Dimensionality Reduction

Spring 2024

Hongchang Gao

Why is dimensionality reduction important?

- 1. Exploratory data analysis
    - Visualize high-dimensional data

Car	MPG	Cylinders	Displacemen	Horsepower	Weight	Acceleration	Model	Origin
Chevrolet Ch	18	8	307	130	3504	12	70	US
Buick Skylark	15	8	350	165	3693	11.5	70	US
Plymouth Sa	18	8	318	150	3436	11	70	US
AMC Rebel S	16	8	304	150	3433	12	70	US
Ford Torino	17	8	302	140	3449	10.5	70	US
Ford Galaxie	15	8	429	198	4341	10	70	US
Chevrolet Im	14	8	454	220	4354	9	70	US
Plymouth Fu	14	8	440	215	4312	8.5	70	US
Pontiac Cata	14	8	455	225	4425	10	70	US
AMC Ambas	15	8	390	190	3850	8.5	70	US
Citroen DS-2	0	4	133	115	3090	17.5	70	Europe
Chevrolet Ch	0	8	350	165	4142	11.5	70	US
Ford Torino (	0	8	351	153	4034	11	70	US
Plymouth Sa	0	8	383	175	4166	10.5	70	US
AMC Rebel S	0	8	360	175	3850	11	70	US

Why is dimensionality reduction important?

- 2. Curse of dimensionality
  - Distance concentration refers to the problem of all the pairwise distances between different samples/points in the space converging to the same value as the dimensionality of the data increases.

$$\|x_1 - x_2\|_2 \rightarrow d$$

$$\|x_1 - x_3\|_2 \rightarrow d$$

Why is dimensionality reduction important?

- 3. Computational cost
  - data science algorithms scale linearly with the number of attributes, but very often the scaling is quadratic, or even worse
  - some of the more costly data science algorithms could be impossible to run on larger-sized data.

Why is dimensionality reduction important?

- 4. Noise reduction
    - the real-life data are often noisy and looking at any individual attribute might not provide any insight.
    - if we smartly combine a large number of noisy attributes into a small set of new ones, those new attributes might reveal some useful properties of the data that are not obvious in the original data.

How to reduce the dimensionality?

- How to reduce the dimensionality?
  - Eliminate features?
  - Lose information?

Car	MPG	~~Cylinders~~	Displacemen	Horsepower	~~Weight~~	Acceleration	Model	Origin
Chevrolet Ch	18	8	307	130	3504	12	70	US
Buick Skylark	15	8	350	165	3693	11.5	70	US
Plymouth Sa	18	8	318	150	3436	11	70	US
AMC Rebel S	16	8	304	150	3433	12	70	US
Ford Torino	17	8	302	140	3449	10.5	70	US
Ford Galaxie	15	8	429	198	4341	10	70	US
Chevrolet Im	14	8	454	220	4354	9	70	US
Plymouth Fu	14	8	440	215	4312	8.5	70	US
Pontiac Cata	14	8	455	225	4425	10	70	US
AMC Ambas	15	8	390	190	3850	8.5	70	US
Citroen DS-2	0	4	133	115	3090	17.5	70	Europe
Chevrolet Ch	0	8	350	165	4142	11.5	70	US
Ford Torino (	0	8	351	153	4034	11	70	US
Plymouth Sa	0	8	383	175	4166	10.5	70	US
AMC Rebel S	0	8	360	175	3850	11	70	US

Preserve information when reducing the dimensionality!

