Dimensionality Reduction: Principal Component Analysis (PAC)

Dr. Yuzana Win (Nagasaki University, Japan)
Lecturer
Department of Computer Engineering and Information Technology

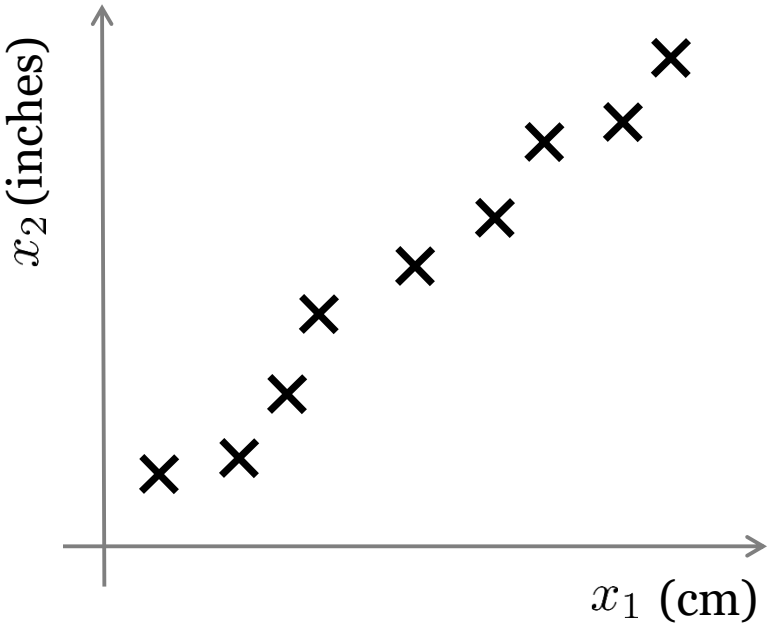
Lecture Objectives

- To introduce
  - Dimensionality Reduction
  - What is Principle Components Analysis?
  - How does Principal Component Analysis works and why would we used?
  - Advantages of using PCA

Dimensionality Reduction

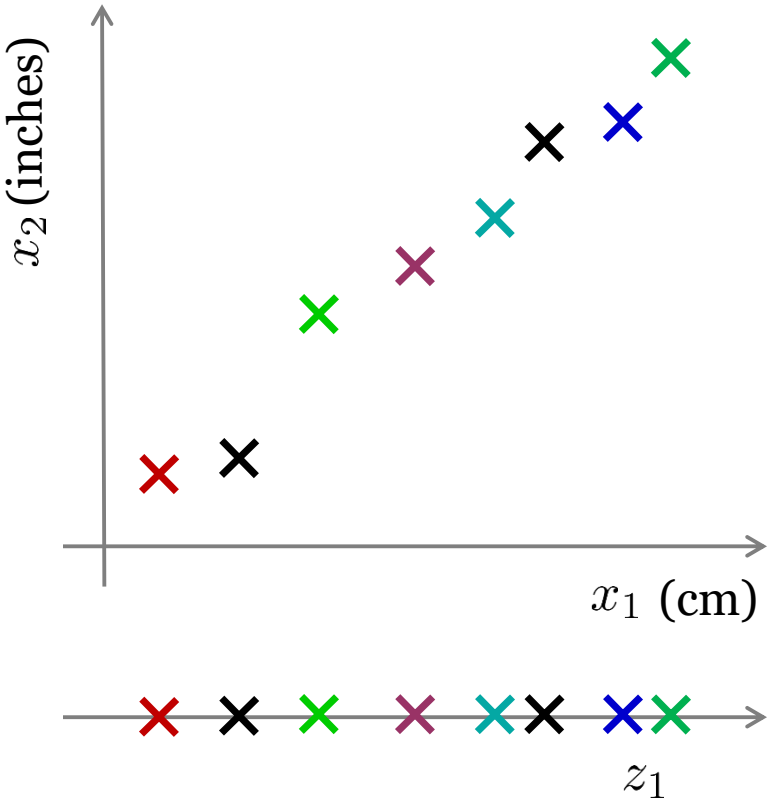
- Dimensionality reduction is to
  - simplify complex high-dimensional data
  - summarize data with a lower dimensional real valued vector
  - Provide a framework for interpretability of the results

