Inspire…Educate…Transform.

Essentials Engineering Skills for Big Data Analytics

PCA & SVD

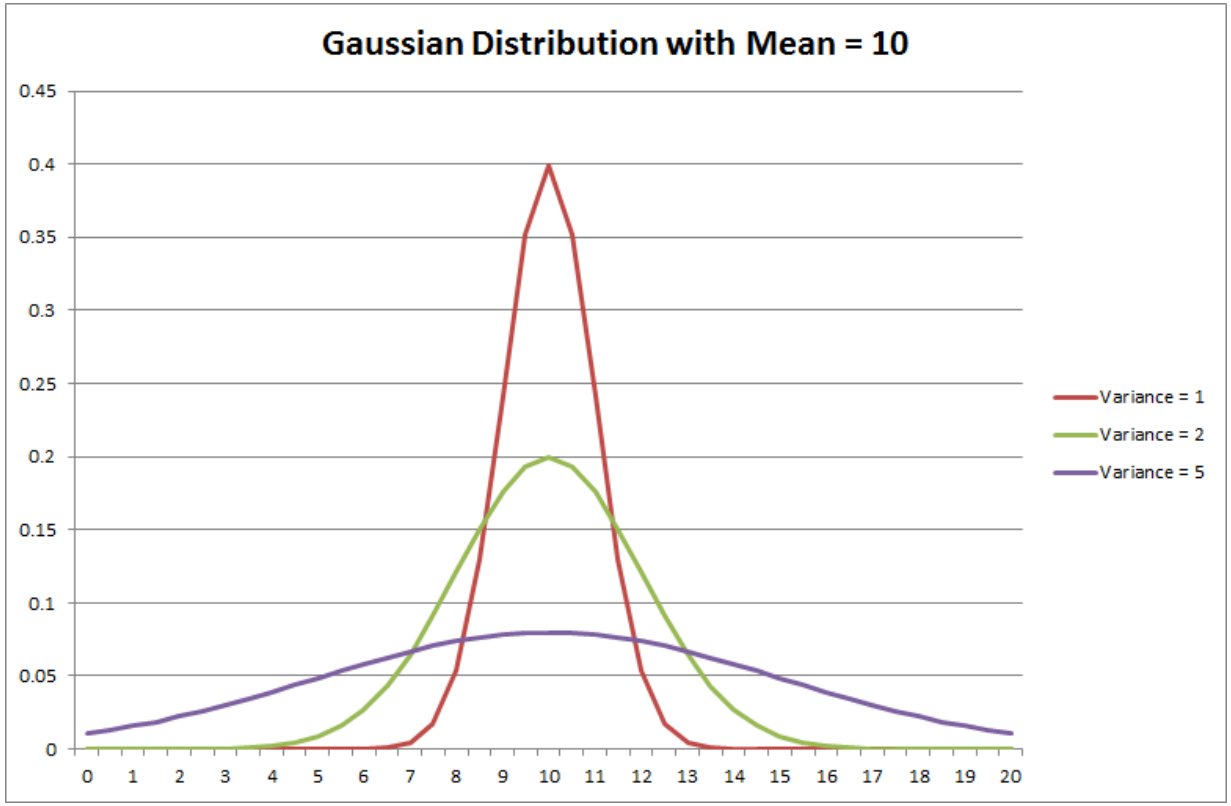
Suryaprakash Kompalli
Associate Professor, INSOFE

