BIM2005 Fall 2021

Principal Component Analysis (PCA): Principle and Implementation

Haoran Sun (USTF)

November 24, 2021

CUHK-Shenzhen

Outline

Motivation

Mathematical background

Simple implementation

Outline

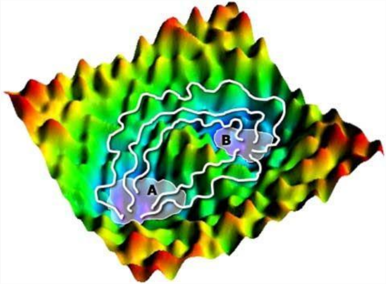
Motivation

Mathematical background

Simple implementation

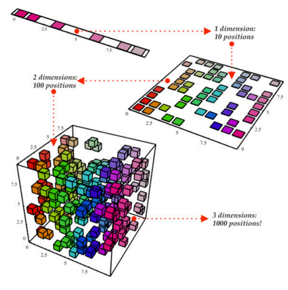
Why we reduce the dimension of a dataset?

- Curse of dimensionality
  - Difficult to understand: which dimension should we focus on? Where is the slow motion? How is the potential energy?
  - Noise exists: how to eliminate the meaningless fluctuation within a dataset?
  - Hard to visualize: how to understand them intuitively?



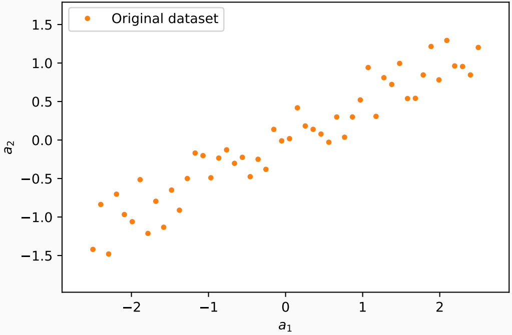
Why we reduce the dimension of a dataset?

- Dimensionality reduction
  - Find a discriptive low dimensional space.
  - Used for visualizing high-dimensional dataset, reduce the N-D data to 2-D or 3-D.
  - Identify low-dimensional space that contains information as much as possible (or, eliminate the noise).



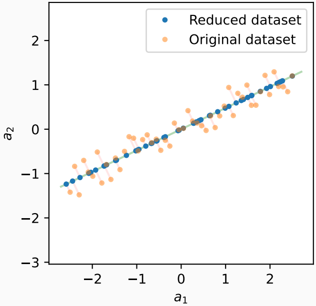
Motivation: reduce the dimension by projection

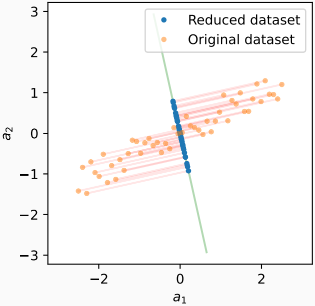
- Assume that we are going to reduce a 2-D dataset to 1-D by projection (MAT2040).



Motivation: reduce the dimension by projection

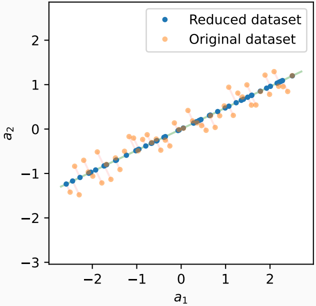
- Assume that we are going to reduce a 2-D dataset to 1-D by projection (MAT2040).

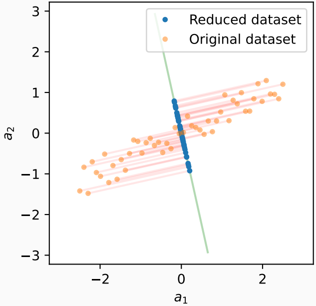




Motivation: reduce the dimension by projection

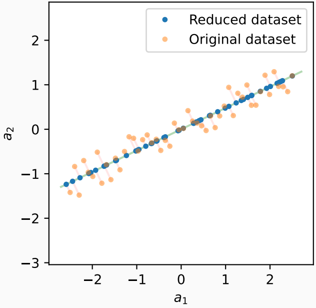
- We want this low-dimensional representation contains information as much as possible.
- Which of these reduced datasets contains ‘more’ information?