Toy Example

- 3D points
  - If each component is stored in a byte, 18 bytes are needed

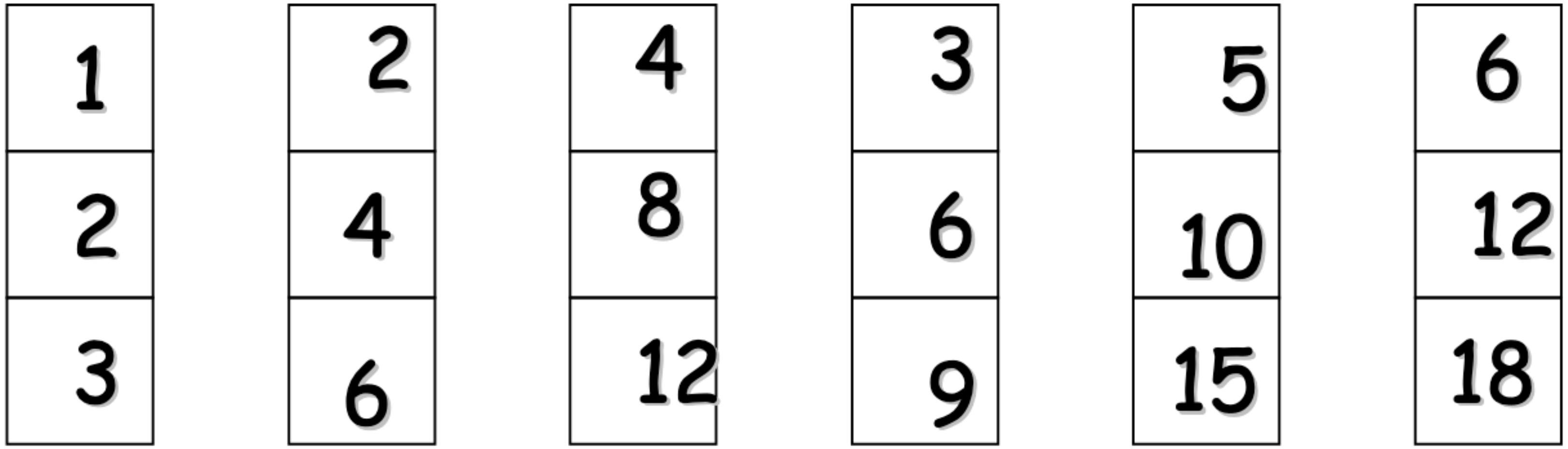


Fig: https://www.cse.psu.edu/~rtc12/CSE586Spring2010/lectures/pcaLectureShort.pdf

Toy Example

- All points are correlated: a common point scaled by different factors

$$\begin{bmatrix}1\\2\\3\end{bmatrix} = 1 * \begin{bmatrix}1\\2\\3\end{bmatrix} \quad \begin{bmatrix}2\\4\\6\end{bmatrix} = 2 * \begin{bmatrix}1\\2\\3\end{bmatrix} \quad \begin{bmatrix}4\\8\\12\end{bmatrix} = 4 * \begin{bmatrix}1\\2\\3\end{bmatrix}$$

$$\begin{bmatrix}3\\6\\9\end{bmatrix} = 3 * \begin{bmatrix}1\\2\\3\end{bmatrix} \quad \begin{bmatrix}5\\10\\15\end{bmatrix} = 5 * \begin{bmatrix}1\\2\\3\end{bmatrix} \quad \begin{bmatrix}6\\12\\18\end{bmatrix} = 6 * \begin{bmatrix}1\\2\\3\end{bmatrix}$$

Toy Example

- To store them, only 9 bytes are needed
  - Store one point and six scaling factors

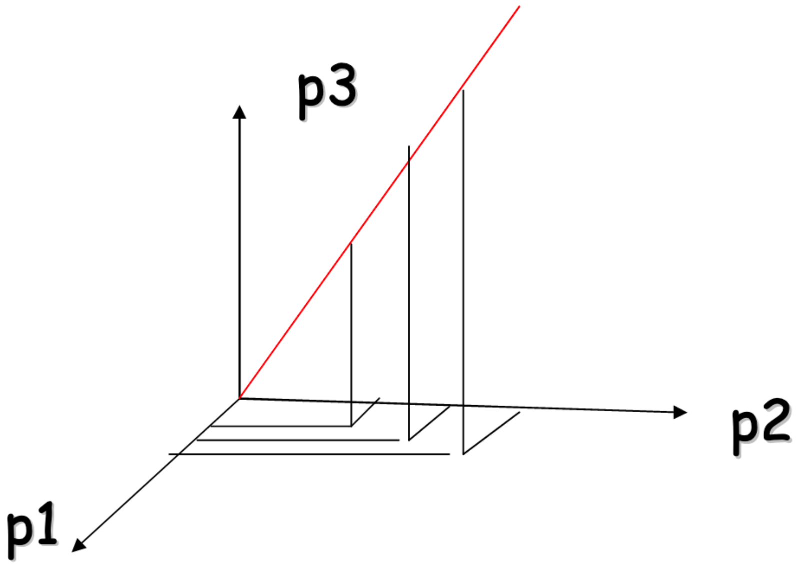
$$\begin{bmatrix}1\\2\\3\end{bmatrix} = 1 * \begin{bmatrix}1\\2\\3\end{bmatrix} \quad \begin{bmatrix}2\\4\\6\end{bmatrix} = 2 * \begin{bmatrix}1\\2\\3\end{bmatrix} \quad \begin{bmatrix}4\\8\\12\end{bmatrix} = 4 * \begin{bmatrix}1\\2\\3\end{bmatrix}$$

