國立政治大學資訊管理學系所
Department of Management Information Systems, NCCU

DSFTA-07: Principal Components Analysis

Prof. Shun-Wen Hsiao

Spring 2017

hsiaom@ncc.edu.tw

PCA

- PCA is concerned with
  - explaining the variance-covariance structure of a set of variables
  - through a few linear combinations of these variables.

- General objectives
  - Data reduction
  - Classification

Step by step PCA

1. Get original data
2. Standardize the data
3. Calculate the covariance matrix
4. Calculate the eigenvectors and eigenvalues of the covariance matrix
5. Choosing components and forming a feature vector
6. Deriving the new data set

Principal Components Analysis

Feature that has (relative) larger variance is more valuable to distinguish different instances.

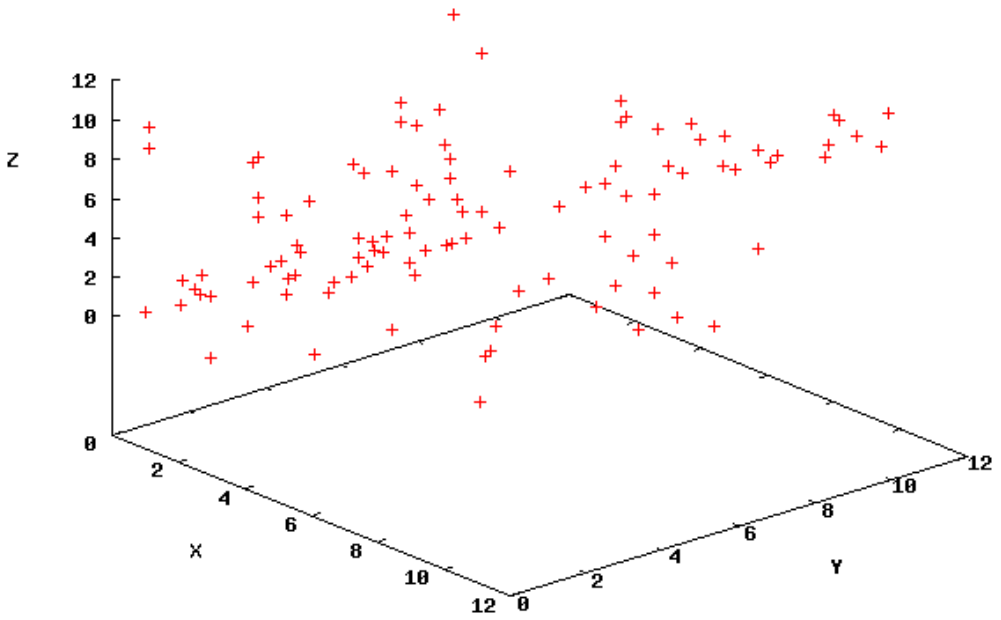
	W	X	Y	Z
instance 1				
instance 2				
instance 3				
instance 4				
instance 5				
…				
Average Variance	V_W	V_X	V_Y	V_Z