CENTRAL TENDENCIES

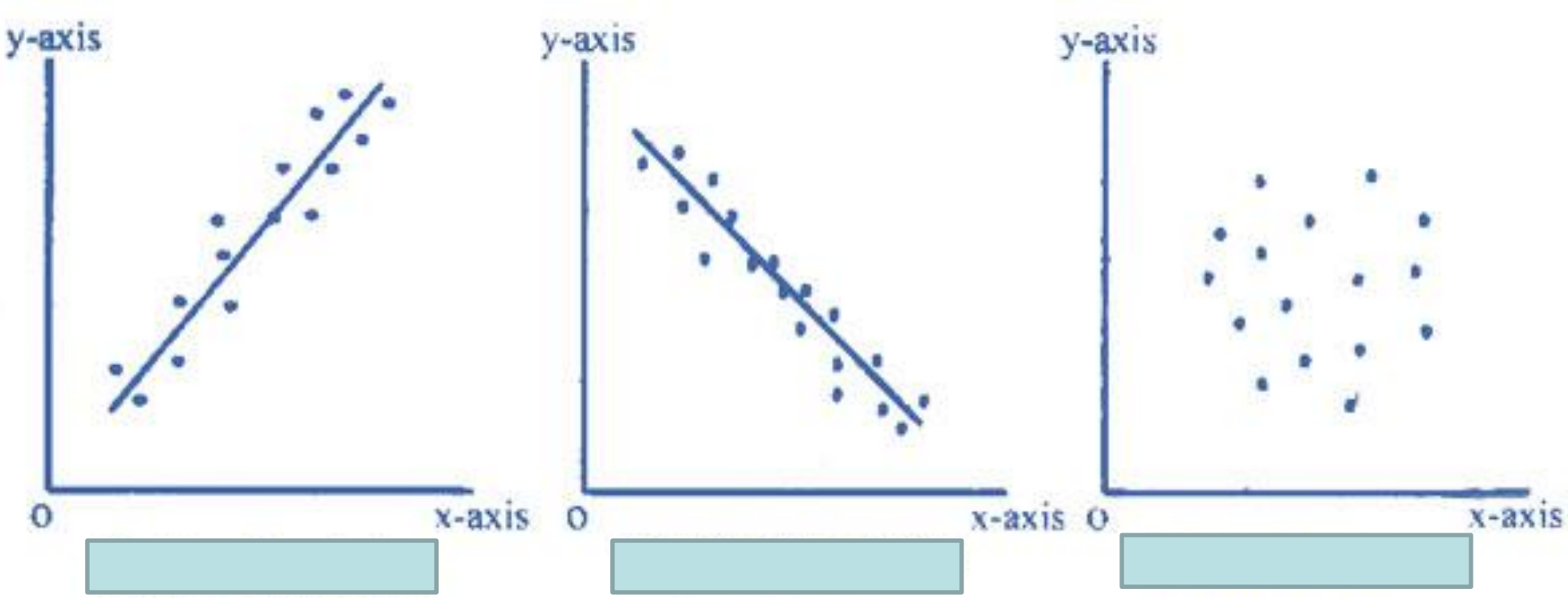
Rigorous analysis of data

- Mean
- Median or 50^{th} percentile
- Mode

Variance (Data spread around mean)



Covariance and correlation



Can you tell how the values are correlated?

What could be the approximate value in these cases?

Analysis of two attributes

- Covariance

$$\sigma_{X,Y} = \sum_{i=1}^{N} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{N}$$

- Correlation

$$\rho_{X,Y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

LET US REDUCE THE DIMENSIONALITY

Multiple tools

- Correlation analysis
  - If there are 100 attributes, we need to check 5000 combinations
    - 100c2: n! / (n-2)!2!
- Principal component analysis

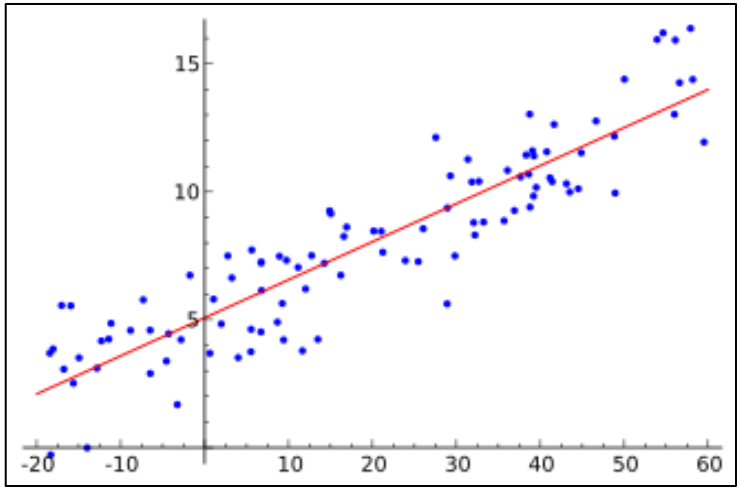
Three steps

- Visualize the data as points and attributes as dimensions
- Identify the **best dimensions** to represent the data
- Remove **unwanted dimensions**

A few basics before PCA

- Linear Regression
  - Fit a line to points

$$f_i = mx_i + b$$



- Measure of a good fit:

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2,$$

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2$$

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}.$$

What does PCA do?

- Given a set of dimensions, and a result

Dimensions, or the “x” part							Result, or the “y” part
age	exp	inc	family	edu	mortgage	ccAvg	loan
45	19	34	3	1	0	18.87937	0
50	24	22	1	3	0	3.929491	0
35	10	81	3	2	104	6.94175	0
34	9	180	1	3	0	104.5955	1
60	30	22	1	3	0	19.25372	0
38	14	130	4	3	134	61.89438	1
42	18	81	4	1	0	31.51252	0
46	21	193	2	3	0	97.62623	1

  - You can try $y = mx + c$ on all features

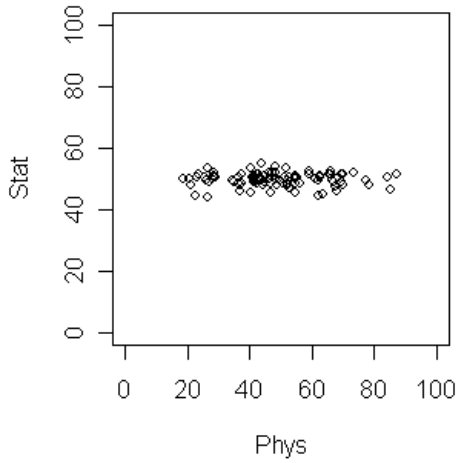
- In our case $|x| = 7$; are all of them needed?

What does PCA do?