$$\begin{bmatrix}3\\6\\9\end{bmatrix} = 3 * \begin{bmatrix}1\\2\\3\end{bmatrix} \quad \begin{bmatrix}5\\10\\15\end{bmatrix} = 5 * \begin{bmatrix}1\\2\\3\end{bmatrix} \quad \begin{bmatrix}6\\12\\18\end{bmatrix} = 6 * \begin{bmatrix}1\\2\\3\end{bmatrix}$$

Toy Example

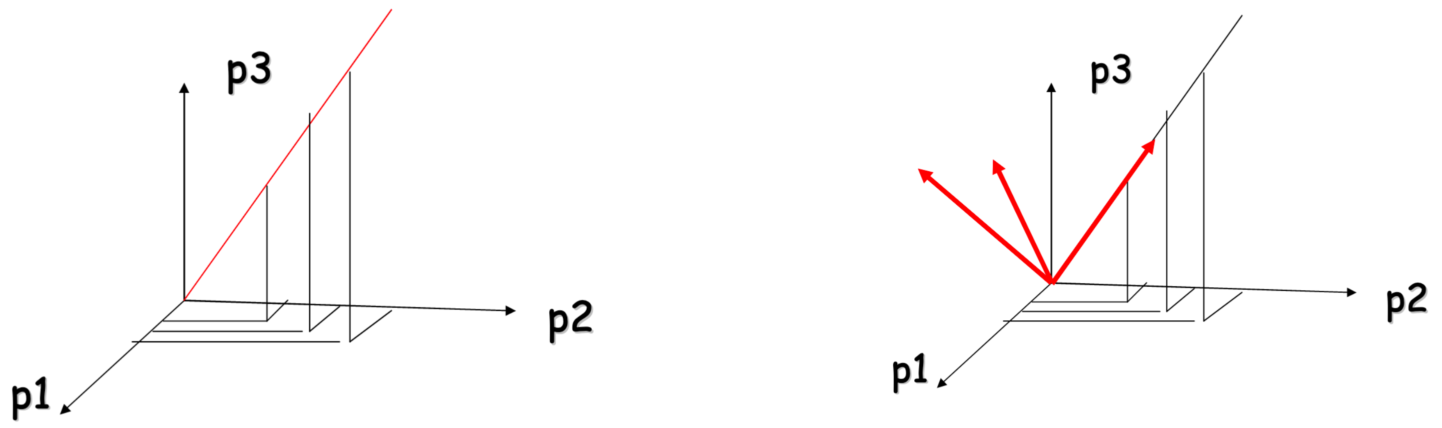
- Geometrical interpretation
  - All data points lie in a 1D subspace of the original 3D space

1	2	4	3	5	6
2	4	8	6	10	12
3	6	12	9	15	18



Toy Example

- Geometrical interpretation
    - Find a new coordinate system where one of the axes is along the direction of the line
    - In the new coordinate system, every data point has only one non-zero coordinate



Principal Components Analysis (PCA)

- PCA is a dimensionality reduction method
  - A linear transformation
    - Find a new coordinate system for the dataset
      - Only use a small part of coordinates to represent data points
    - Preserve as much of the data's variance as possible
- Formally, given a dataset with n samples $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ $\mathbf{x}_i \in \mathbb{R}^d$
  - Find a linear transformation $W^{d \times k}$ where $k < d$
    - d is the number of features in the original data
    - k is the number of new features
  - Preserve the variance as much as possible

Principal Components Analysis

x=(1, 0), y=(0, 1), a=(2, 3)