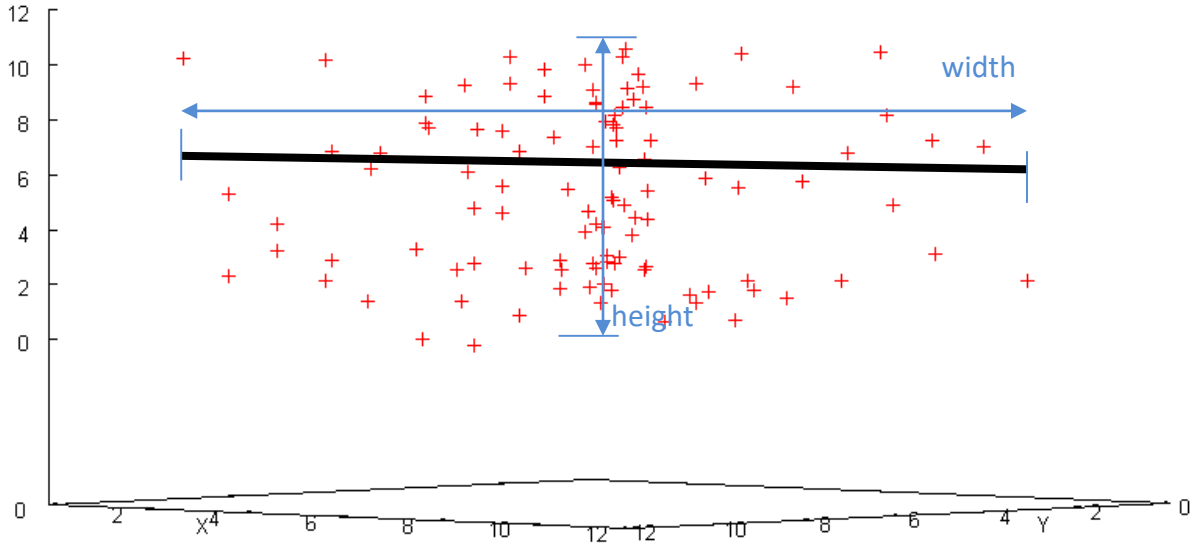
PCA redistributes features' variances

$PC1 = a_1*W+b_1*X+c_1*Y+d_1*Z$
$PC2 = a_2*W+b_2*X+c_2*Y+d_2*Z$
$PC3 = a_3*W+b_3*X+c_3*Y+d_3*Z$
$PC4 = a_4*W+b_4*X+c_4*Y+d_4*Z$

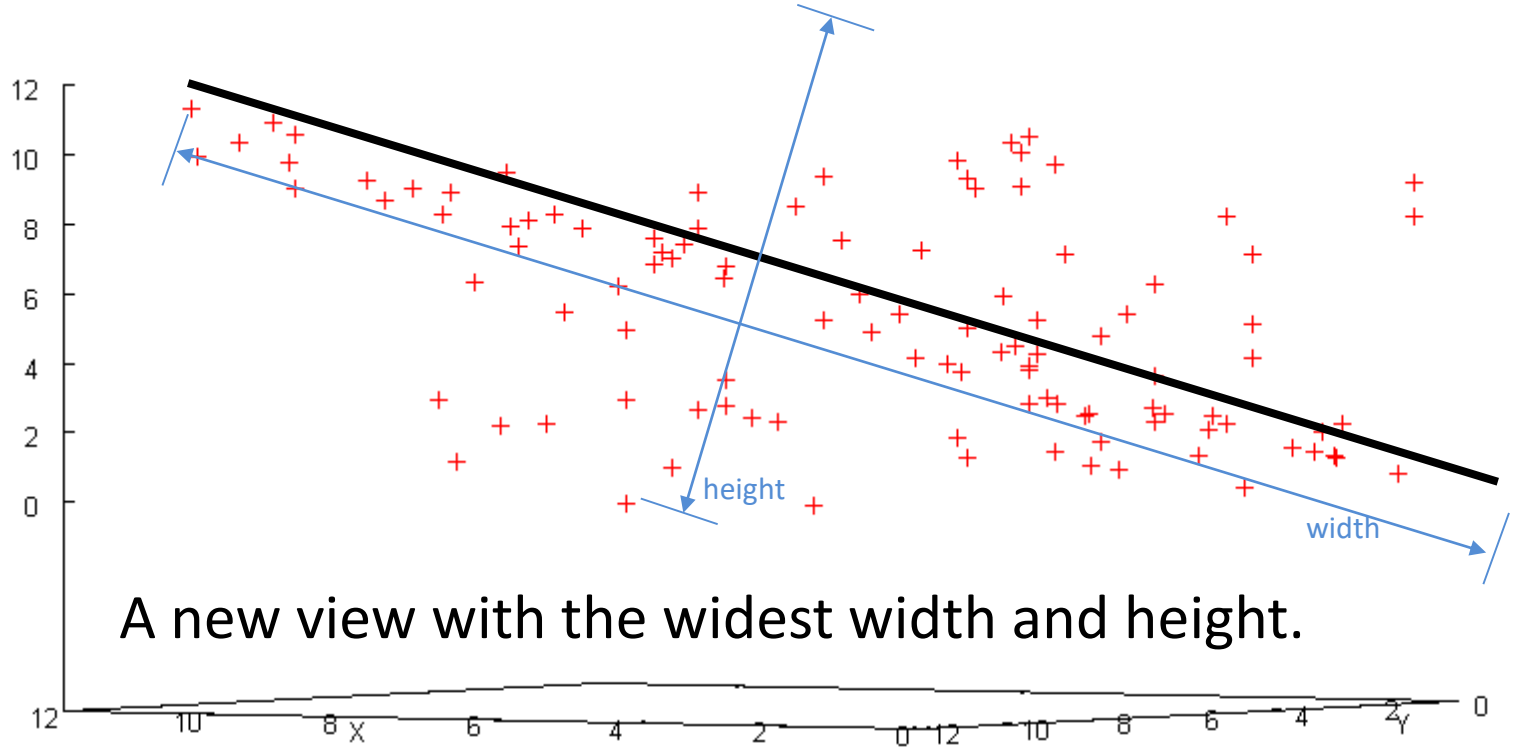
$$\mathbf{V_W+V_X+V_Y+V_Z = V_{PC1}+V_{PC2}+V_{PC3}+V_{PC4}}$$
$$\mathbf{V_{PC1} >= V_{PC2} >= V_{PC3} >= V_{PC4}}$$