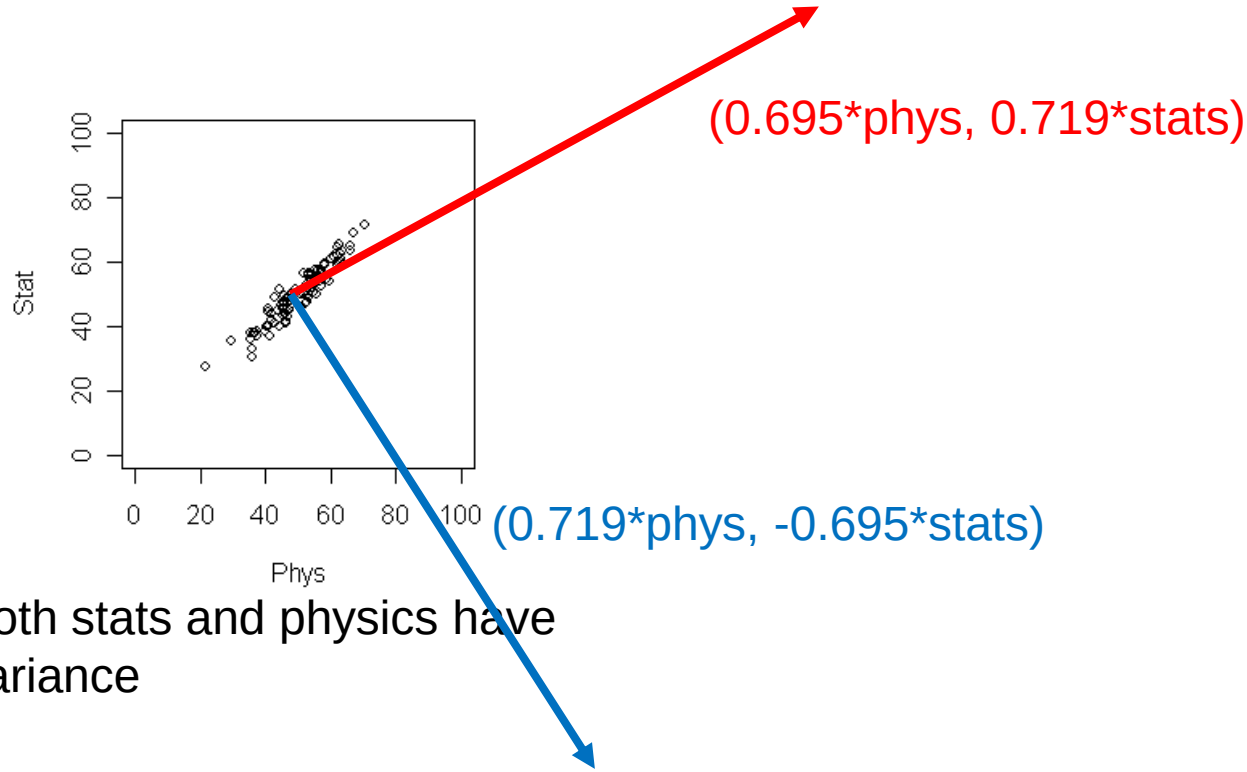
Dimensions, or the "x" part							Result, or the "y" part
age	exp	inc	family	edu	mortgage	ccAvg	loan
45	19	34	3	1	0	18.87937	0
50	24	22	1	3	0	3.929491	0
35	10	81	3	2	104	6.94175	0
34	9	180	1	3	0	104.5955	1
60	30	22	1	3	0	19.25372	0
38	14	130	4	3	134	61.89438	1
42	18	81	4	1	0	31.51252	0
46	21	193	2	3	0	97.62623	1

- PCA will do g(X)
  - $g_1 = a_1 x_1 + a_2 x_2 + a_3 x_3 .... a_7 x_7$
  - $g_2 = b_1 x_1 + b_2 x_2 + b_3 x_3 .... b_7 x_7$
  - So on, till a g_n where n <= 7
- Linear regression between y and g_n will be a good fit

What does PCA do?



