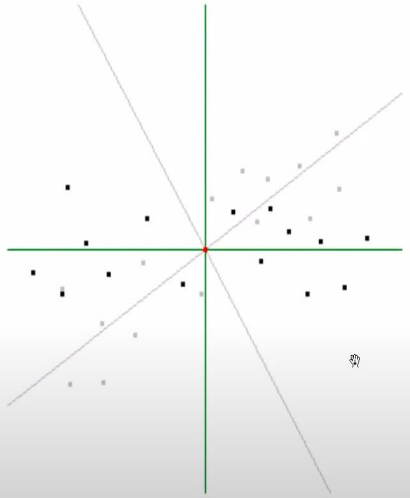


calculate the direction in which the variance is maximal …

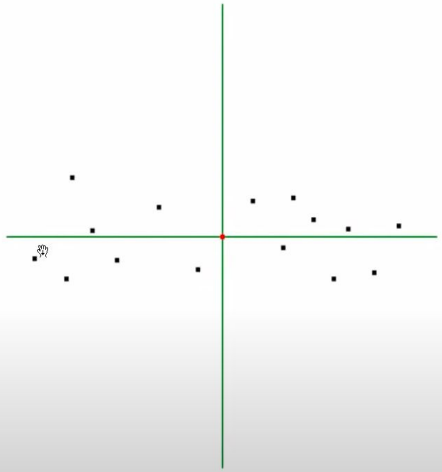


leaving us with a rotated grid …

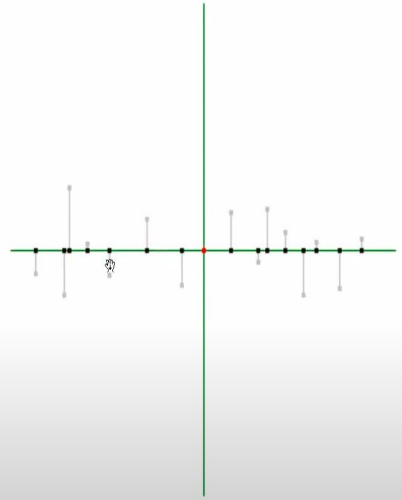
STEPS FOR PCA



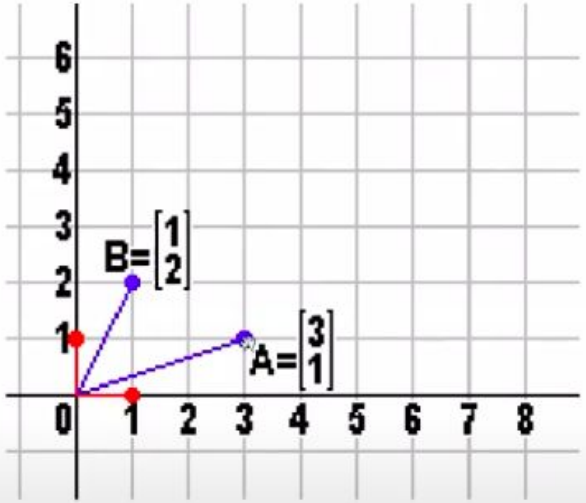
which we can rotate to a 'normal' position …



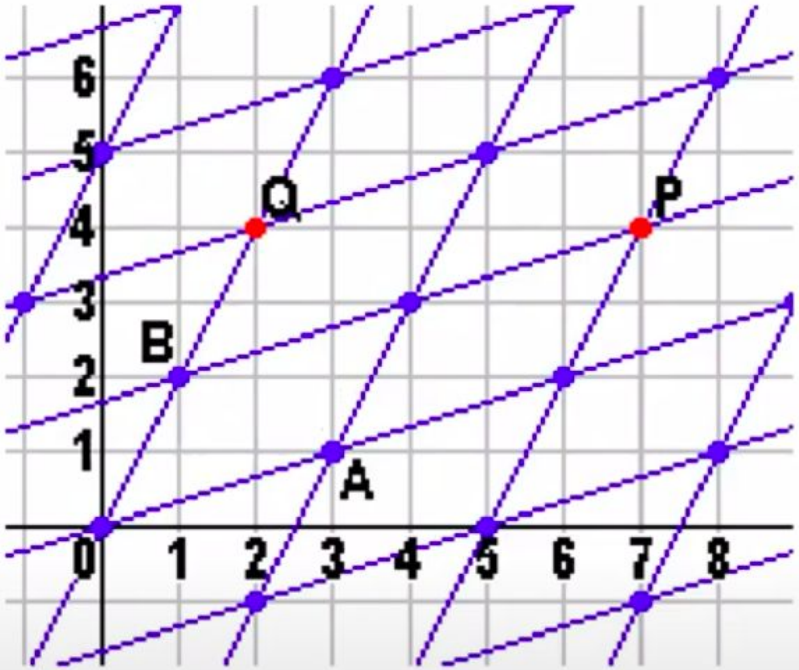
showing us maximal variance.