Dimensionality Reduction



Reduce data from 2D to 1D

Dimensionality Reduction

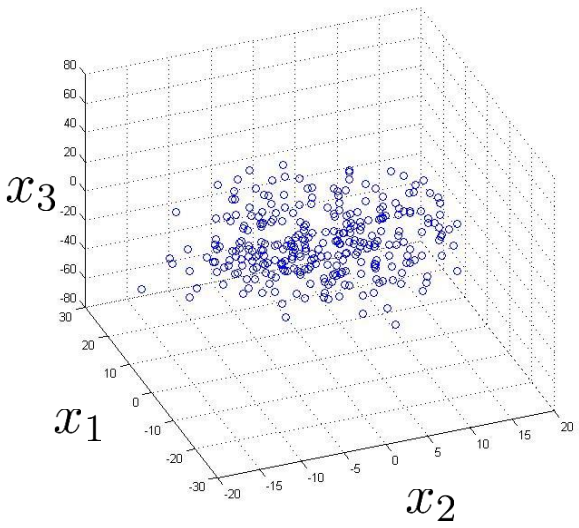


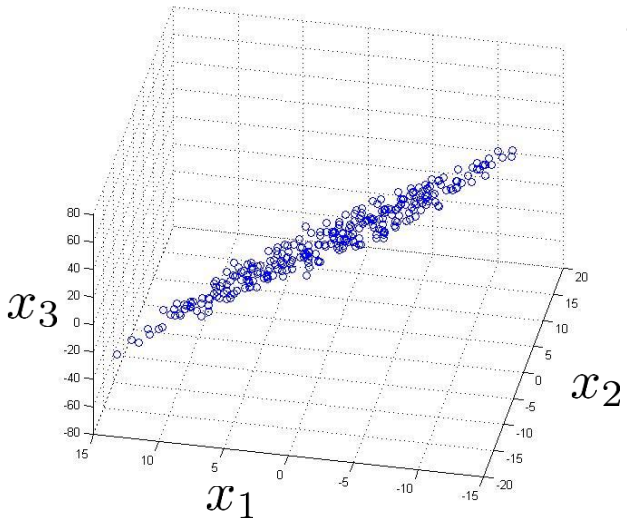
Reduce data from 2D to 1D

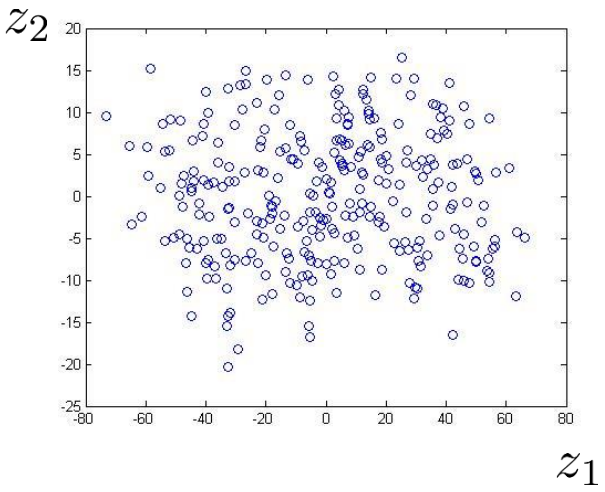
$$x^{(1)} \rightarrow z^{(1)}$$
$$x^{(2)} \rightarrow z^{(2)}$$
$$\vdots$$
$$x^{(m)} \rightarrow z^{(m)}$$

Dimensionality Reduction

Reduce data from 3D to 2D







Dimensionality Reduction Methods

- PCA (Principle Component Analysis)
- SVD (Singular Value Decomposition)
- LSA (Latent Semantic Analysis)
- LDA (Latent Dirichlet Allocation)

What is Principle Component Analysis?