Only one parameter physics has significant variance



Both stats and physics have variance

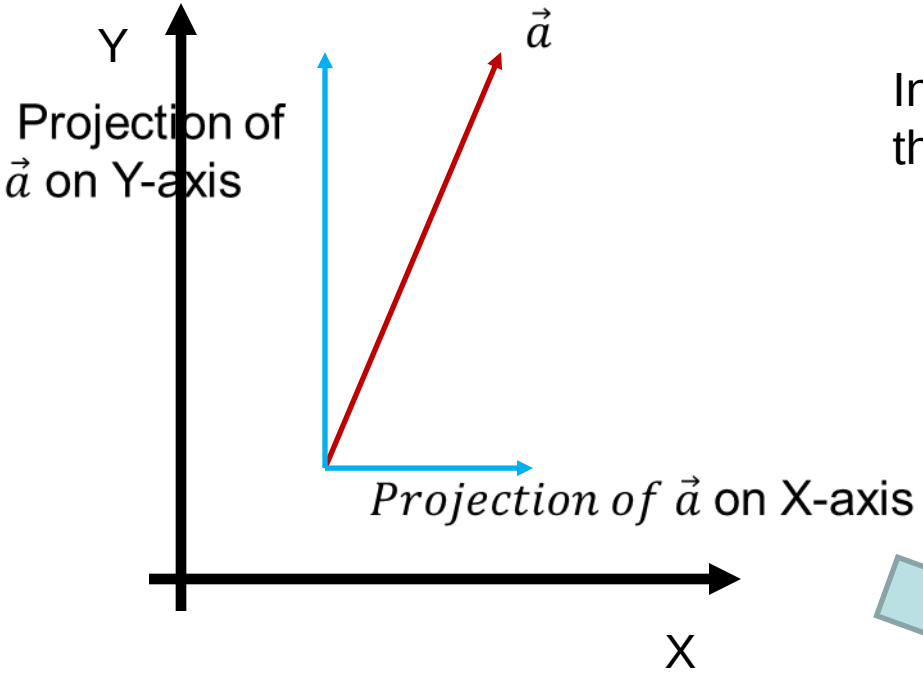
pcaPhysicsStatsMarks.R

http://astrostatistics.psu.edu/su09/lecturenotes/pca.html

Aha!

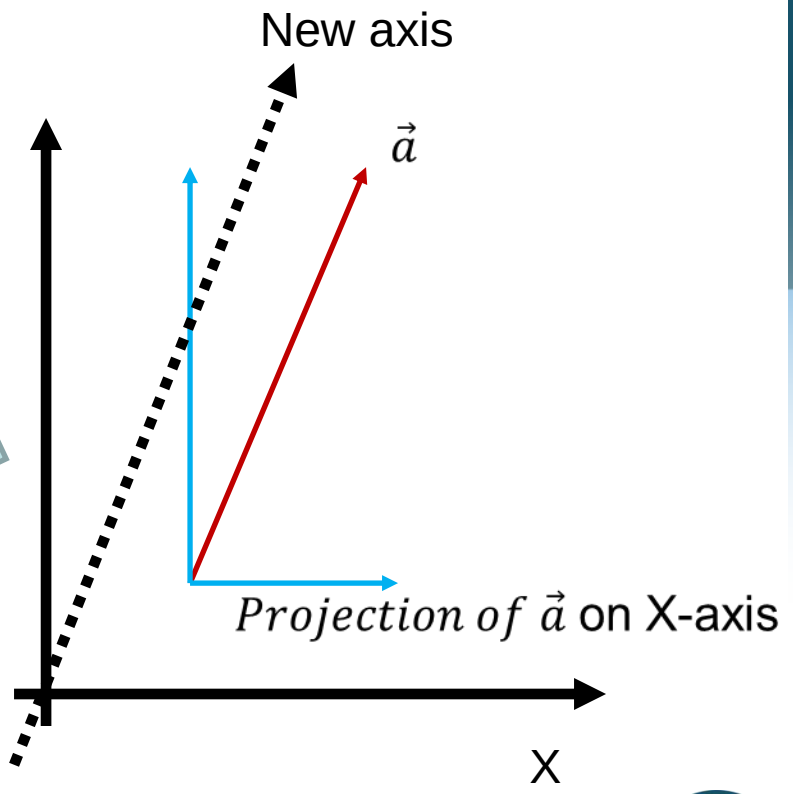
- From the original attributes, the data may look complex and need lot of dimensions
- But, if we change the basis (axis of the features), suddenly the data may need a lot fewer dimensions

Which are those basis?



Think of an axis where projection is maximum !!!

In multiple dimensions, you may get many axes that are somewhat better than your current axes.



Which are those basis

- Covariance matrix

$$C = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix}$$

- Significant members of the Cov matrix!!!

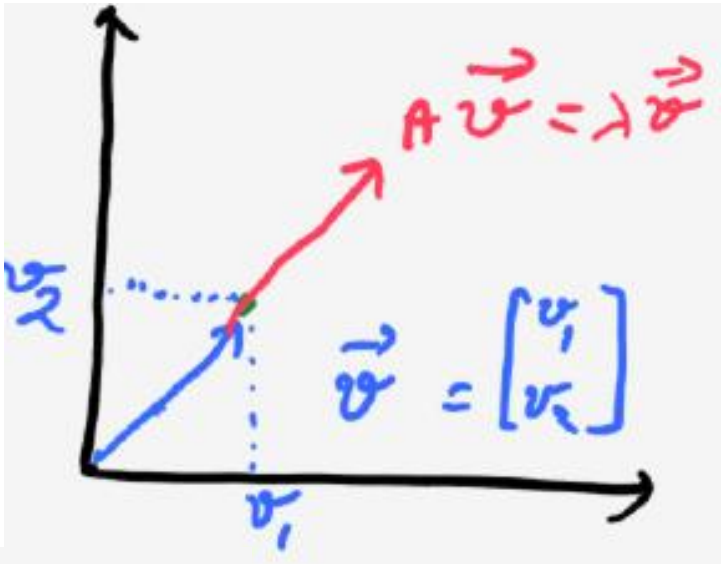
Eigen values of covariance matrix is the principal component

How does it do it?