	PC1	PC2	PC3	PC4
instance 1				
instance 2				
instance 3				
instance 4				
instance 5				
…				
Average Variance	V_{PC1}	V_{PC2}	V_{PC3}	V_{PC4}





PCA would like to choose a widest width, because with this point of view, the widest width makes us have a better chance to differentiate these data points.

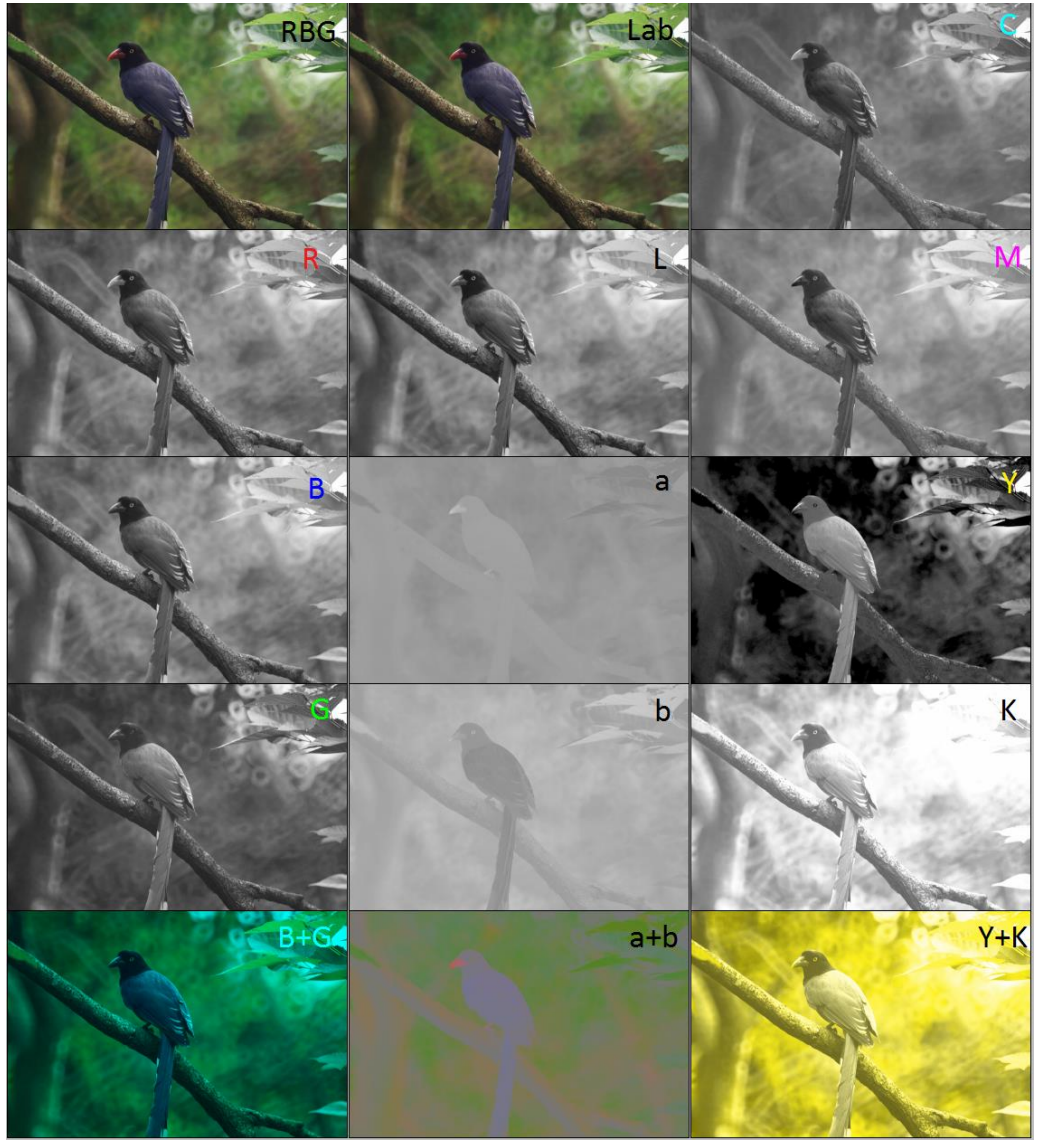


A new view with the widest width and height.

PCA

- Features are used to describe data and can distinguish a data point from another.
- For a set of data points, a feature with zero variance is meaningless.
  - A feature with larger variance is more meaningful.
- Generally, PCA combines original features (e.g., X, Y, …) to populate new features (e.g., PC1, PC2, …).
  - The variances of the original features (i.e., var(X), var(Y), …) are redistributed to the variances of new features (i.e., var(PC1), var(PC2), …) in order.
  - A data point can be described by original features or new features without losing any information.