- a dimension-reduction tool
- Reduce a large set of variables to a small set

  (Transform the data to a linearly uncorrelated new coordinate system)
- Coordinate system components are called Principal Components
- The new coordinate system has the same dimensionality or a lower dimensionality compared to the original data

Principle Component Analysis Ideas

- PCA is a technique that can be used to simplify a dataset
- For a matrix of *m* samples x *n* genes, create a new covariance matrix of size *n* x *n*
- Transform some large number of variables into a smaller number of uncorrelated variables called principal components (PCs)
- developed to capture as much of the variation in data as possible
  - greatest variance by any projection of the data set comes to lie on the first axis (then called the first principal component),
  - the second greatest variance on the second axis, and so on.

Why would Principle Component Analysis used?

- Can be used to:
  - Reduce number of dimensions in data
  - Find patterns in high-dimensional data
  - Visualize data of high dimensionality

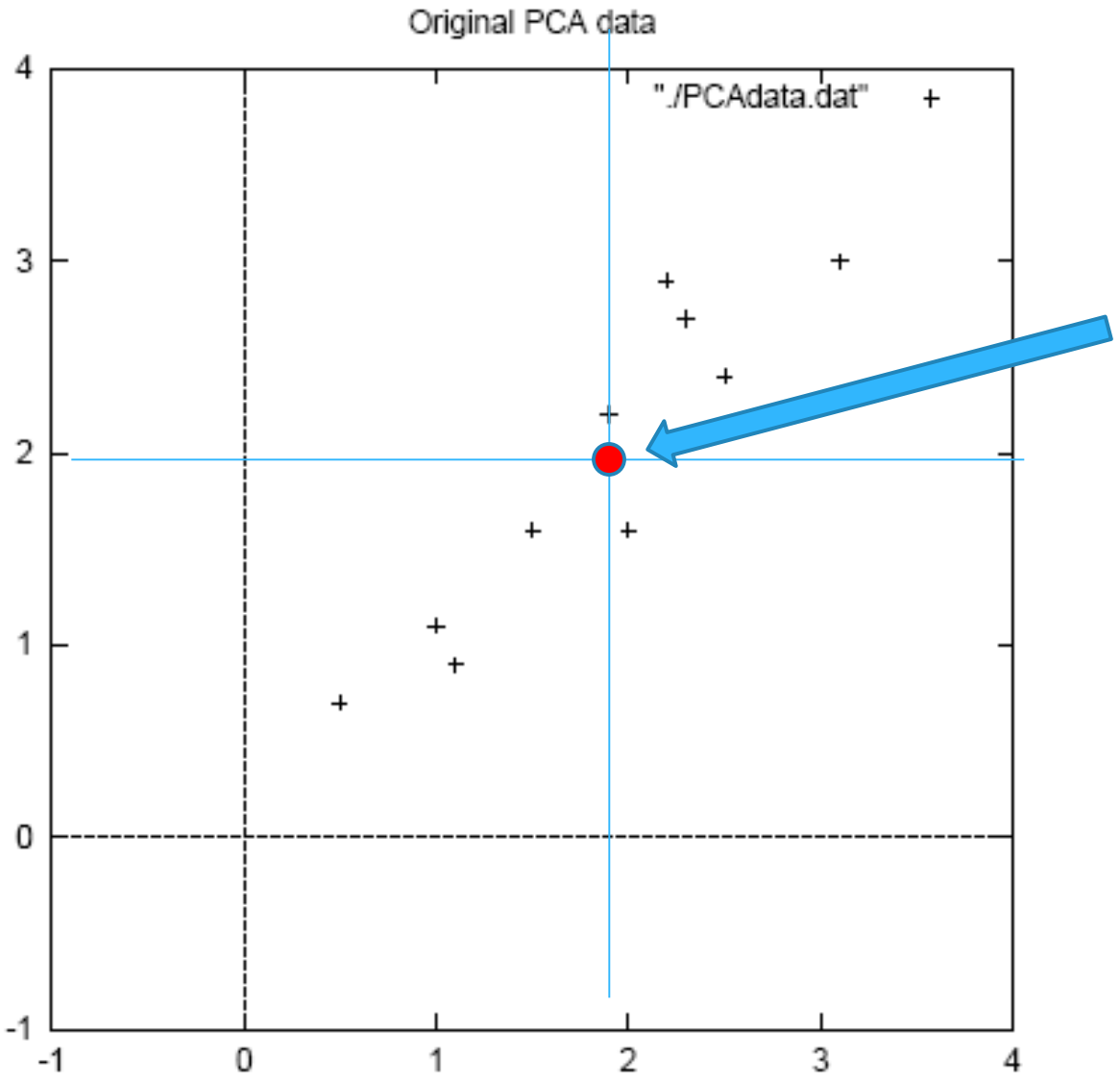
- Example applications:
  - Face recognition
  - Image compression
  - Gene expression analysis