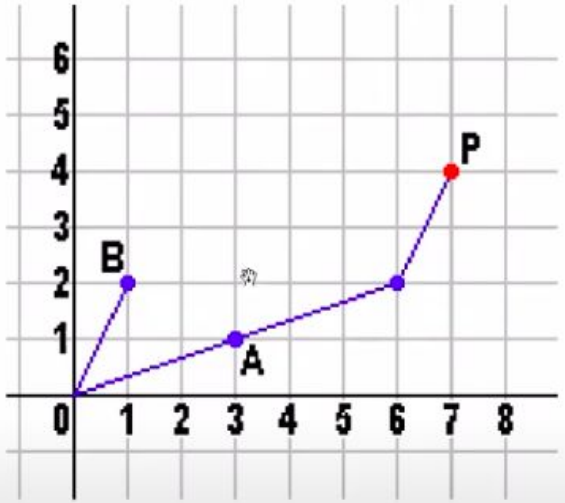


by eliminating a number of axis by projection of the points.



$$A = 3 \bullet \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 1 \bullet \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$





$$\mathbf{p} = 2 \bullet \mathbf{a} + 1 \bullet \mathbf{b}$$

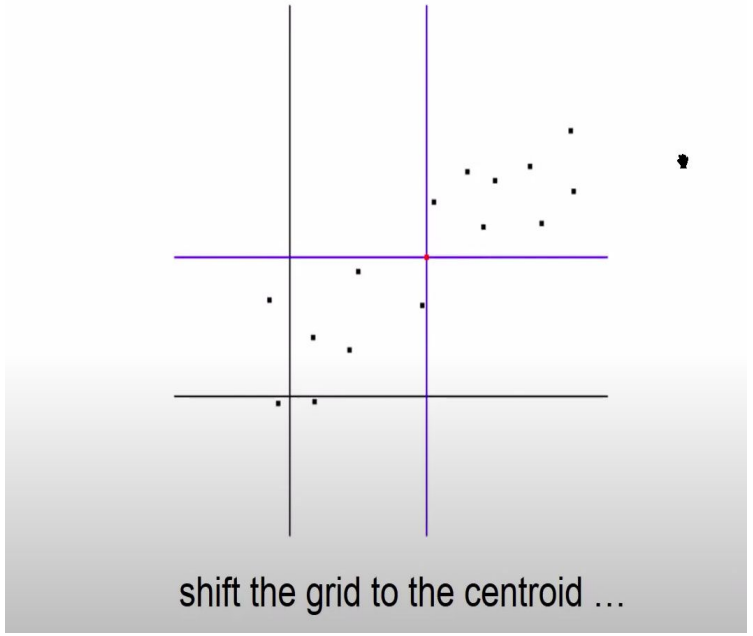
$$\mathbf{v}' = \begin{bmatrix} u \\ w \end{bmatrix} \rightarrow \mathbf{v} = u \bullet \begin{bmatrix} 3 \\ 1 \end{bmatrix} + w \bullet \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix} \bullet \begin{bmatrix} u \\ w \end{bmatrix} = \begin{bmatrix} 3 \bullet u + 1 \bullet w \\ 1 \bullet u + 2 \bullet w \end{bmatrix}$$

$$\mathbf{v} = \mathbf{A} \bullet \mathbf{v}' \qquad \mathbf{v}' = \mathbf{A}^{-1} \bullet \mathbf{v}$$