  - But for the convenience of analyzing, a small set of features having large aggregated variance is better than a larger set of features having the same variance.
    - For example, color can be represented as (R, G, B), (L, a, b), (C, M, Y, K), or (H, S, V).

Change viewpoint (Rotate data)

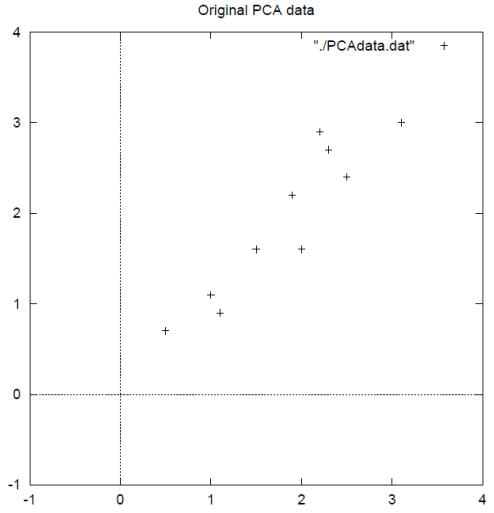


PCA, example of calculation 1/3

(1) Get original data

Data =

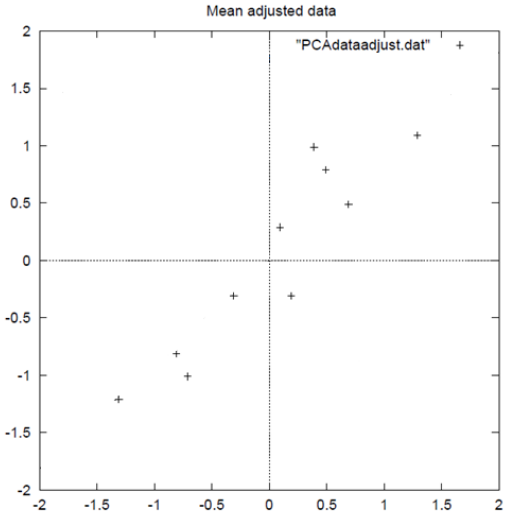
x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9



(2) Standardization
- let mean = 0
- variances are not changed.

DataAdjust =

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, example of calculation 2/3

(3) Calculate covariance matrix

DataAdjust =

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

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

$$cov = \begin{pmatrix} cov(X,X) & cov(X,Y) \\ cov(Y,X) & cov(Y,Y) \end{pmatrix}$$

