Linear Dimensionality Reduction: PCA

Shangsong Liang

Sun Yat-sen University

Outline

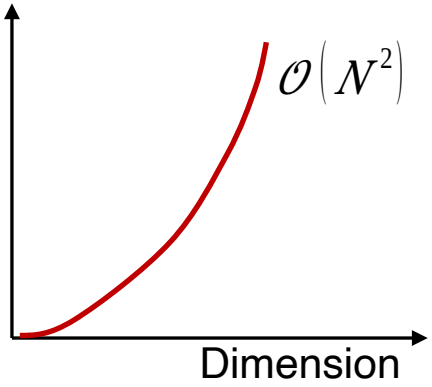
- Motivation
- Perspective 1: Minimizing Reconstruction Error
- Perspective 2: Maximizing Variance
- Perspective 3: SVD
- Other Applications of PCA

Motivation

- The dimensionality of many types of data is very high, *e.g.*, the dimension of each image below is as high as

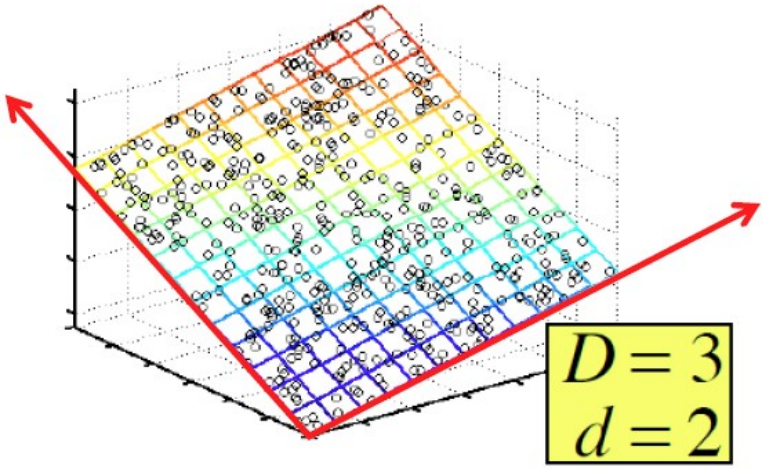
$$256 \times 256 = 65536$$

- If we work on the raw data directly, the complexity of subsequent tasks (e.g. classification) could be extremely high

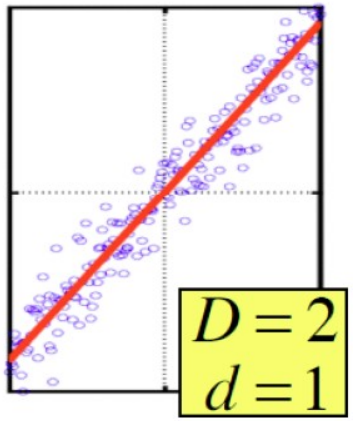


The complexity of logistic regression

- The high-dimensional data often resides on a low-dimensional intrinsic space approximately



3-dimensional data lies on a 2-dimensional plane

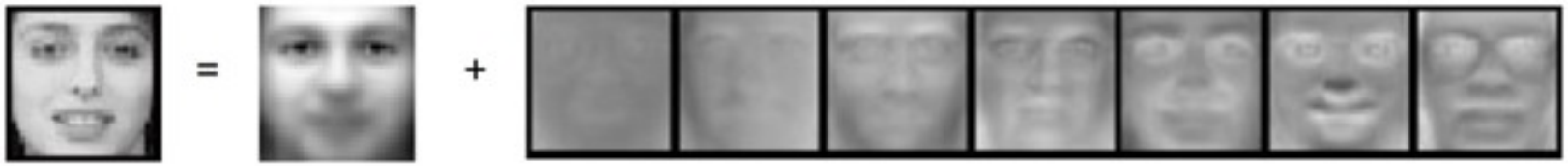


2-dimensional data lies on a 1-dimensional line

Finding *the principal directions* so that the dimensions of data represented under the new directions can be reduced significantly

- For the real-world data, this is also possible

e.g., an face image can be represented well *by only several values* if appropriate principal directions can be found



$$\boldsymbol{x} \approx \boldsymbol{\mu}_0 + a_1 \boldsymbol{\mu}_1 + \cdots + a_7 \boldsymbol{\mu}_7$$