- Eigen values and Eigen vectors

$$A\vec{v} = \lambda\vec{v}$$

eigen value



$$A\vec{v} - \lambda\vec{v} = 0$$

$$(A - \lambda I_n)\vec{v} = 0$$

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$ if $A = 2 \times 2$.

A is a matrix, $\vec{\vartheta}$ is a unit vector on which A is projected. If A is a covariance matrix, $\vec{\vartheta}$ would represent principal components

http://www.doc.ic.ac.uk/~dfg/ProbabilisticInference/IDAPILecture15.pdf
https://www.math.hmc.edu/calculus/tutorials/eigenstuff/

How does it do it?

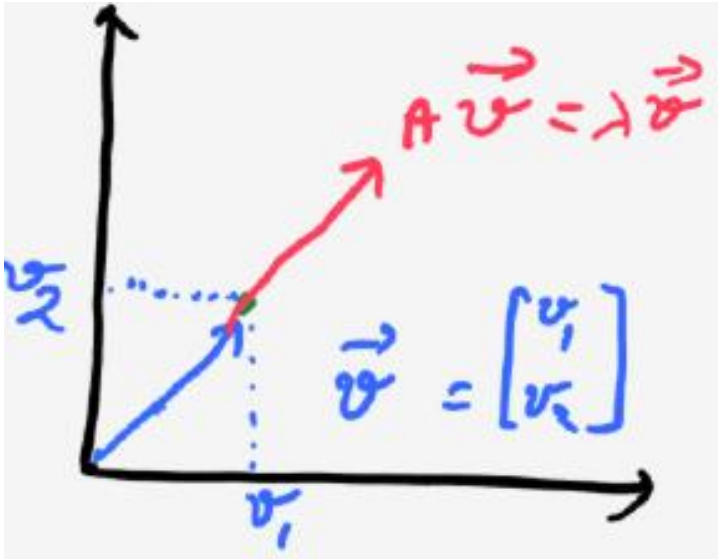
- Eigen values and Eigen vectors

$$A\vec{v} - \lambda\vec{v} = 0$$

$$(A - \lambda I_n)\vec{v} = 0$$

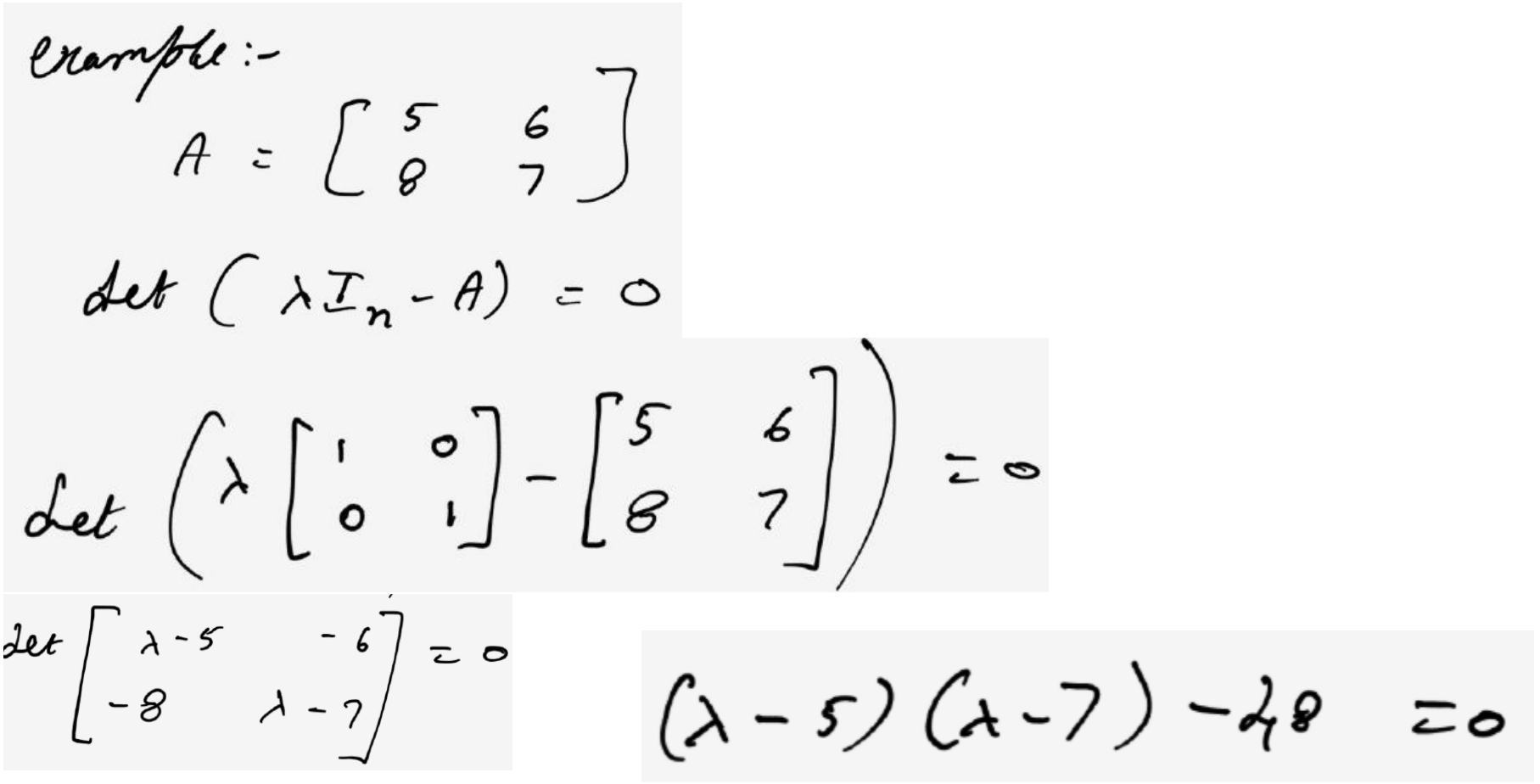
$\vec{v}$ is NON ZERO

$$\det(\lambda I_n - A) = 0$$



How does it do it?

- Eigen values and Eigen vectors



How does it do it?

$$A = \begin{bmatrix} 5 & 6 \\ 8 & 7 \end{bmatrix}$$

- Eigen values:
  - λ is 13 or -1
- Eigen vectors:
  - For 13: 0.6, 0.8
  - For -1: -0.71, 0.71
- Which Eigen value / vector?
  - Larger/smaller?

$$A\vec{v} - \lambda\vec{v} = 0$$

$$(A - \lambda I_n)\vec{v} = 0$$

How to get Eigen Vector?