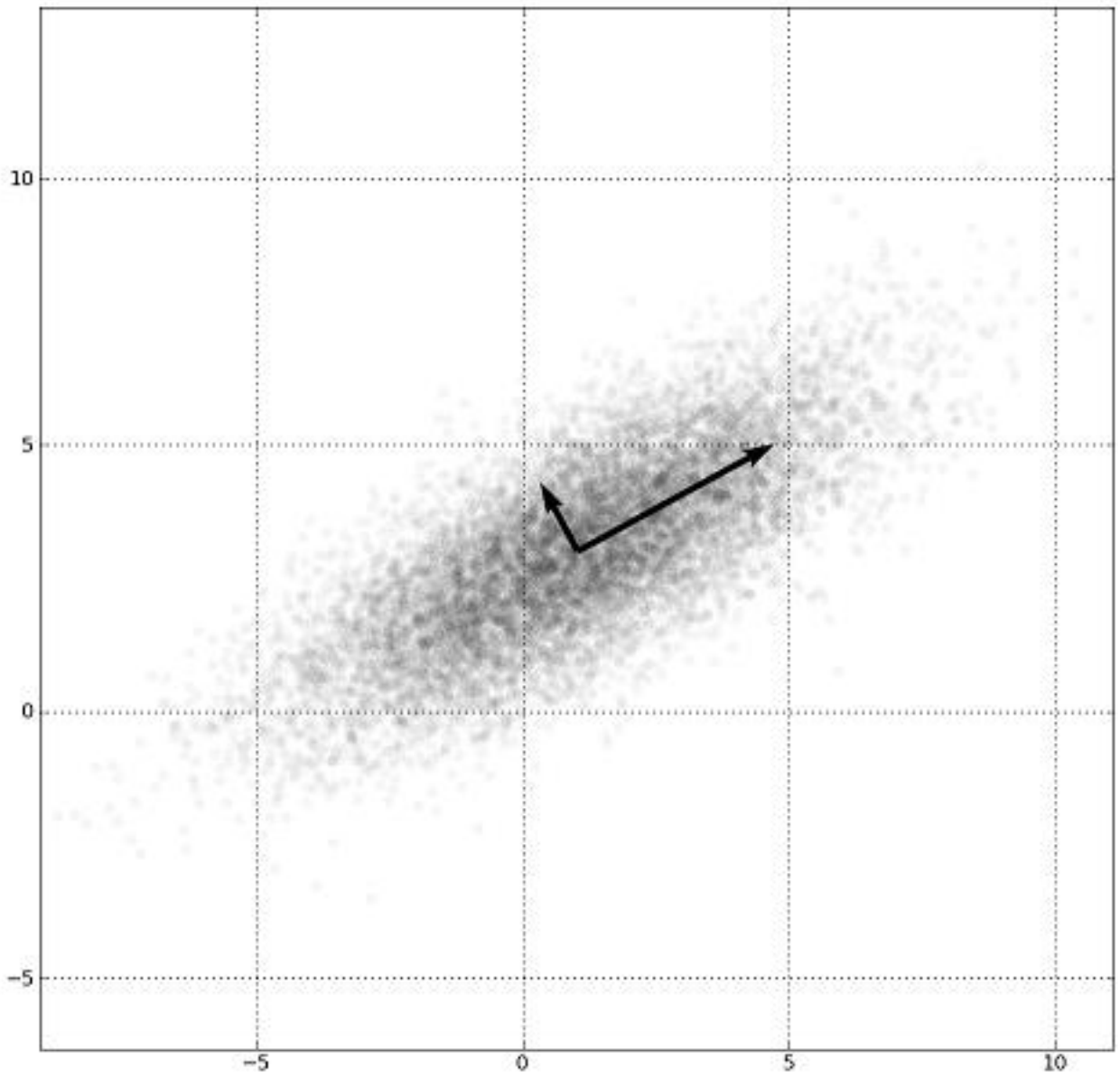
$$= \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

(4) Eigenvector and Eigenvalue
- try to solve AV = λV
- A be an n*n matrix (i.e., ***cov***)
- λ is the eigenvalue of A
- V is the eigenvector of A

AV = λV

$$\begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix} \cdot \begin{pmatrix} -.735178656 \\ .677873399 \end{pmatrix} = 0.490833989 \cdot \begin{pmatrix} -.735178656 \\ .677873399 \end{pmatrix}$$

$$\begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix} \cdot \begin{pmatrix} -.677873399 \\ -.735178656 \end{pmatrix} = 1.28402771 \cdot \begin{pmatrix} -.677873399 \\ -.735178656 \end{pmatrix}$$



A scatter plot of samples that are distributed according a multivariate (bivariate) Gaussian distribution centered at (1, 3) with a standard deviation of 3 in roughly the (0.878, 0.478) direction and of 1 in the orthogonal direction. The directions represent the Principal Components (PC) associated with the sample.
~ Ben FrantzDale