Examples

	X	Y
Gene A	4	2
Gene B	0	1
Gene C	8	7
Gene D	2	2
Gene E	6	3
μ	μ1= 4	μ2=3

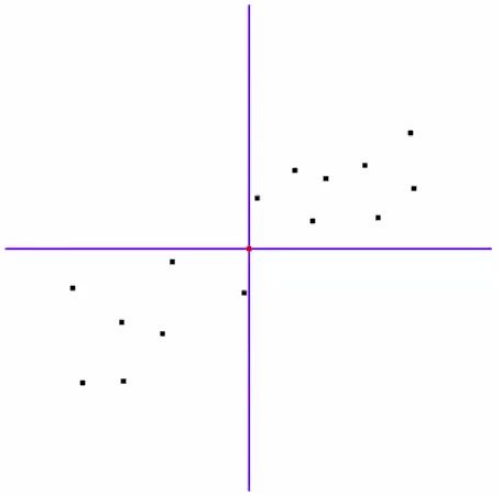
	X - μ1	Y - μ2
Gene A	4 - 4 = 0	2 - 3 = -1
Gene B	0 - 4 = -4	1 - 3 = -2
Gene C	8 - 4 = 4	7 - 3 = 4
Gene D	2 - 4 = -2	2 - 3 = -1
Gene E	6 - 4 = 2	3 - 3 = 0



PCA step 2: rotating the grid, base on variance

$\sigma^2_{x,y} = E(xy) - E(x)E(y)$ covariance

	x_1 (=x')	x_2 (=y')
gene A	0	-1
gene B	-4	-2
gene C	4	4
gene D	-2	-1
gene E	2	0



$\sigma^2_{x1,x2} = E(x_1x_2) = (0+8+16+2+0)/5 = 5.2$

$\sigma^2_{x2,x1} = \sigma^2_{x1,x2} = 5.2$

$\sigma^2_{x1,x1} = E(x_1x_1) = (0+16+16+4+4)/5 = 8$

$\sigma^2_{x2,x2} = E(x_2x_2) = (1+4+16+1+0)/5 = 4.4$

Covariance Matrix

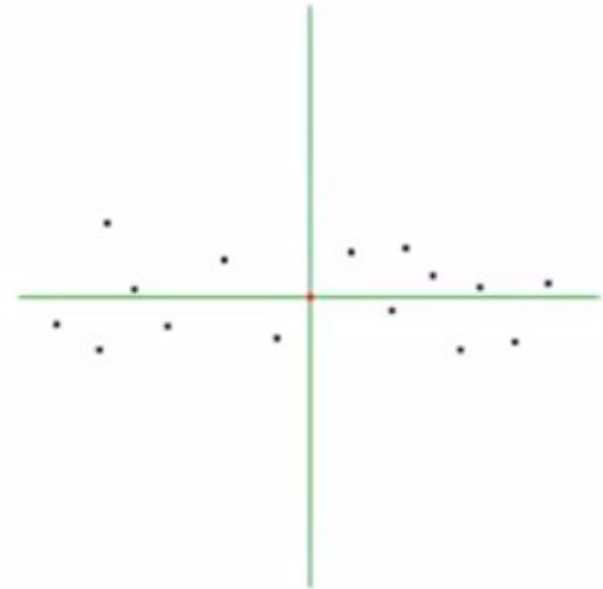
$$C = \begin{bmatrix} \sigma^2_{1,1} & \sigma^2_{1,2} \\ \sigma^2_{2,1} & \sigma^2_{2,2} \end{bmatrix} = \begin{bmatrix} 8 & 5.2 \\ 5.2 & 4.4 \end{bmatrix}$$

Principal Component Analysis

$$\sigma^2_{p'} = \sigma^2_{x'} + \sigma^2_{y'}$$

$$C' = \begin{bmatrix} e'_1 & 0 \\ 0 & e'_2 \end{bmatrix}$$

$$X = \begin{bmatrix} ev_{1,1} & ev_{1,2} \\ ev_{2,1} & ev_{2,2} \end{bmatrix} = \begin{bmatrix} 1' & 0' \\ 0' & 1' \end{bmatrix}$$



for each **v'** on x'-axis: $\mathbf{v'} = \begin{bmatrix} v' \\ 0 \end{bmatrix}$