$$For\ 13: \begin{bmatrix} -8 & 6 \\ 8 & -6 \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

Or $-8v_1 + 6v_2 = 0$

If $v_2 = t, v_1 = 0.75t$

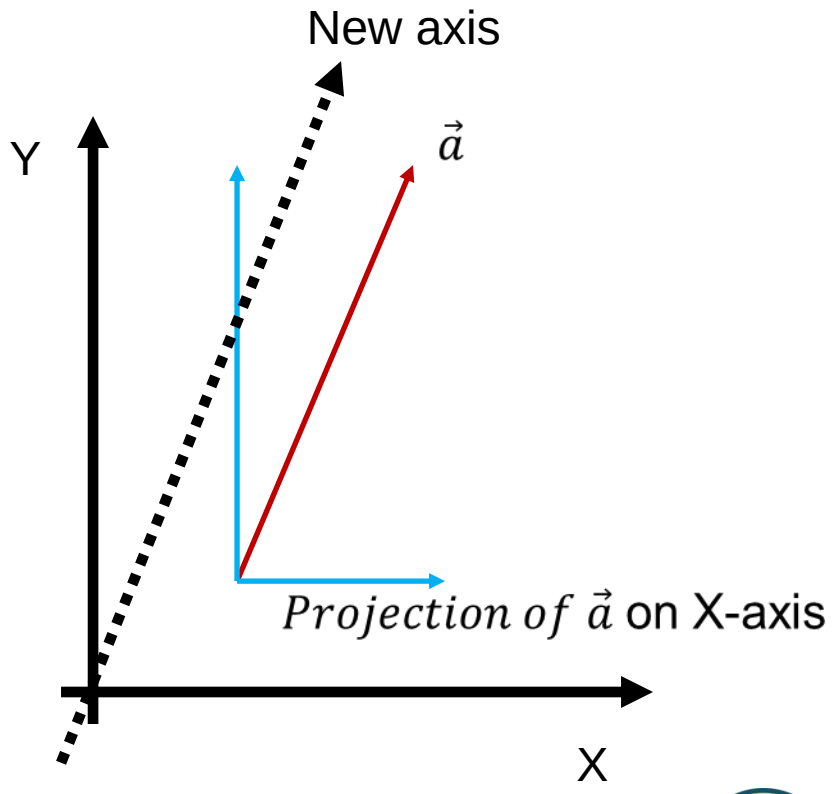
Hence $\vec{v} = (0.75,1)$ The unit vector along $\vec{v}$ = (0.6,0.8)

$A\vec{v} = \lambda\vec{v}$ → eigen value

How does it do it?

- Which Eigen value / vector?
  - Larger/smaller?
- Hint: Larger Eigen value means projection is longer
  - Longer projection means more variance is captured

$A\vec{v} = \lambda\vec{v}$ — λ: eigen value



Which are those basis – example

corIris:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.0000000	-0.1175698	0.8717538	0.8179411
Sepal.Width	-0.1175698	1.0000000	-0.4284401	-0.3661259
Petal.Length	0.8717538	-0.4284401	1.0000000	0.9628654
Petal.Width	0.8179411	-0.3661259	0.9628654	1.0000000

pcIris$loadings

	Comp.1	Comp.2	Comp.3	Comp.4
Sepal.Length	0.361	-0.657	-0.582	0.315
Sepal.Width		-0.730	0.598	-0.320
Petal.Length	0.857	0.173		-0.480
Petal.Width	0.358		0.546	0.754

eigen(covIris):

[,1]	[,2]	[,3]	[,4]
0.36138659	-0.65658877	-0.58202985	0.3154872
-0.08452251	-0.73016143	0.59791083	-0.3197231
0.85667061	0.17337266	0.07623608	-0.4798390
0.35828920	0.07548102	0.54583143	0.7536574

summary(pcIris):