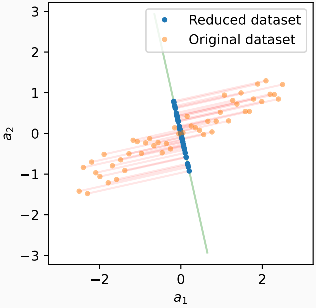




Motivation: choose the standard

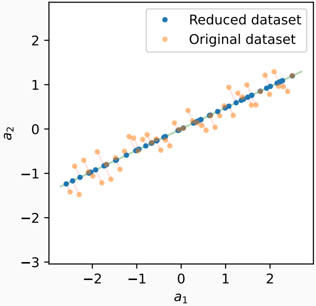
- Intuitively, spanned widely $\rightarrow$ more information.
- How to verify 'spanned widely' quantitatively? By which standard?

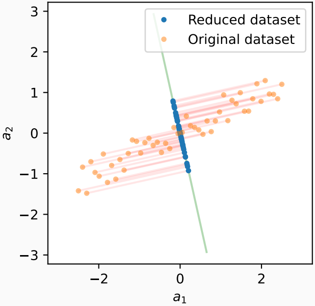




Motivation: choose the standard

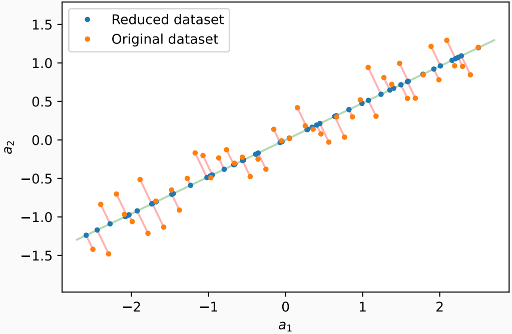
- In statistical inference, variance always used as an measurement that how data points spread around their mean value (STA2001).
- Thus, reduced dataset on left has higher variance, while the right has lower variance.





Motivation: choose the standard

- Therefore, we choose variance as the standard when choosing the direction which we are going to project data on.
- We want more information $\rightarrow$ we want to get reduced dataset large variance $\rightarrow$ **we want to maximize variance.**



Outline

Motivation

Mathematical background

Simple implementation

Notations: dataset

- We have a d-D dataset $\mathcal{A}$ contains N data points, represent this dataset as matrix A.
- Each column of a matrix represents a data point (vector).

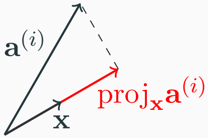
$$A = \begin{bmatrix} \mathbf{a}^{(1)} & \mathbf{a}^{(2)} & \cdots & \mathbf{a}^{(N)} \end{bmatrix} \in \mathbb{R}^{d \times N}$$

- $\mathbf{a}^{(i)}$ is the ith data point, it could be represented in column vector form.

$$\mathbf{a}^{(i)} = \begin{bmatrix} a_1^{(i)} \\ a_2^{(i)} \\ \vdots \\ a_d^{(i)} \end{bmatrix} \in \mathbb{R}^{d \times 1}$$

Notations: projection vector

- We would like to project dataset $\mathcal{A}$ onto a unit vector $\mathbf{x}$, i.e., we project each $\mathbf{a}^{(i)}$ onto $\mathbf{x}$.