由於 x 和 y 分量共變，x 與 y 的變異數不能完全描述該分布；箭頭的方向對應的共變異數矩陣的特徵向量，其長度為特徵值的平方根。

特徵值越大，說明樣本在對應的特徵向量上投影後的平方差越大，樣本點越離散，越容易區分，訊息量也就越多。

PCA, example of calculation 3/3

(5,6) Transformed data

Eigenvector

$$(x \quad y)\cdot\begin{pmatrix} -.677873399 & -.735178656 \\ -.735178656 & .677873399 \end{pmatrix} = (PC1 \quad PC2)$$

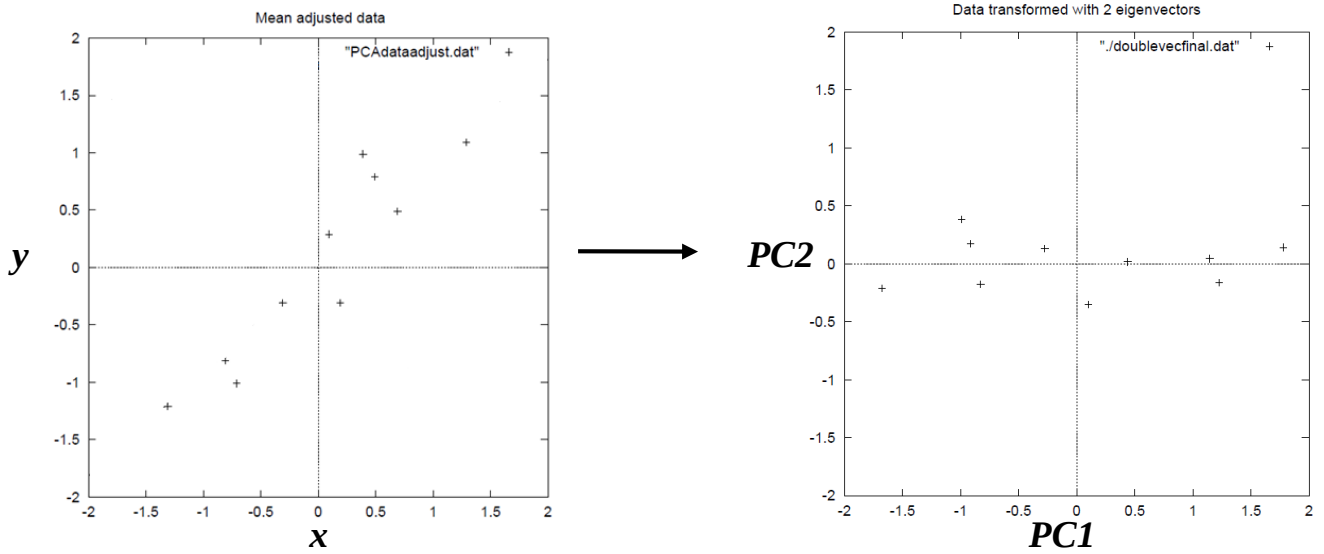
or

$$PC1 = -.677873399x - .735178656y$$

$$PC2 = -.735178656x + 0.677873399y$$

Transformed Data=

PC1	*PC2*
-.827970186	-.175115307
1.77758033	.142857227
-.992197494	.384374989
-.274210416	.130417207
-1.67580142	-.209498461
-.912949103	.175282444
.0991094375	-.349824698
1.14457216	.0464172582
.438046137	.0177646297
1.22382056	-.162675287



Var(x) + Var(y)
= Var(PC1) + Var(PC2)

What is a “good” subspace?

- Each of those eigenvectors is associated with an eigenvalue, which tell us about the “length” or “magnitude” of the eigenvectors.
- If we observe that all the eigenvalues are of very similar magnitude, this is a good indicator that our data is already in a “good” subspace.
- Or if some of the eigenvalues are much much higher than others, we might be interested in keeping only those eigenvectors with the much larger eigenvalues, since they contain more information about our data distribution.
- Vice versa, eigenvalues that are close to 0 are less informative and we might consider in dropping those when we construct the new feature subspace.