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	2.0494032	0.49097143	0.27872586	0.153870700
Proportion of Variance	0.9246187	0.05306648	0.01710261	0.005212184
Cumulative Proportion	0.9246187	0.97768521	0.99478782	1.000000000

pcaPhysicsStatsMarks.R

Quasar Dataset

charRecog.R
pcaPhysicsStatsMarks.R

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10	Comp.11
R.A.							0.102	0.987			
Dec.	-0.697	-0.711									
z				0.241	-0.247	-0.301			0.787		
u_mag				0.247	-0.731	0.168	0.519		-0.162		
sig_u							0.137			-0.150	
g_mag					-0.333	0.184	-0.298		-0.239	0.554	0.211
sig_g											
r_mag					-0.215	0.176	-0.449		0.203	0.369	
sig_r											
i_mag					-0.199	0.172	-0.397		0.198	-0.250	-0.277
sig_i											
z_mag					-0.174	0.140	-0.458		-0.288	-0.659	0.131
sig_z											
Radio			0.987	-0.118							
X.ray			-0.105	-0.822	-0.399	-0.383					
J_mag	0.426	-0.420									0.555
sig_J											
H_mag	0.411	-0.403									
sig_H											
K_mag	0.390	-0.384									-0.705
sig_K											-0.157
M_i				-0.407		0.786	0.196		0.344		

12 out of 22 features are not needed to describe 97.77% of data

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	11.9895830	9.5851787	3.65511626	2.0697063	1.78656841
Proportion of Variance	0.5549115	0.3546631	0.05157247	0.0165361	0.01232125
Cumulative Proportion	0.5549115	0.9095746	0.96114706	0.9776832	0.99000441

http://astrostatistics.psu.edu/su09/lecturenotes/pca.html

Disadvantages

- Explicability

SVD

SINGULAR VALUE DECOMPOSITION

SVD

- Decompose matrix A in the form:

$-A_{mn} = U_{mm} D_{mn} V_{nn}^T$

A has m Rows, n columns
U has m rows and columns etc.

  - U: Columns are orthonormal eigen vectors of AA^T
  - V: Columns are orthonormal eigen vectors of $A^T A$
  - D: Diagonal matrix containing sqr roots of Eigen values from U or V in decreasing order

https://www.ling.ohio-state.edu/~kbaker/pubs/Singular_Value_Decomposition_Tutorial.pdf

SVD

- Where is this useful?

		Customers			
		Cust 1	Cust 2	Cust 3	Cust 4
Number of times movie was seen	Matrix	2			2
	WallE	1	3	4	
	Notebook		5		
	Planes		3	3	1
	Aviator	2	1		
	Aliens	1			
	Predator		1		1
	Interstellar	3			2

$$A_{mn} = U_{mm} D_{mn} V_{nn}^T$$

U: Eigen vectors of AA^T, i.e. movies that co-occur

V: Eigen vectors of $A^T A$, i.e. customers with similar movie watching pattern

$d
[1] 7.754896 4.863228 3.418626 1.877123

$v

	[,1]	[,2]	[,3]	[,4]
[1,]	-0.1831444	**-0.82064480**	0.05248445	0.538744482
[2,]	-0.8215331	0.22656669	0.52300014	0.014890526
[3,]	-0.5232137	0.06423226	-0.84977295	0.002762113
[4,]	-0.1333751	**-0.52065715**	0.04002726	-0.842333085

Similar customers:

1 and 4

$u

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]
[1,]	-0.08163090	**-0.55160972**	0.05412216	-0.32346157	-0.23135900	-0.11911823	-0.16340417	-0.69791795
[2,]	-0.61130391	0.02384925	**-0.51997697**	0.31668914	-0.38423360	-0.19048329	0.26424259	-0.03643070
[3,]	-0.52968673	0.23293858	0.76492733	0.03966316	-0.23685755	-0.02159801	-0.14392224	0.03468454
[4,]	-0.53741731	0.07232639	**-0.27504942**	-0.42052391	0.51231146	0.25397773	-0.35232345	0.04857427
[5,]	-0.15317058	-0.29090202	0.18369046	0.58194343	0.65085839	-0.16890413	0.12957274	-0.22490020
[6,]	-0.02361662	-0.16874487	0.01535250	0.28700540	-0.15658134	0.91975627	0.06614412	-0.11625523
[7,]	-0.12313617	-0.06047227	0.16469405	-0.44080355	0.14919576	0.08641090	0.85428102	0.01504681
[8,]	-0.10524751	**-0.72035459**	0.06947465	-0.03645617	-0.09939462	-0.05107608	-0.08757462	0.66610741