PCA process –STEP 1

DATA:

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

PCA process –STEP 1



Means of two data variables

PCA process (Step-1)

- Subtract the means from the corresponding data component to re-center the data set
- Re-construct the scatterplot to view
- Write ‘adjusted’ data as a matrix to calculate the covariance matrix
- Note that the ‘adjusted’ data set will have means zero

PCA process (Step-1)

DATA:

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Mean = Total/Number

X* = 1.81

Y* = 1.91

ZERO MEAN DATA:

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

PCA process (Step-2)

- Compute the sample variance-covariance matrix C

$$C = \frac{1}{N-1}(x-\bar{x})'(x-\bar{x}) = \frac{1}{N-1}X'X$$

- Use covariance matrix for PCA if variables are of the same scale or unit
- Use correlation matrix for PCA if variables are of the different scale or unit

PCA process (Step-2)

- Calculate the covariance matrix

$$\text{cov}(C) = \begin{pmatrix} 0.616555556 & 0.615444444 \\ 0.615444444 & 0.716555556 \end{pmatrix}$$

- since the non-diagonal elements in this covariance matrix are positive, we should expect that both the x and y variable increase together.

PCA process (Step-3)

- Calculate the eigenvectors and eigenvalues of the covariance matrix

$$\text{eigenvalues} = \begin{pmatrix} 1.2840 \\ 0.0490 \end{pmatrix}$$

$$\text{eigenvectors} = \begin{pmatrix} 0.678 & 0.735 \\ 0.735 & -0.678 \end{pmatrix}$$

PCA process (Step-3)