$x^T a = 2$, $y^T a = 3$

- First principal component
  - Subtract the mean

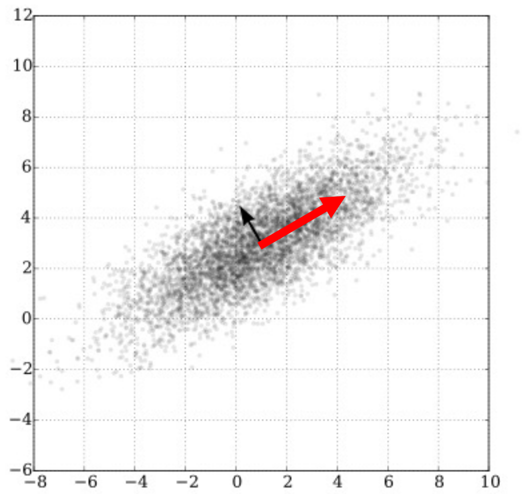
$$\tilde{X} = [\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \cdots, \tilde{\mathbf{x}}_n] \qquad \tilde{\mathbf{x}}_i = \mathbf{x}_i - \bar{\mathbf{x}}$$

  - Apply the linear transformation
    - Get data points in the new coordinate system

$$[\mathbf{w}^T\tilde{\mathbf{x}}_1, \mathbf{w}^T\tilde{\mathbf{x}}_2, \cdots, \mathbf{w}^T\tilde{\mathbf{x}}_n]$$

  - Compute the variance in the new coordinate system

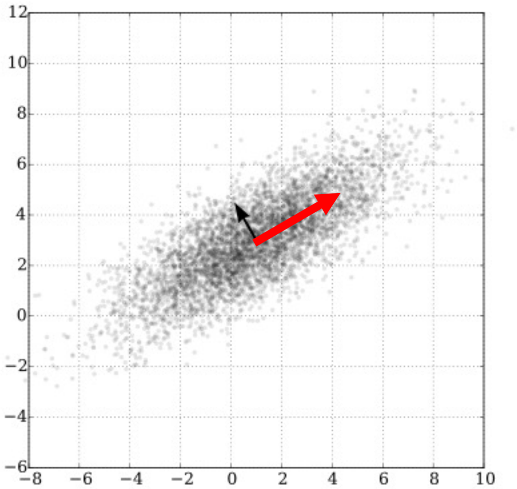
$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^T\tilde{\mathbf{x}}_1)^2$$



Principal Components Analysis

- How to find this new coordinate system/linear transformation?
  - Maximize the variance in the new coordinate system

$$\max_{\|\mathbf{w}\|_2=1} \frac{1}{n} \sum_{i=1}^{n} (\mathbf{w}^T \tilde{\mathbf{x}}_1)^2$$

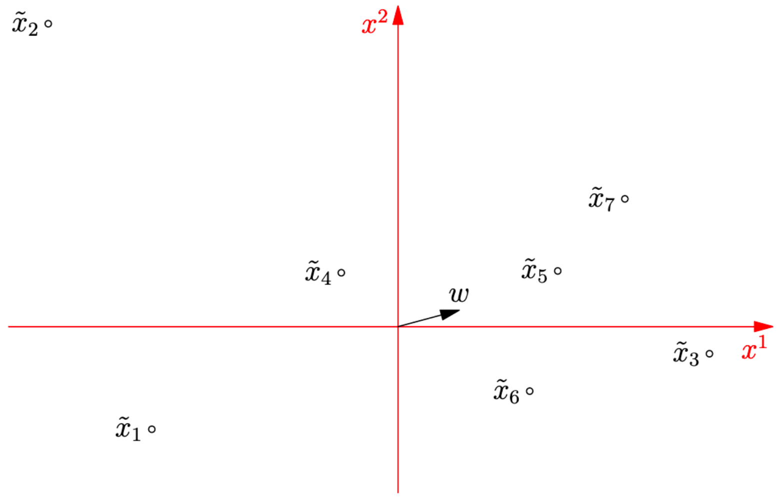


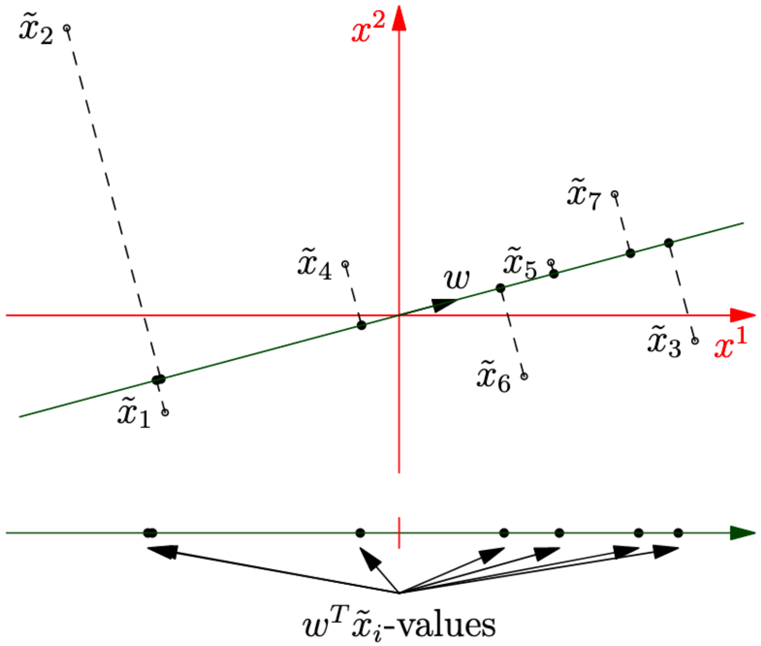
  - Optimizing this problem can find the new coordinate system
    - The first principal component: d → 1