The raw image that has 65536 values can be represented by only 7 values of

Outline

- Motivation
- Perspective 1: Minimizing the Reconstruction Error
- Perspective 2: Maximizing Variance
- Perspective 3: SVD
- Other Applications of PCA

Re-representation under the New Directions

- Orthogonal directions in high dimensional space

A set of vectors satisfying

$$\boldsymbol{u}_i^T \boldsymbol{u}_j = \delta_{ij}$$

where if ; 0 otherwise

Theorem: Under the given orthogonal directions , the *best approximation* to a data sample is

$$\tilde{\boldsymbol{x}} = \alpha_1 \boldsymbol{u}_1 + \alpha_2 \boldsymbol{u}_2 + \cdots + \alpha_M \boldsymbol{u}_M$$

with the equals to

$$\alpha_i = \boldsymbol{u}_i^T \boldsymbol{x}$$

Proof:

$$\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|^2=\left\|\boldsymbol{x}-\sum_{i=1}^{M}\alpha_i\,\boldsymbol{u}_i\right\|^2$$

$$¿\,\|\boldsymbol{x}\|^2-2\sum_{i=1}^{M}\alpha_i\,\boldsymbol{u}_i^T\,\boldsymbol{x}+\sum_{i=1}^{M}\alpha_i^2$$

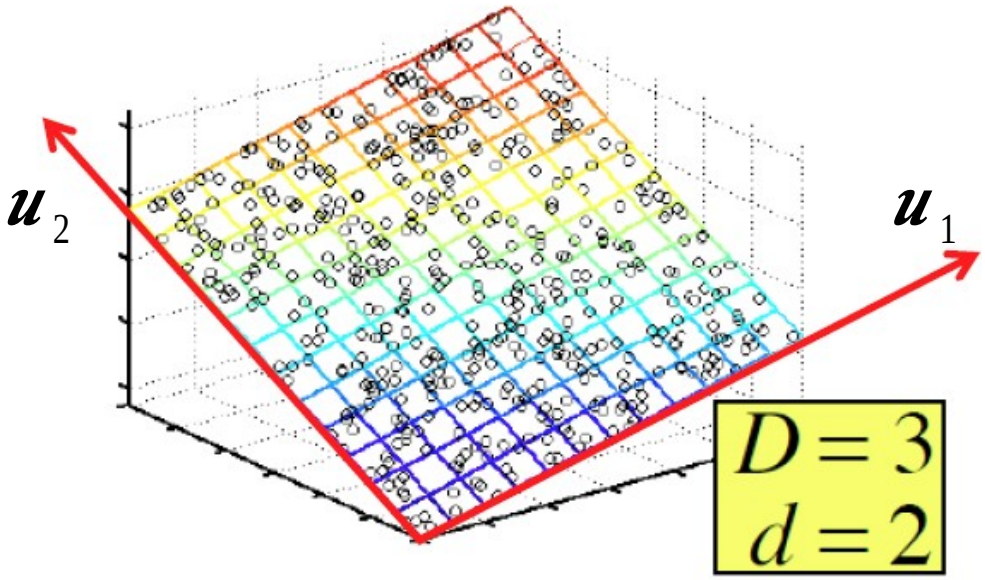
where we used for and for

This is a quadratic function, and can be minimized when

Given the directions , the best coefficient is . But *which directions are the best is still unknown*

Finding the Best Directions

- **Objective:** Given data from , finding the orthogonal directions under which the original data can be represented best



$$x^{(n)} \approx \sum_{i=1}^{M} \alpha_i^{(n)} u_i$$

- Suppose the best directions are given, what is the coefficients ?

$$\alpha_i^{(n)} = \boldsymbol{u}_i^T \boldsymbol{x}^{(n)}$$

Instead of representing the data directly, we first center the data to the origin, *i.e.*, representing data

,

with

- The objective can be formulated as minimizing the error between data and its approximant in

where the best coefficient is known equal to

$$\alpha_i = \boldsymbol{u}_i^T \left(\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}} \right)$$