Movies seen together

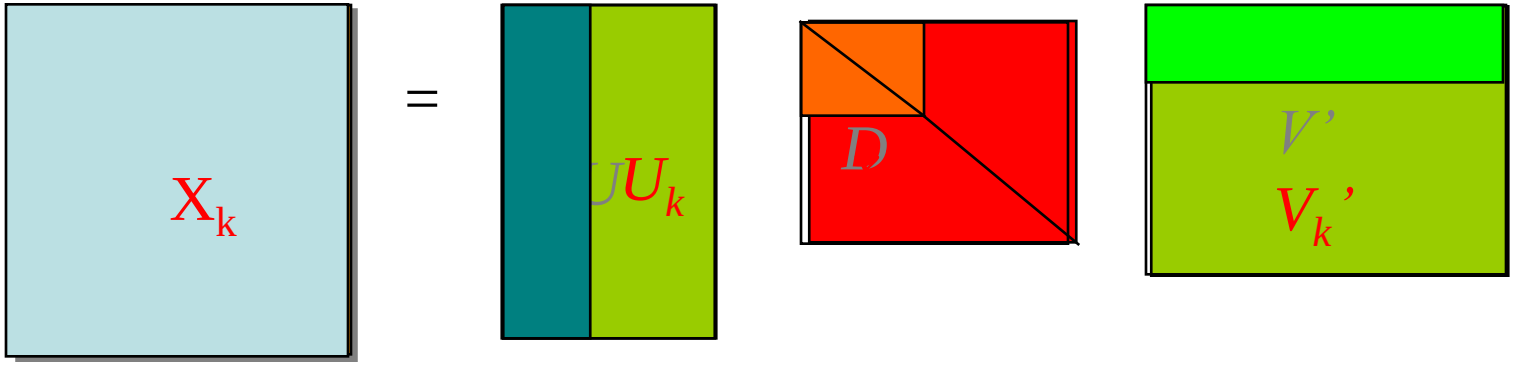
Matrix and Interstellar

WallE and Planes

https://www.igvita.com/2007/01/15/svd-recommendation-system-in-ruby/

SVD

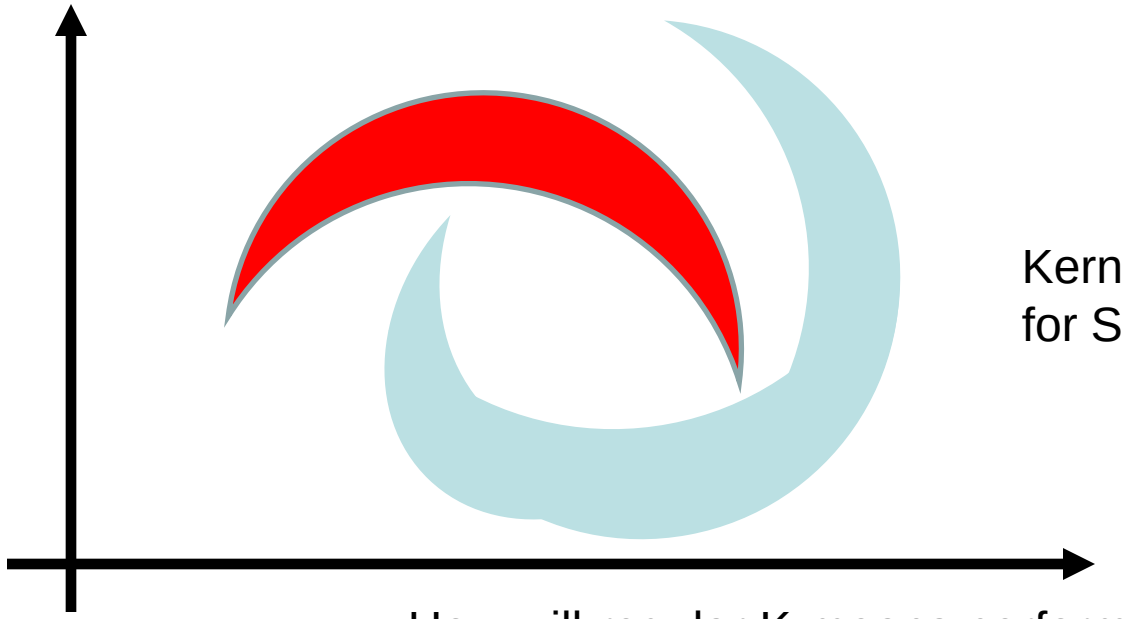
- Other uses:
  - Text Analysis: Columns are documents, rows count of words in each document
  - Image processing: Columns are character classes, rows are count of features



Only a part of U, D, and V is retained

Spectral Clustering

- Do dimensionality reduction via PCA or SVD
- Apply K-means clustering on reduced dimensions
- Note: You are not just "reducing" dimensions, you are mapping them to different space



Kernlab, clusterSim R commands for Spectral clustering

How will regular K-means perform on above data?

Inspire...Educate...Transform.

International School of Engineering

Plot 63/A, 1st Floor, Road # 13, Film Nagar, Jubilee Hills, Hyderabad - 500 033

For Individuals: +91-9502334561/63 or 040-65743991

For Corporates: +91-9618483483

Web: http://www.insofe.edu.in

Facebook: https://www.facebook.com/insofe

Twitter: https://twitter.com/Insofeedu

YouTube: http://www.youtube.com/InsofeVideos

SlideShare: http://www.slideshare.net/INSOFE

LinkedIn: http://www.linkedin.com/company/international-school-of-engineering