$$\{\mathbf{w}^T \tilde{\mathbf{x}}_1, \mathbf{w}^T \tilde{\mathbf{x}}_2, \cdots, \mathbf{w}^T \tilde{\mathbf{x}}_n\}$$

Principal Components Analysis

- Visualization

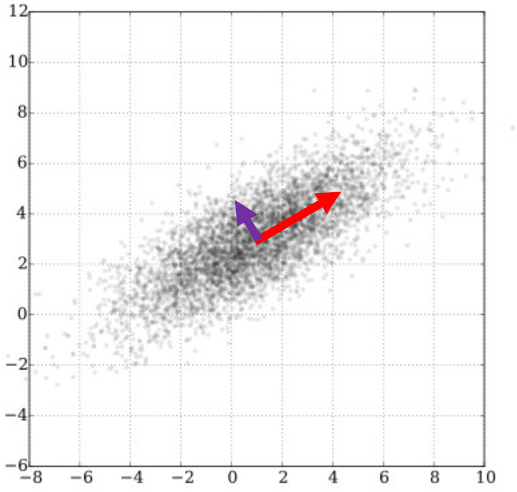




Principal Components Analysis

- Second principal component

$$\max_{\substack{\|\mathbf{w}\|_2=1,\\ \mathbf{w}\perp\mathbf{w}_1}} \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^T\tilde{\mathbf{x}}_i)^2$$



  - The first principal component

$$\{\mathbf{w}_1^T\tilde{\mathbf{x}}_1, \mathbf{w}_1^T\tilde{\mathbf{x}}_2, \cdots, \mathbf{w}_1^T\tilde{\mathbf{x}}_n\}$$

  - The second principal component

$$\{\mathbf{w}_2^T\tilde{\mathbf{x}}_1, \mathbf{w}_2^T\tilde{\mathbf{x}}_2, \cdots, \mathbf{w}_3^T\tilde{\mathbf{x}}_n\}$$

$$X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$$

$$W = [\mathbf{w}_1, \mathbf{w}_2] \in \mathbb{R}^{d\times 2}$$

$$\hat{X} = W^T\tilde{X} \in \mathbb{R}^{2\times n}$$

Principal Components Analysis

- The k-th principal component

$$\max_{\substack{\|\mathbf{w}\|_2=1, \\ \mathbf{w}\perp\mathbf{w}_1,\mathbf{w}_2,\cdots,\mathbf{w}_{k-1}}} \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^T\tilde{\mathbf{x}}_i)^2$$

$$X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$$

$$W = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_k] \in \mathbb{R}^{d\times k}$$

$$\Downarrow$$

$$\hat{X} = W^T\tilde{X} \in \mathbb{R}^{k\times n}$$

Principal Components Analysis

- Matrix form

$$
\begin{aligned}
&\textstyle\sum_{i=1}^{n} (\mathbf{w}^T \tilde{\mathbf{x}}_i)^2 \\
&= \textstyle\sum_{i=1}^{n} (\mathbf{w}^T \tilde{\mathbf{x}}_i)(\mathbf{w}^T \tilde{\mathbf{x}}_i) \\
&= \textstyle\sum_{i=1}^{n} (\mathbf{w}^T \tilde{\mathbf{x}}_i)(\tilde{\mathbf{x}}_i^T \mathbf{w}) \\
&= \mathbf{w}^T (\textstyle\sum_{i=1}^{n} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T) \mathbf{w} \\
&= \mathbf{w}^T \tilde{X} \tilde{X}^T \mathbf{w}
\end{aligned}
$$

$\Rightarrow$

Objective function of PCA

$$
\max_{W^T W = I} W^T \underbrace{\tilde{X} \tilde{X}^T}_{\text{Covariance matrix}} W
$$

How to optimize this model?