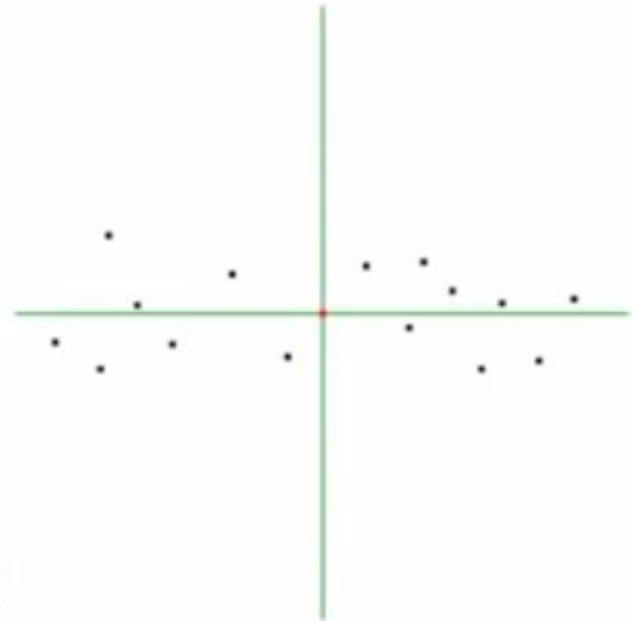
$$\text{cov}(\mathbf{v'}) = C' \bullet \mathbf{v'} = \begin{bmatrix} e'_1 & 0 \\ 0 & e'_2 \end{bmatrix} \bullet \begin{bmatrix} v' \\ 0 \end{bmatrix} = \begin{bmatrix} v' \bullet e'_1 \\ 0 \end{bmatrix} = e'_1 \bullet \begin{bmatrix} v' \\ 0 \end{bmatrix} = e'_1 \bullet \mathbf{v'}$$

$$C' \bullet \mathbf{v'} = \lambda_1 \bullet \mathbf{v'}$$

Principal Component Analysis

v_1' on x-axis: $C' \bullet \mathbf{v_1'} = \lambda_1 \bullet \mathbf{v_1'}$

v_2' on y-axis: $C' \bullet \mathbf{v_2'} = \lambda_2 \bullet \mathbf{v_2'}$



λ_i = eigenvalue of C'
v_i = eigenvector corresponding to λ_i

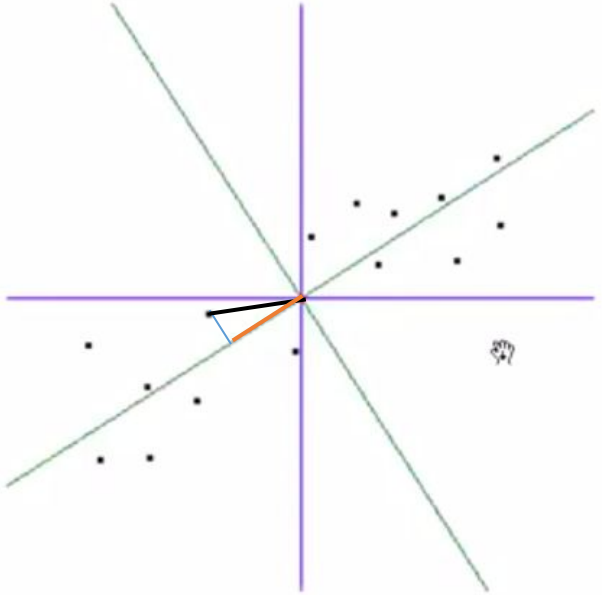
The value of λ_i corresponds to the variance on the x_i-axis

$$C = \begin{bmatrix} \sigma^2_{1,1} & \sigma^2_{1,2} \\ \sigma^2_{2,1} & \sigma^2_{2,2} \end{bmatrix} = \begin{bmatrix} 8 & 5.2 \\ 5.2 & 4.4 \end{bmatrix} \qquad X = \begin{bmatrix} ev_{1,1} & ev_{1,2} \\ ev_{2,1} & ev_{2,2} \end{bmatrix}$$