- Reformulating the reconstruction error

  a) Substituting into and using gives

  b) Substituting gives

c) Rewritting it in a matrix form gives

where and is the Frobenius norm

- Minimizing is equivalent to maximize

$$\max_{\boldsymbol{u}_1 \cdots \boldsymbol{u}_M} \sum_{i=1}^{M} \boldsymbol{u}_i^T \boldsymbol{S} \boldsymbol{u}_i$$

$$s.t.: \boldsymbol{u}_i^T \boldsymbol{u}_j = \delta_{ij}$$

- Consider the simple case with . The problem is reduced to:

  - This is equivalent to maximize

$$\boldsymbol{u}_1^T \boldsymbol{S}\, \boldsymbol{u}_1 - \lambda \left(\boldsymbol{u}_1^T \boldsymbol{u}_1 - 1 \right)$$

  - Taking the derivative *w.r.t.* and setting it to 0 gives

$$\boldsymbol{S}\, \boldsymbol{u}_1 = \lambda\, \boldsymbol{u}_1,$$

  from which we can see that should be the eigenvector of *w.r.t. to the largest eigenvalue*

- For the case with , the problem becomes

$$\max_{\boldsymbol{u}_1,\boldsymbol{u}_2} \boldsymbol{u}_1^T \boldsymbol{S}\, \boldsymbol{u}_1 + \boldsymbol{u}_2^T \boldsymbol{S}\, \boldsymbol{u}_2$$

$$s.t.: \boldsymbol{u}_1^T \boldsymbol{u}_1 = 1, \boldsymbol{u}_2^T \boldsymbol{u}_2 = 1, \boldsymbol{u}_1^T \boldsymbol{u}_2 = 0$$

  - This is equivalent to maximize

$$\boldsymbol{u}_1^T \boldsymbol{S}\, \boldsymbol{u}_1 - \lambda_1 \left(\boldsymbol{u}_1^T \boldsymbol{u}_1 - 1 \right) + \boldsymbol{u}_2^T \boldsymbol{S}\, \boldsymbol{u}_2 - \lambda_2 \left(\boldsymbol{u}_2^T \boldsymbol{u}_2 - 1 \right)$$

    under the constraint

  - Taking the derivative *w.r.t.* and and setting it to 0 gives

$$\boldsymbol{S}\, \boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1, \boldsymbol{S}\, \boldsymbol{u}_2 = \lambda_2 \boldsymbol{u}_2,$$

    $\Rightarrow$ and must be the eigenvectors of

    $\Rightarrow$ In fact, to have maximized, and must be the eigenvectors *corresponding to the largest two eigenvalues*

For the case , the directions are *the eigenvectors of corresponding to the largest eigenvalues*

Question: Does the eigenvectors of satisfy ?

- For any semi-positive definite matrix , it has eigenvectors, and they are orthogonal to each other

- For every , it can be decomposed as

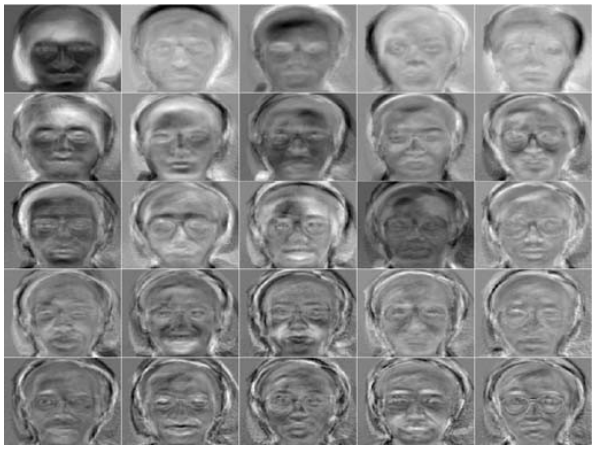
$$\boldsymbol{S} = \boldsymbol{U}\ \boldsymbol{\Lambda}\ \boldsymbol{U}^T$$

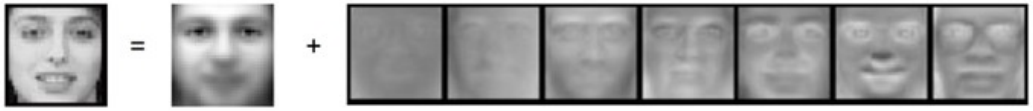
where consists of the eigenvectors and ; is a diagonal matrix consisting of eigenvalues of