- Since $\mathbf{x}$ is an unit vector, i.e., $\|\mathbf{x}\| = 1$, then

$$\mathrm{proj}_{\mathbf{x}}\mathbf{a}^{(i)} = (\mathbf{x} \cdot \mathbf{a}^{(i)})\mathbf{x}$$

where the inner product $\mathbf{x} \cdot \mathbf{a}^{(i)}$ is equivalent to

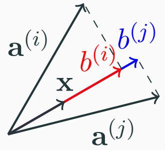
$$\mathbf{x} \cdot \mathbf{a}^{(i)} = \mathbf{x}^T \mathbf{a}^{(i)}$$

Also

$$\|\mathrm{proj}_{\mathbf{x}}\mathbf{a}^{(i)}\| = \mathbf{x} \cdot \mathbf{a}^{(i)} = \mathbf{x}^T \mathbf{a}^{(i)}$$

Notations: reduced dataset

- We can build an axis along the unit vector $\mathbf{x}$, using the length of projected vector as coordinate value.

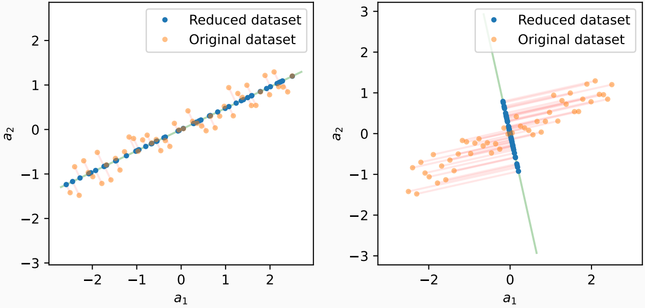


- Define reduced dataset $\mathcal{B}$ with respect to coordinate b, using matrix B to represent $\mathcal{B}$.

$$B = \begin{bmatrix} b_1 & b_2 & \cdots & b_N \end{bmatrix} = \mathbf{x}^T A = \begin{bmatrix} \mathbf{x}^T \mathbf{a}^{(1)} & \cdots & \mathbf{x}^T \mathbf{a}^{(N)} \end{bmatrix}$$

Notations: reduced dataset

- By projection, we can obtain 1-D dataset $\mathcal{B}$ from 2-D dataset $\mathcal{A}$.



Notations: mean and variance

- Recall the definition of mean and variance, given a random variable X, the mean μ and variance σ^2 is defined as

$$\mu = E(X)$$
$$\sigma^2 = E\left[(X-\mu)^2\right]$$

Notations: sample variance

- Recall our goal: **choosing appropriate unit vector $\mathbf{x}$ which maximize the variance of 1-D dataset $\mathcal{B}$.**
- Note that we would use sample variance $S^2_{\mathcal{B}}$, which is an approximate of variance $\sigma^2_{\mathcal{B}}$ (we assumed that $\mu_{\mathcal{A}} = \mathbf{0}$).

$$\begin{aligned}
\sigma^2_{\mathcal{B}} \approx S^2_{\mathcal{B}} &= \frac{1}{N-1}\sum_{i=1}^{N}(b^{(i)} - \bar{b})^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N} {b^{(i)}}^2 && (\bar{b} = 0 \text{ if } \mu_{\mathcal{A}} = \mathbf{0}) \\
&= \frac{1}{N-1}\mathbf{x}^T A A^T \mathbf{x} \approx \frac{1}{N}\mathbf{x}^T A A^T \mathbf{x} && (\text{large } N)
\end{aligned}$$

- $\mathbf{S} = \frac{1}{N} A A^T$ is usually called the covariance matrix.

Notations: Lagrange multipliers

- Thus, our question becomes

$$\begin{array}{ll} \text{Maximize} & S_{\mathcal{B}}^2 = \frac{1}{N}\mathbf{x}^T A A^T \mathbf{x} \\ \text{Constrained by} & \|\mathbf{x}\| = 1 \Leftrightarrow \|\mathbf{x}\| - 1 = 0 \end{array}$$

- Recall the knowledge in calculus. Generally, we solve extreme problem with constraint by the method of Lagrange multipliers.
- More specifically, we are going to solve

$$\begin{array}{ll} \text{Minimize/maximize} & f(\mathbf{x}) \\ \text{Constrained by} & g(\mathbf{x}) = 0 \end{array}$$