- Eigen-decomposition for the covariance matrix

$$A = U\Sigma U^T$$

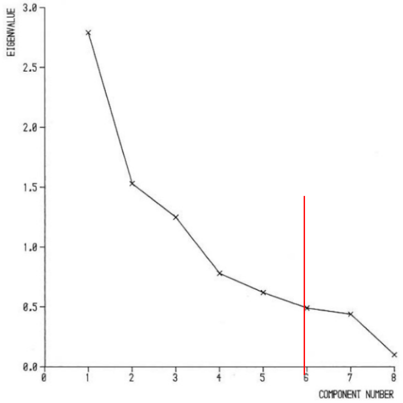
  - $A = \tilde{X}\tilde{X}^T \in \mathbb{R}^{d\times d}$ is the covariance matrix
  - $U = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_d] \in \mathbb{R}^{d\times d}$ where $\mathbf{u}_i$ is the i-th largest eigenvector, $\mathbf{u}_i^T\mathbf{u}_i = 0$, $\|\mathbf{u}_i\|_2 = 1$
  - $\Sigma = diag\{\lambda_1, \lambda_2, \cdots, \lambda_d\}$ where λ_i is the i-th largest eigenvalue $0 \le \lambda_d \le \cdots \le \lambda_2 \le \lambda_1$

- The solution is the largest k eigenvectors

$$\max_{W^TW=I} W^T\tilde{X}\tilde{X}^T W \quad \Longrightarrow \quad W = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_k] \in \mathbb{R}^{d\times k}$$

How to optimize this model?

- Interpretation



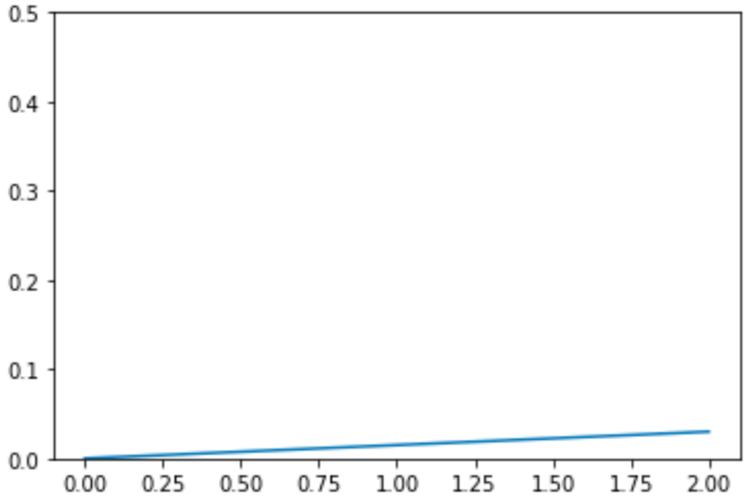
Covariance matrix $A = U\Sigma U^T$

$$A = \lambda_1 \mathbf{u}_1 \mathbf{u}_1^T + \lambda_2 \mathbf{u}_2 \mathbf{u}_2^T + \cdots + \lambda_k \mathbf{u}_k \mathbf{u}_k^T + \cdots + \lambda_d \mathbf{u}_d \mathbf{u}_d^T$$

Only keep the largest k eigenvalue $0 \leq \lambda_d \leq \cdots \boxed{\leq \lambda_2 \leq \lambda_1}$

$$A \approx \lambda_1 \mathbf{u}_1 \mathbf{u}_1^T + \lambda_2 \mathbf{u}_2 \mathbf{u}_2^T + \cdots + \lambda_k \mathbf{u}_k \mathbf{u}_k^T$$

$$W = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_k] \in \mathbb{R}^{d \times k}$$



$$\begin{pmatrix} 2 \\ 0.03 \end{pmatrix} = 2 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + 0.03 \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Summary

- Step 1: Mean subtraction

$$\tilde{X} = X - \frac{1}{n} X \mathbf{1}\mathbf{1}^T$$

- Step 2: Compute the covariance matrix

$$A = \tilde{X}\tilde{X}^T$$

- Step 3: Eigen-decomposition

$$A = U \Sigma U^T$$

- Step 4: Keep the largest k eigenvectors

$$W = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_k] \in \mathbb{R}^{d \times k}$$