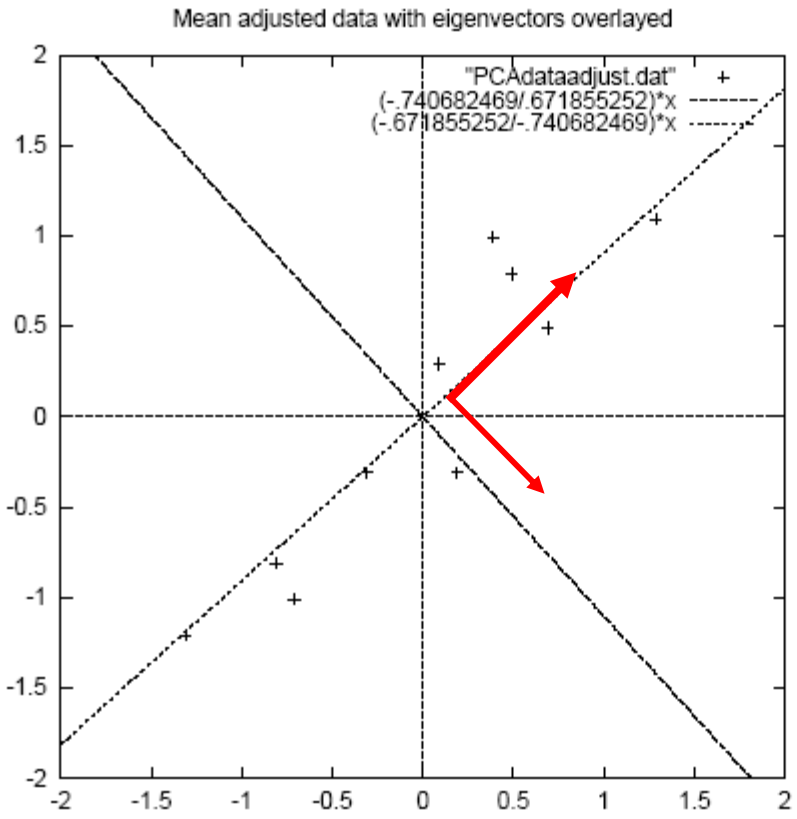


Figure 3.2: A plot of the normalised data (mean subtracted) with the eigenvectors of the covariance matrix overlayed on top.

- eigenvectors are plotted as diagonal dotted lines on the plot.
- Note they are perpendicular to each other.
- Note one of the eigenvectors goes through the middle of the points, like drawing a line of best fit.
- The second eigenvector gives us the other, less important, pattern in the data, that all the points follow the main line, but are off to the side of the main line by some amount.

PCA process (Step-3)

- Total sample variance = sum of eigenvalues

  0.6166 + 0.7166 = 1.2840 + 0.0490

  1.333 = 1.333

- The first eigenvector will go through the middle of the data points, as if it is the line of best fit
- The second eigenvector will the less important pattern

(all the data points follow the main line, but are off to the side of the main line by some amount)

Variable	Eigenvector-1	Eigenvector-2
X1 x2	0.678 0.735	0.735 -0.678
Eigenvalues	1.2840	0.0490
% of total variance	1.2840 / 1.333 = 96 %	0.0490 / 1.333 = 37 %

PCA process (Step-4)

- Choose the components and form the eigenvector matrix V
- By ordering the eigenvectors according to the eigenvalues, this gives the components in order of their significance.
- The eigenvector with the highest eigenvalue is the principal component.
- The components of lesser significance can be ignored, so as to reduce the dimensions of the data set.
- Transformation matrix V = $\begin{pmatrix} 0.678 \\ 0.735 \end{pmatrix}$

PCA process (Step-5)

- Derive the new data set by taking Y = XV
- Basically we have transformed our data so that it is expressed in terms of the patterns between them, where the patterns are the lines that most closely describe the relationships between the data.
- $Y = \begin{pmatrix} 2.5 & 2.4 \\ 0.5 & 0.7 \\ . & . \\ . & . \\ 1.1 & 0.9 \end{pmatrix} \begin{pmatrix} 0.678 \\ 0.735 \end{pmatrix} = \begin{pmatrix} 3.459 \\ 0.854 \\ . \\ . \\ 1.407 \end{pmatrix}$