- We can solve this problem by define the Lagrange function $\mathcal{L}$, and solve the equation set $\nabla_{\mathbf{x},\lambda}\mathcal{L} = \mathbf{0}$.

$$\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda g(\mathbf{x}), \nabla_{\mathbf{x},\lambda}\mathcal{L} = \mathbf{0}$$

Simplification

- To simplify calculation, we modify our problem to

$$\begin{array}{ll} \text{Maximize} & f(\mathbf{x}) = \mathbf{x}^T A A^T \mathbf{x} \\ \text{Constrained by} & g(\mathbf{x}) = \mathbf{x}^T \mathbf{x} - 1 = 0 \end{array}$$

  since N is a constant, $\mathbf{x}^T\mathbf{x} = 1$ is equivalent to $\|\mathbf{x}\| = 1$

- Therefore, the Lagrange function would be

$$\mathcal{L}(\mathbf{x}, \lambda) = \mathbf{x}^T A A^T \mathbf{x} - \lambda(\mathbf{x}^T \mathbf{x} - 1)$$

Calculate the gradient

- Note that the gradient would be

$$\nabla_{\mathbf{x},\lambda}\mathcal{L} = \begin{bmatrix} \left[\frac{\partial \mathcal{L}}{\partial \mathbf{x}}\right]^T \\ 1 - \mathbf{x}^T\mathbf{x} \end{bmatrix} = \mathbf{0}$$

- It could be shown that

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{x}} &= \frac{\partial[\mathbf{x}^T AA^T\mathbf{x} - \lambda(\mathbf{x}^T\mathbf{x} - 1)]}{\partial \mathbf{x}} \\ &= 2\mathbf{x}^T AA^T - 2\lambda\mathbf{x}^T = 0 \\ \Rightarrow & AA^T\mathbf{x} = \lambda\mathbf{x} \end{aligned}$$

- Therefore

$$\begin{cases} AA^T\mathbf{x} = \lambda\mathbf{x} \\ 1 - \mathbf{x}^T\mathbf{x} = 0 \end{cases}$$

Eigenvalue

- For a square matrix $M \in \mathbb{R}^{n \times n}$, nonzero vector $\mathbf{x}$, and real value λ, if

$$M\mathbf{x} = \lambda\mathbf{x}$$

  then λ is a eigenvalue of matrix M and $\mathbf{x}$ is the eigenvector corresponding to eigenvalue λ.
- Thus, λ is the eigenvalue of AA^T and $\mathbf{x}$ is the eigenvector of AA^T.
- Recall that the expression of sample variance could be rewritten as

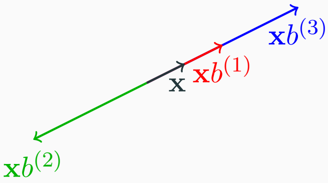
$$\sigma_{\mathcal{B}}^2 \approx S_{\mathcal{B}}^2 = \frac{1}{N}\mathbf{x}^T AA^T\mathbf{x} = \mathbf{x}^T\lambda\mathbf{x} = \lambda\mathbf{x}^T\mathbf{x} = \lambda$$

  Thus, the sample variance of reduced dataset $\mathcal{B}$ is exactly λ.

Conclusion

- To find $\mathbf{x}$, we
  - first find the largest eigenvalue λ of AA^T, then
  - find its corresponding eigenvector $\mathbf{x}$.
- Project $\mathcal{A}$ onto $\mathbf{x}$ to obtain $\mathcal{B}$.
- To represent $\mathcal{B}$ on axis along $\mathbf{x}$, remap $\mathcal{B}$ onto original data space by multiply $\mathbf{x}$ by B.

$$B' = \mathbf{x}B = \begin{bmatrix} \mathbf{x}b^{(1)} & \mathbf{x}b^{(2)} & \cdots & \mathbf{x}b^{(N)} \end{bmatrix}$$



Outline

Motivation

Mathematical background

Simple implementation

Numpy implementation

- Please check `PCA.ipynb` for detailed instruction.