Examples

Input data: each face image is a data point

Top 25 principal directions





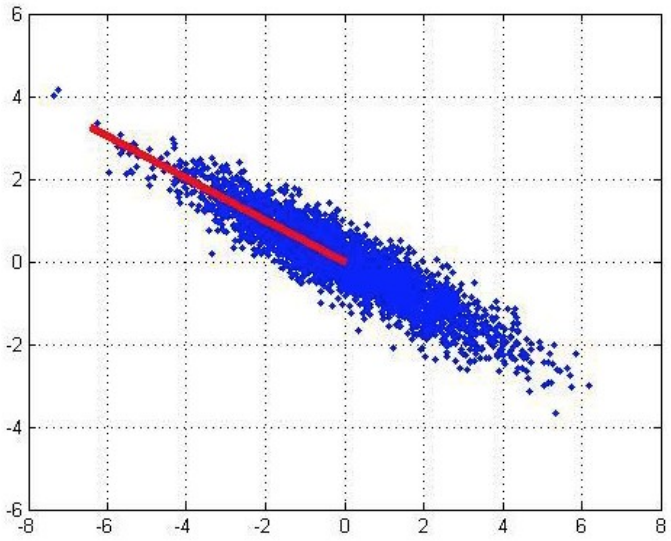
$$\boldsymbol{x} \approx \overline{\boldsymbol{x}} + \alpha_1 \boldsymbol{\mu}_1 + \cdots + \alpha_7 \boldsymbol{\mu}_7$$

Outline

- Motivation
- Perspective 1: Minimizing the Reconstruction Error
- Perspective 2: Maximizing Variance
- Perspective 3: SVD
- Other Applications of PCA

Problem Formulation

- **Objective:** Given data from , finding the orthogonal directions onto which the variance of data projected is maximized



Maximizing the variance is equivalent to *preserve the information of the original data as much as possible*

- For the first direction , we hope the variance in data projected onto the direction , *i.e.,* is maximized

  - The variance expression

$$var = \frac{1}{N}\sum_{n=1}^{N}\left(\boldsymbol{u}_1^T\left(\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}}\right)\right)^2$$

$$\propto \boldsymbol{u}_1^T \frac{1}{N}\sum_{n=1}^{N}\left(\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}}\right)\left(\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}}\right)^T \boldsymbol{u}_1$$

$$\propto \boldsymbol{u}_1^T \boldsymbol{S}\, \boldsymbol{u}_1$$

  - Subjecting to , as derived before, the variance is maximized when is *the eigenvector of corresponding to the largest eigenvalue*

- For the second direction , it also should maximize the variance

$$var = \boldsymbol{u}_2^T \boldsymbol{S}\, \boldsymbol{u}_2,$$

but should subject to the constraints , that is,

$$\boldsymbol{u}_2^T \boldsymbol{u}_2 = 1 \quad \boldsymbol{u}_1^T \boldsymbol{u}_2 = 0$$

- Due to is the eigenvector *w.r.t.* the largest eigenvalue, it can be proved that *is the eigenvector of corresponding the second largest eigenvalue*

is the eigenvector of corresponding the i-th largest eigenvalue

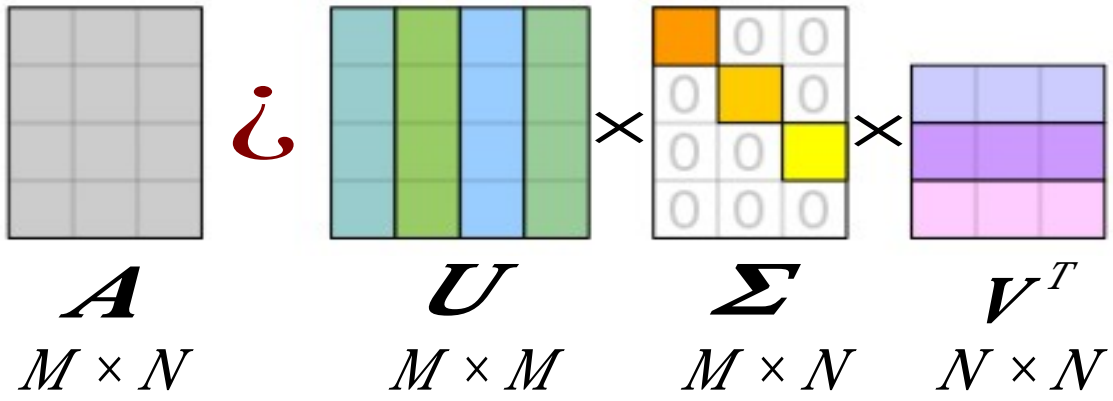
Outline

- Motivation
- Perspective 1: Minimizing Reconstruction Error
- Perspective 2: Maximizing Variance
- Perspective 3: SVD
- Other Applications of PCA