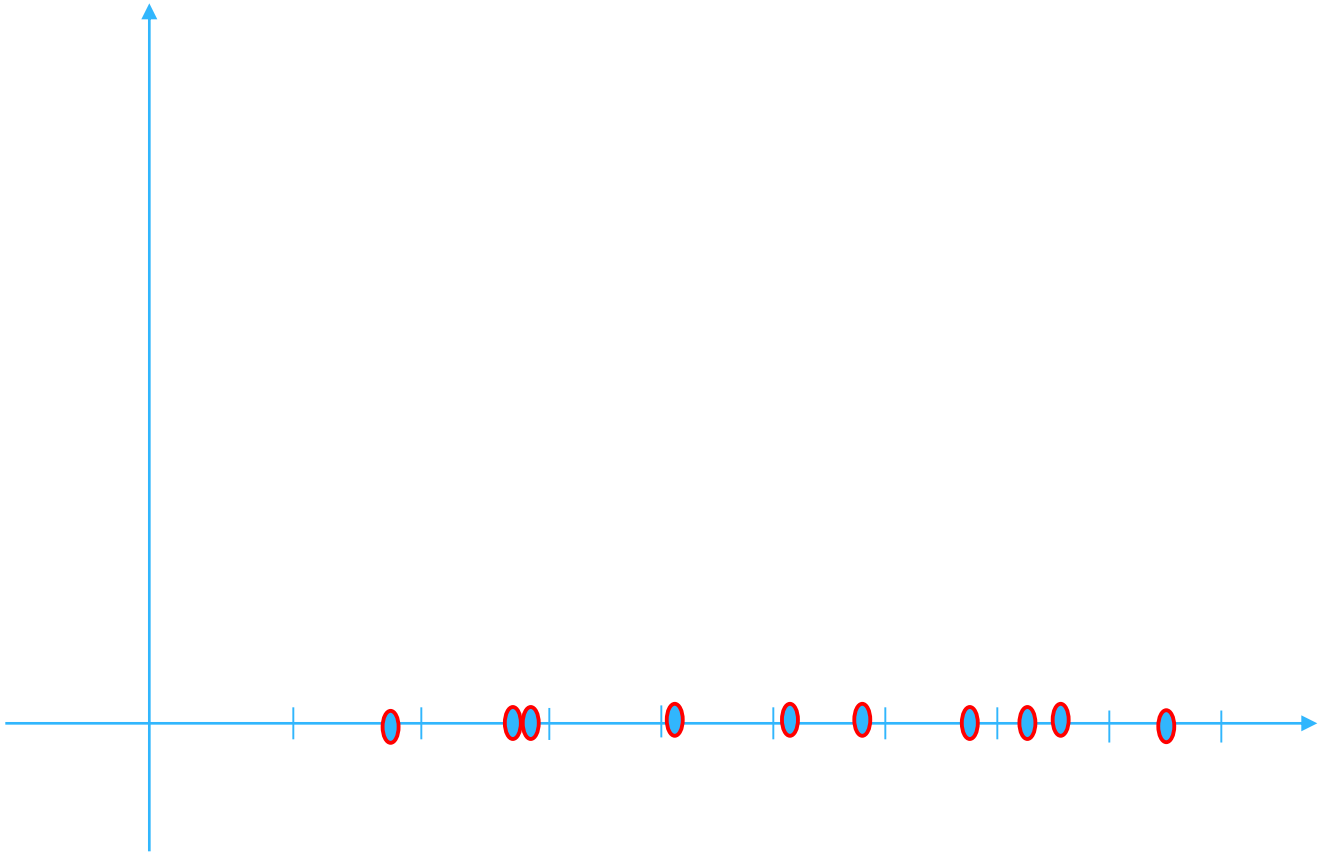
So, Y = 0.678 (x1) + 0.735 (x2)

PCA process (Step-5)

- FinalData is the final data set, with data items in columns, and dimensions along rows.
- What will this give us?
  - It will give us the original data *solely in terms of the vectors we chose.*
- We have changed our data from being in terms of the axes x and y , and now they are in terms of our 2 eigenvectors.

Reconstruction of original Data

$$x = \begin{pmatrix} 3.459 \\ 0.854 \\ 3.623 \\ 2.905 \\ 4.306 \\ 3.544 \\ 2.532 \\ 1.486 \\ 2.193 \\ 1.407 \end{pmatrix}$$



Advantages of PCA

- Reduce the duplicate data
- Reduce the risk of over fitting
- Reduce the computational complexity
- Improve training time

References

- Jiawei Han and Micheline Kamber. *Data Mining - Concepts and Techniques*. MorganKaufmann Publishers, 2001
- Pang-Ning Tan, Michael Steinbach, Vipin Kumar:*Introduction to Data Mining*, Addison-Wesley

Next Week Lecture

- Model Evaluation and Improvement

Thank You