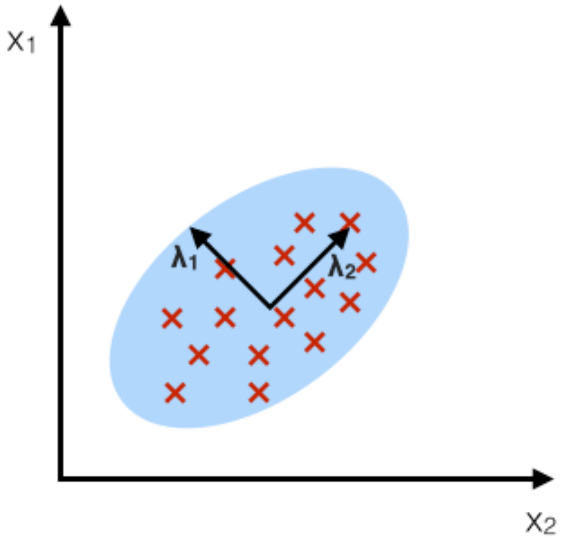
Dimension Reduction

- 主要目的乃是希望用較少的變數去解釋原來資料中的大部份變異，亦即期望能將我們手中許多相關性很高的變數轉化成彼此互相獨立的變數，能由其中選取較原始變數個數少，能解釋大部份資料之變異的幾個新變數。

1. 迴歸分析常遇到一個困擾，就是所謂共線性問題，這是由於預測變數間有高度相關所造成，主成份分析是解決共線性問題之一。
2. 所謂主成份分析是尋找幾個解釋變數的線性組合，一方面要能保有原來變數大部分的資訊，而且主成份間彼此是獨立的。
3. 能以“少數”幾個主成份代替原來“多個”解釋變數。