Singular Value Decomposition (SVD)

- For any matrix , it can always be decomposed as

$$\boldsymbol{A} = \boldsymbol{U}\ \boldsymbol{\Sigma}\ \boldsymbol{V}^T$$



- and , with and being the -th eigenvector of and , *and* and

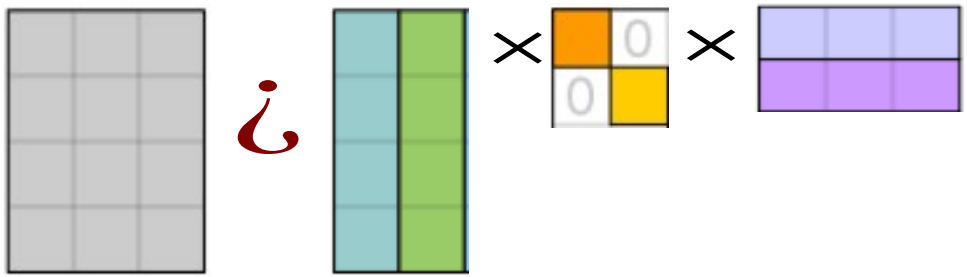
- has nonzero values on the diagonal, which are the squared roots of the eigenvalues of or *(They are the same)*

  is called *singular values* and are stored in a decreasing order

- Because only has nonzero values on the diagonal, can be expressed as

$$\boldsymbol{A} = \boldsymbol{U}' \boldsymbol{\Sigma}' \boldsymbol{V}'^{T} = \sum_{i=1}^{r} \Sigma_{ii} \boldsymbol{u}_i \boldsymbol{v}_i^{T}$$

where and are the -th column of and ; is the number of nonzero diagonal elements in



$\boldsymbol{A}$ $M \times N$ $\boldsymbol{U}'$ $M \times r$ $\boldsymbol{\Sigma}'$ $r \times r$ $\boldsymbol{V}'^{T}$ $r \times N$

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} \approx \begin{bmatrix} 0.13 & 0.02 & -0.01 \\ 0.41 & 0.07 & -0.03 \\ 0.55 & 0.09 & -0.04 \\ 0.68 & 0.11 & -0.05 \\ 0.15 & -0.59 & 0.65 \\ 0.07 & -0.73 & -0.67 \\ 0.07 & -0.29 & 0.32 \end{bmatrix} \begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix} \begin{bmatrix} 0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & -0.69 & -0.69 \\ 0.40 & -0.80 & 0.40 & 0.09 & 0.09 \end{bmatrix}$$

- The vector in the SVD decomposition of is the eigenvector of *w.r.t.* its -th largest eigenvalues

- By defining , we can see that

which has the same eigenvectors as the matrix

If we do SVD on , we can obtain the principal directions of the data

Outline

- Motivation
- Perspective 1: Minimizing Reconstruction Error
- Perspective 2: Maximizing Variance
- Perspective 3: SVD
- Other Applications of PCA