We have to solve: $C \bullet \mathbf{x} = \lambda \bullet \mathbf{x}$ for all λ and $\mathbf{x}$ (with $|\mathbf{x}_i| \equiv 1$)

$$\begin{bmatrix} 8 & 5.2 \\ 5.2 & 4.4 \end{bmatrix} \bullet \mathbf{x} = \lambda \bullet \mathbf{x} = \lambda \bullet \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \bullet \mathbf{x} = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \bullet \mathbf{x}$$

$$\Leftrightarrow \begin{bmatrix} 8-\lambda & 5.2 \\ 5.2 & 4.4-\lambda \end{bmatrix} \bullet \mathbf{x} = 0 \Leftrightarrow \begin{cases} 8\mathbf{x}_1 - \lambda\mathbf{x}_1 + 5.2\mathbf{x}_2 = 0 \\ 5.2\mathbf{x}_1 + 4.4\mathbf{x}_2 - \lambda\mathbf{x}_2 = 0 \\ \mathbf{x}_1^2 + \mathbf{x}_2^2 = 1 \end{cases}$$



$$\lambda_1 \approx 73.59 \qquad \mathbf{x}_1 \approx \begin{bmatrix} 0.8428 \\ 0.5383 \end{bmatrix}$$

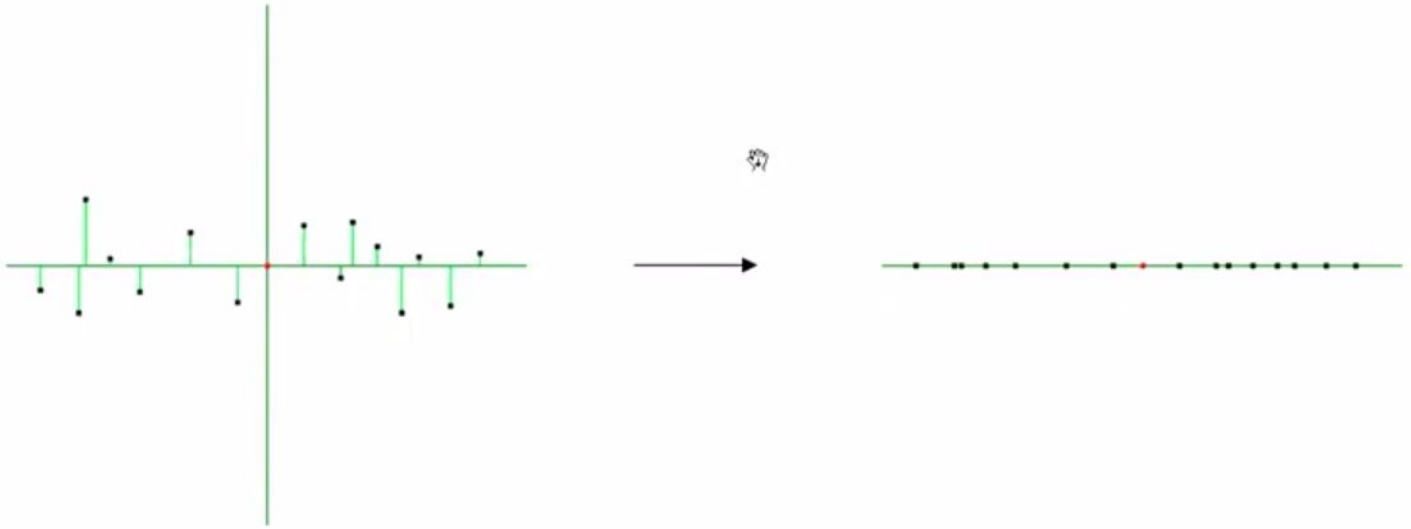
$$\lambda_2 \approx 4.33 \qquad \mathbf{x}_2 \approx \begin{bmatrix} -0.5383 \\ 0.8428 \end{bmatrix}$$

$$\sigma^2_{p'} = \sigma^2_{x'} + \sigma^2_{y'}$$

$\lambda_1 \approx 73.59 \cong 94.44\%$ of the total variance

$\lambda_2 \approx 4.33 \cong 5.56\%$ of the total variance

PCA step 3: reducing complexity



Reducing complexity = removing dimensions:

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \rightarrow \mathbf{v'} = \begin{bmatrix} v_1 \\ 0 \end{bmatrix} \cong [v_1]$$

Dimensionality Reduction

- One approach to deal with high dimensional data is by reducing their dimensionality.
- Project high dimensional data onto a lower dimensional sub-space using linear or non-linear transformations.

$$x = \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_N \end{bmatrix} \text{--->} \textit{reduce dimensionality} \text{--->} y = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_K \end{bmatrix} \quad (K << N)$$

Dimensionality Reduction

- Linear transformations are simple to compute and tractable.

$$Y = U \; X \qquad (b_i = u_i^t a_i)$$

k x 1 k x d d x 1 (k<<d)

- Classical –linear- approaches:
  - Principal Component Analysis (PCA)
  - Fisher Discriminant Analysis (FDA)

Principal Component Analysis (PCA)

- Find a basis in a low dimensional sub-space:
  - Approximate vectors by projecting them in a low dimensional sub-space:

(1) Original space representation:

$$\begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_N \end{bmatrix}$$

$$x = a_1 v_1 + a_2 v_2 + \dots + a_N v_N$$

where $v_1, v_2, \dots, v_n$ is a base in the original N-dimensional space

(2) Lower-dimensional sub-space representation:

$$\begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_K \end{bmatrix}$$

$$\hat{x} = b_1 u_1 + b_2 u_2 + \dots + b_K u_K$$

where $u_1, u_2, \dots, u_K$ is a base in the K-dimensional sub-space (K<N)

- *Note:* if K=N, then $\hat{x} = x$

Principal Component Analysis (PCA)

- Example (K=N):

$$v_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, v_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, v_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad \text{(standard basis)}$$

$$x_v = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix} = 3v_1 + 3v_2 + 3v_3$$

$$u_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, u_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, u_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \quad \text{(some other basis)}$$

$$x_u = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix} = 0u_1 + 0u_2 + 3u_3$$

thus, $x_v = x_u$

Principal Component Analysis (PCA)

- Information loss
  - Dimensionality reduction implies information loss !!
  - PCA preserves as much information as possible:

$$\min \| x - \hat{x} \| \quad \text{(reconstruction error)}$$

- What is the "best" lower dimensional sub-space?

> The "best" low-dimensional space is centered at the sample mean and has directions determined by the "best" eigenvectors of the covariance matrix of the data x.

  - By "best" eigenvectors we mean those corresponding to the *largest* eigenvalues (i.e., **"principal components"**).
  - Since the covariance matrix is real and symmetric, these eigenvectors are orthogonal and form a set of basis vectors.

(see pp. 114-117 in textbook for a proof)

Principal Component Analysis (PCA)

- Methodology
  - Suppose $x_1, x_2, ..., x_M$ are N x 1 vectors

Step 1: $\bar{x} = \frac{1}{M}\sum_{i=1}^{M} x_i$

Step 2: subtract the mean: $\Phi_i = x_i - \bar{x}$

Step 3: form the matrix $A = [\Phi_1\ \Phi_2 \cdots \Phi_M]$ (NxM matrix), then compute:

$$C = \frac{1}{M}\sum_{n=1}^{M} \Phi_n \Phi_n^T = AA^T$$

(sample **covariance** matrix, NxN, characterizes the *scatter* of the data)

Step 4: compute the eigenvalues of C: $\lambda_1 > \lambda_2 > \cdots > \lambda_N$

Step 5: compute the eigenvectors of C: $u_1, u_2, \ldots, u_N$

Principal Component Analysis (PCA)

- Methodology – cont.