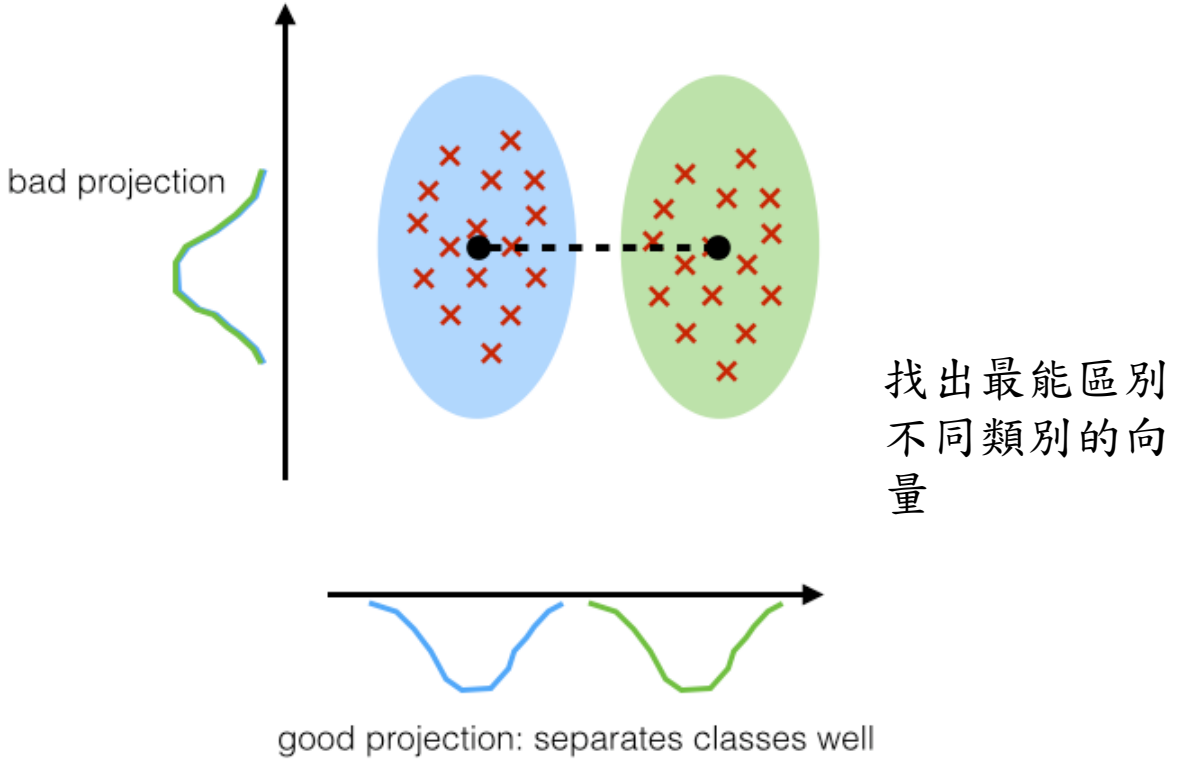
PCA and LDA

PCA:
component axes that maximize the variance



找出最能解釋變數變異情況的向量

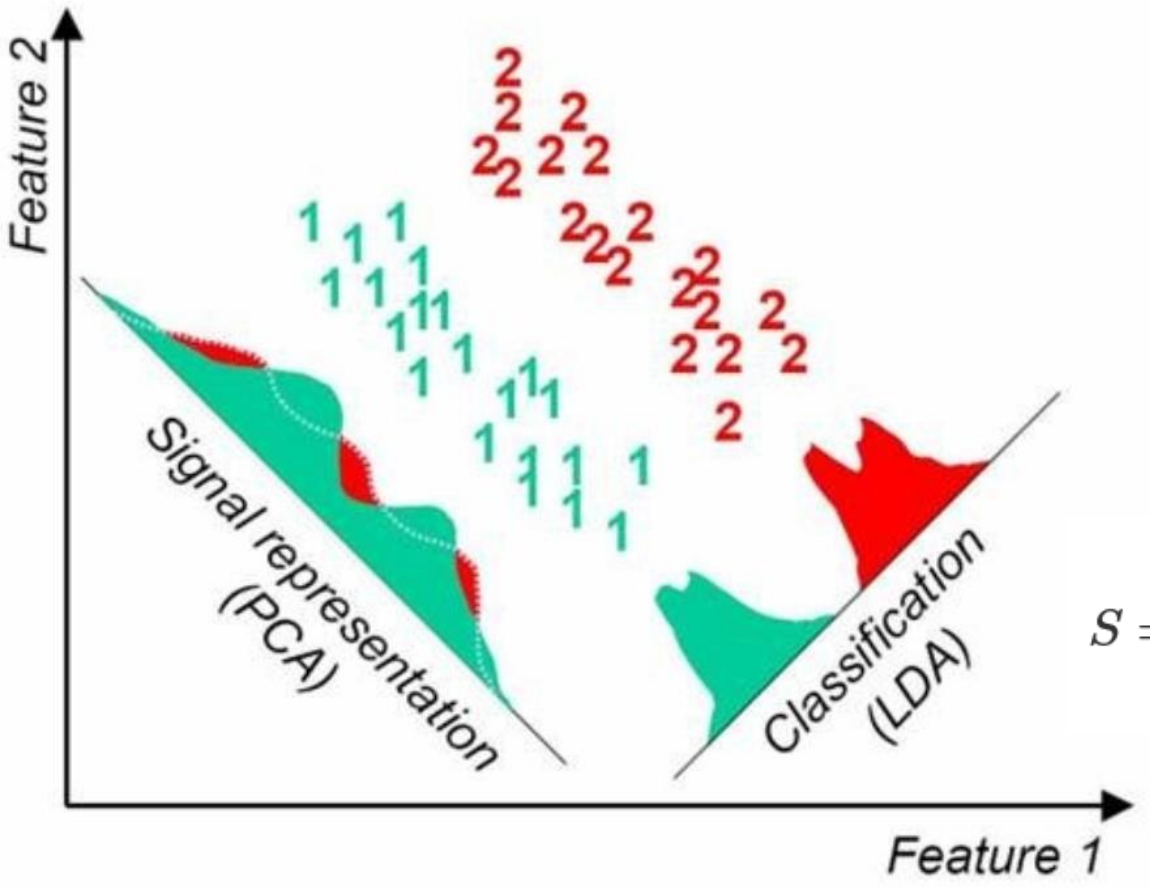
LDA:
maximizing the component axes for class-separation



找出最能區別不同類別的向量

http://sebastianraschka.com/Articles/2014_python_lda.html

PCA and LDA



LDA (Linear Discriminant Analysis): 有目標函數，
欲極大化S值，組間變異值要大且組內變異值要小。

$$S = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}} = \frac{(w \cdot \mu_2 - w \cdot \mu_1)^2}{w^T \Sigma_2 w + w^T \Sigma_1 w} = \frac{(w \cdot (\mu_2 - \mu_1))^2}{w^T (\Sigma_1 + \Sigma_2) w}$$