Image Compression

Divide the image below into many patches

- Each patch is viewed as an data instance
- Performing PCA on the patches

Reconstruction Error vs # PCA components

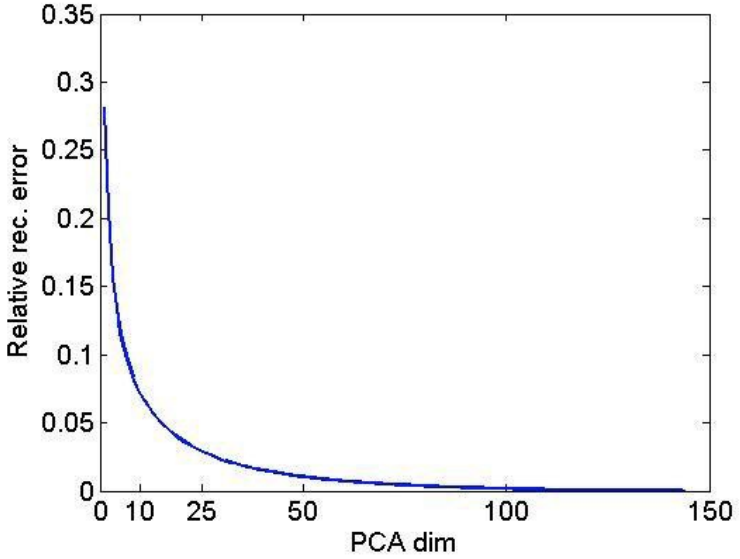
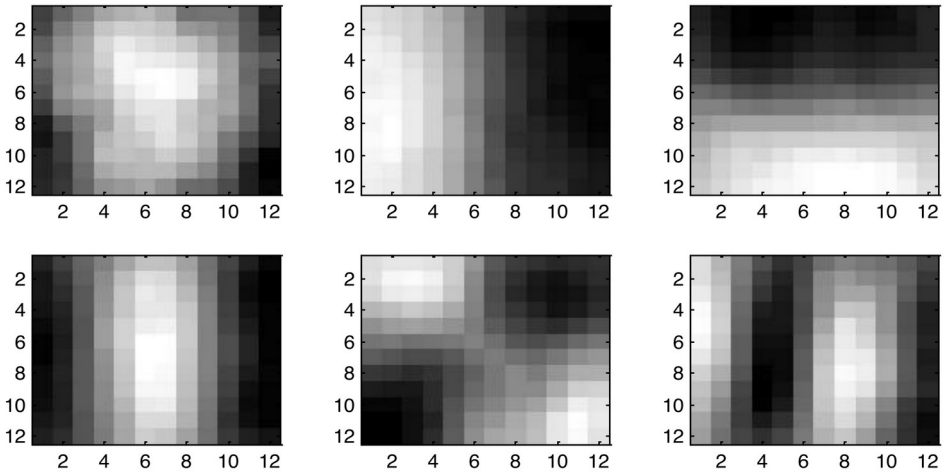