- Since C is symmetric, $u_1, u_2, \ldots, u_N$ form a basis, (i.e., any vector x or actually $(x - \bar{x})$, can be written as a linear combination of the eigenvectors):

$$x - \bar{x} = b_1 u_1 + b_2 u_2 + \cdots + b_N u_N = \sum_{i=1}^{N} b_i u_i \qquad \boxed{b_i = u_i^T (x - \bar{x})}$$

Step 6: (**dimensionality reduction step**) keep only the terms corresponding to the K largest eigenvalues:

$$\hat{x} - \bar{x} = \sum_{i=1}^{K} b_i u_i \text{ where } K << N$$

- The representation of $\hat{x} - \bar{x}$ into the basis $u_1, u_2, ..., u_K$ is thus

$$\begin{bmatrix} b_1 \\ b_2 \\ \ldots \\ b_K \end{bmatrix}$$

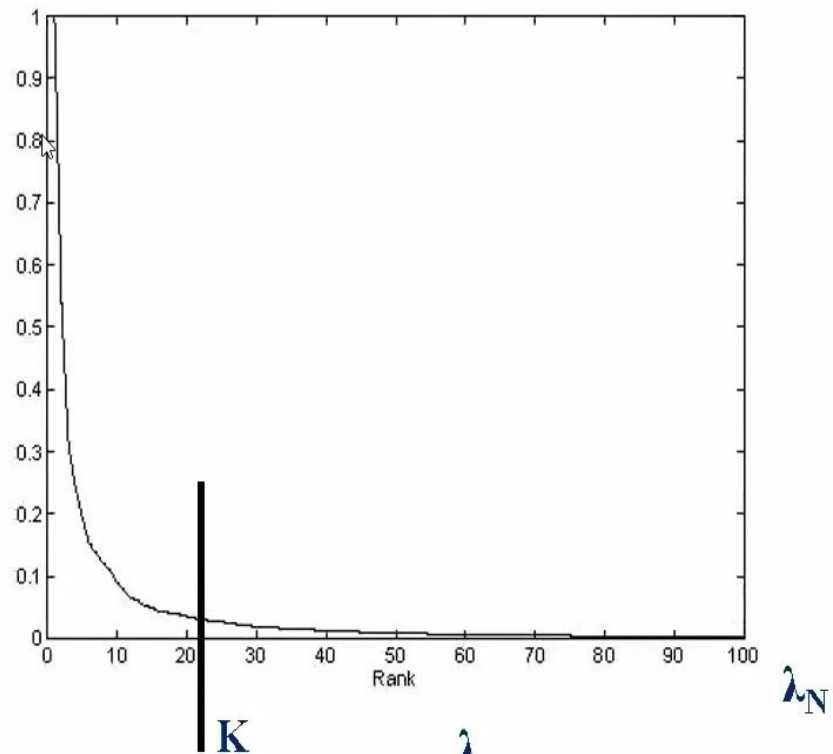
Principal Component Analysis (PCA)

- How many principal components to keep?
  - To choose K, you can use the following criterion:

$$\frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{N} \lambda_i} > Threshold \quad \text{(e.g., 0.9 or 0.95)}$$

Principal Component Analysis (PCA)

- Eigenvalue spectrum



Principal Component Analysis (PCA)

- Linear transformation implied by PCA
  - The linear transformation $R^N \rightarrow R^K$ that performs the dimensionality reduction is:

$$\begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_K \end{bmatrix} = \begin{bmatrix} u_1^T \\ u_2^T \\ \dots \\ u_K^T \end{bmatrix} (x - \bar{x}) = U^T (x - \bar{x})$$

Principal Component Analysis (PCA)

- What is the error due to dimensionality reduction?

$$e = \|x - \hat{x}\|$$

$$\hat{x} - \bar{x} = \sum_{i=1}^{K} b_i u_i \text{ or } \hat{x} = \sum_{i=1}^{K} b_i u_i + \bar{x}$$

- It can be shown that the average error due to dimensionality reduction is equal to:

$$\overline{e} = 1/2 \sum_{i=K+1}^{N} \lambda_i$$

Principal Component Analysis (PCA)

- Standardization
  - The principal components are dependent on the *units* used to measure the original variables as well as on the *range* of values they assume.
  - We should always standardize the data prior to using PCA.
  - A common standardization method is to transform all the data to have zero mean and unit standard deviation:

$$\frac{x_i - \mu}{\sigma} \quad (\mu \text{ and } \sigma \text{ are the mean and standard deviation of } x_i\text{'s})$$