Illustration of the top 6 PCA components



Reconstruction with the top 60 components

Reconstruction with the top 16 components

Denoising

Noisy Image

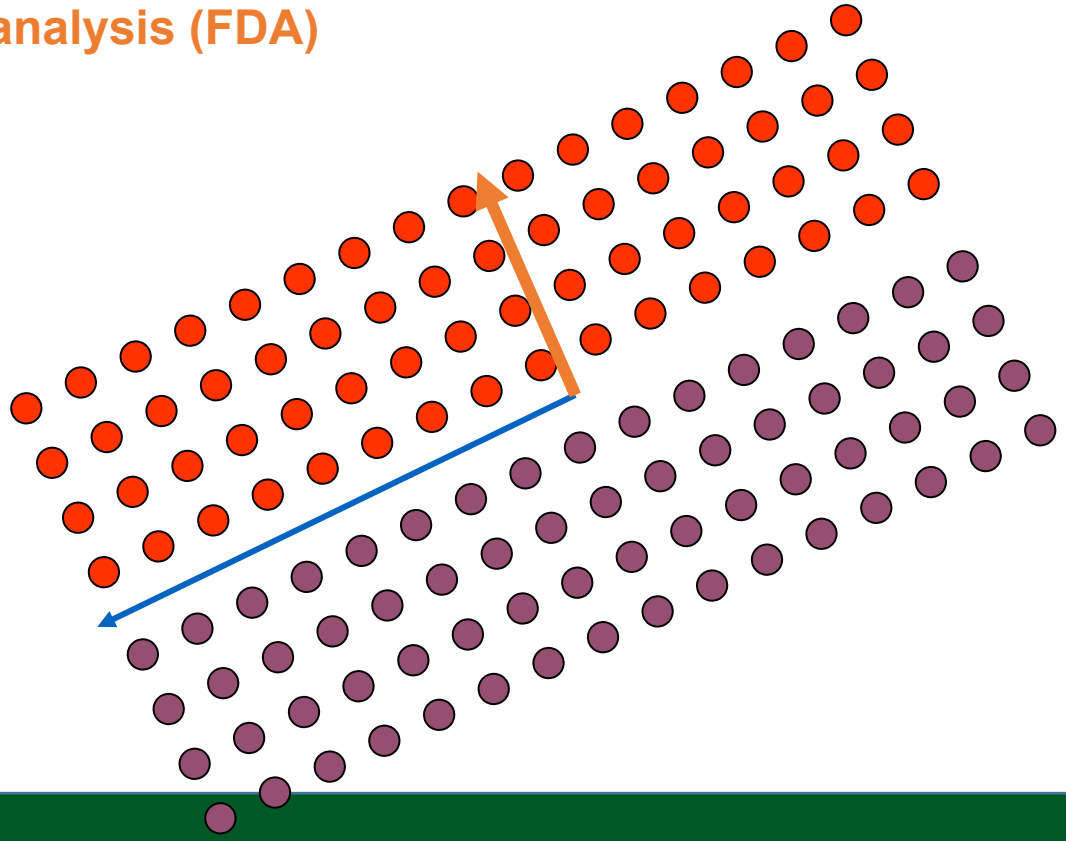
Denoised Image

Reconstructed from the top 15 principal components

Limitations of PCA

Are the maximal variance dimensions the relevant dimensions for preservation?

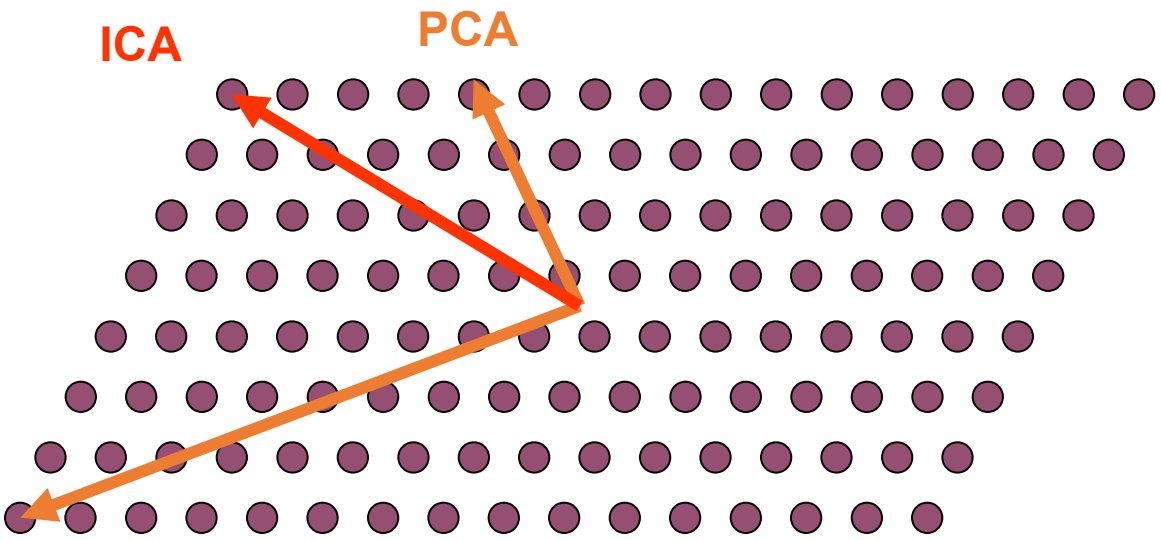
- Relevant Component Analysis (RCA)
- Fisher Discriminant analysis (FDA)



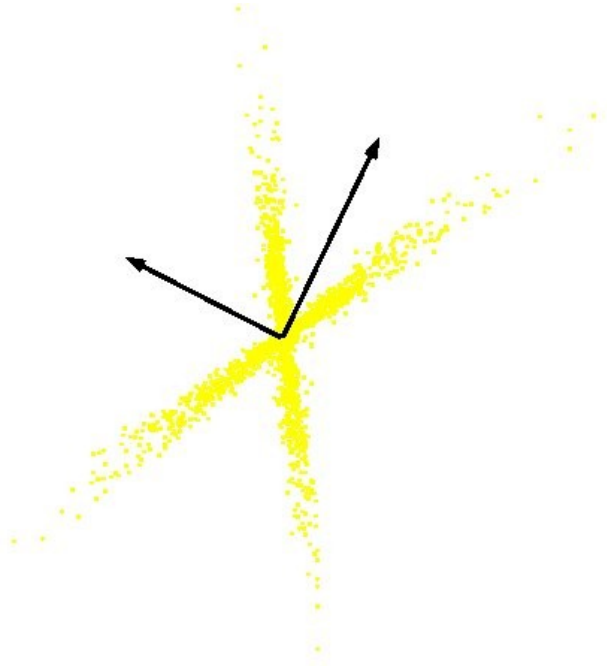
Limitations of PCA

Should the goal be finding independent rather than pair-wise uncorrelated dimensions

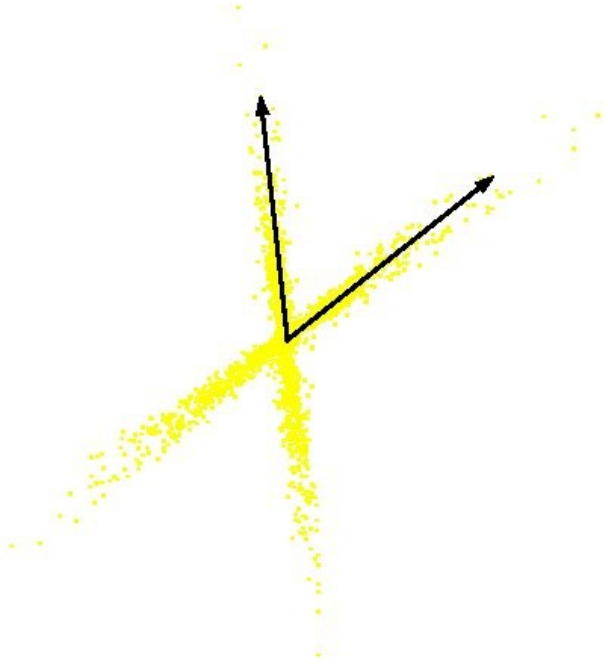
- Independent Component Analysis (ICA)



PCA vs ICA



PCA
(orthogonal coordinate)

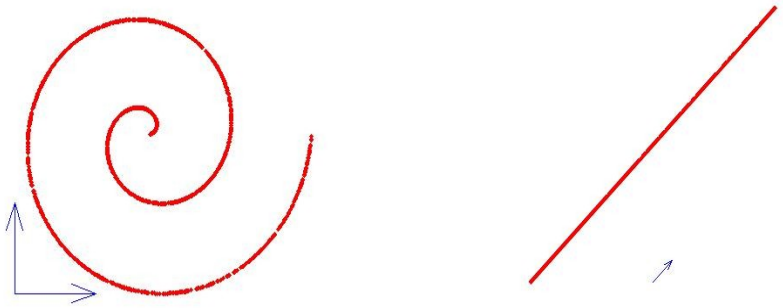


ICA
(non-orthogonal coordinate)

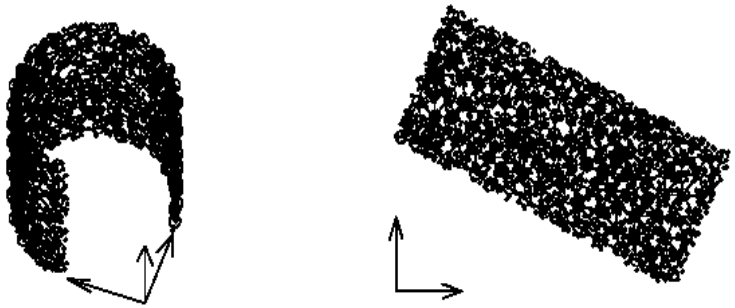
Limitations of PCA

- The reduction of dimensions for complex distributions may need non linear processing
- Curvilenear Component Analysis (CCA)
  - Non linear extension of PCA
  - Preserves the proximity between the points in the input space i.e. local topology of the distribution
  - Enables to unfold some varieties in the input data
  - Keep the local topology

Example of data representation using CCA



Non linear projection of a spiral



Non linear projection of a horseshoe

Thank You!

- Shangsong Liang
- SYSU