Data and Data Exploration

Le Ou-Yang

Shenzhen University

Outline

- Additional remarks on Principal component analysis
- Remaining Data Preprocessing
  1) Feature Subset Selection
  2) Attribute Transformation
- Measure of Similarity & Dissimilarity
- What is data exploration?

Eigenvalues & Eigenvectors

- **Eigenvectors** (for a square $m \times m$ matrix $\mathbf{S}$)

$$\mathbf{S}\mathbf{v} = \lambda \mathbf{v}$$

(right) eigenvector eigenvalue

$$\mathbf{v} \in \mathbb{R}^m \neq \mathbf{0} \qquad \lambda \in \mathbb{R}$$

Example

$$\begin{pmatrix} 6 & -2 \\ 4 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \end{pmatrix} = 2 \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

- How many eigenvalues are there at most?

$$\mathbf{S}\mathbf{v} = \lambda \mathbf{v} \iff (\mathbf{S} - \lambda \mathbf{I})\, \mathbf{v} = \mathbf{0}$$

only has a non-zero solution if $|\mathbf{S} - \lambda \mathbf{I}| = 0$

this is a *m*-th order equation in λ which can have **at most *m* distinct solutions** (roots of the characteristic polynomial) – <u>can be complex even though **S** is real.</u>

Matrix-vector multiplication

$$S = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

has eigenvalues 3, 2, 0 with corresponding eigenvectors

$$v_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \qquad v_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \qquad v_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

On each eigenvector, S acts as a multiple of the identity matrix: but as a different multiple on each.

Any vector (say $x = \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix}$) can be viewed as a combination of the eigenvectors: $x = 2v_1 + 4v_2 + 6v_3$

Matrix vector multiplication

- Thus a matrix-vector multiplication such as Sx can be rewritten in terms of the eigenvalues/vectors:

$$Sx = S(2v_1 + 4v_2 + 6v_3)$$

$$Sx = 2Sv_1 + 4Sv_2 + 6Sv_3 = 2\lambda_1 v_1 + 4\lambda_2 v_2 + 6\lambda_3 v_3$$

- Even though x is an arbitrary vector, the action of S on x is determined by the eigenvalues/vectors.
- Suggestion: the effect of "small" eigenvalues is small.

Eigenvalues & Eigenvectors

For symmetric matrices, eigenvectors for distinct eigenvalues are **orthogonal**

$$Sv_{\{1,2\}} = \lambda_{\{1,2\}} v_{\{1,2\}}, \text{ and } \lambda_1 \neq \lambda_2 \Rightarrow v_1 \bullet v_2 = 0$$

All eigenvalues of a real symmetric matrix are **real**.

All eigenvalues of a positive semidefinite matrix are **non-negative**

$$\overbrace{\forall w \in \Re^n, w^T S w \geq 0}, \text{ then if } \mathrm{Sv} = \lambda \mathrm{v} \Rightarrow \lambda \geq 0$$

Example

- Let

$$S = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

Real, symmetric.

- Then

$$S - \lambda I = \begin{bmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{bmatrix} \Rightarrow (2-\lambda)^2 - 1 = 0.$$

- The eigenvalues are 1 and 3 (nonnegative, real).
- The eigenvectors are orthogonal (and real):

$$\begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Plug in these values and solve for eigenvectors.

Eigen/diagonal Decomposition

- Let $\mathbf{S} \in \mathbb{R}^{m \times m}$ be a **square** matrix with ***m* linearly independent eigenvectors** (a "non-defective" matrix)

- **Theorem**: Exists an **eigen decomposition**

diagonal

$$\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}$$

Unique for distinct eigen-values

  – (cf. matrix diagonalization theorem)

- Columns of ***U*** are **eigenvectors** of ***S***

- Diagonal elements of $\mathbf{\Lambda}$ are **eigenvalues** of $\mathbf{S}$

$$\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m), \quad \lambda_i \geq \lambda_{i+1}$$

Diagonal decomposition - example

Recall $S = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}; \lambda_1 = 1, \lambda_2 = 3.$

The eigenvectors $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ form $U = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$

Inverting, we have $U^{-1} = \begin{bmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{bmatrix}$

Recall $UU^{-1} = 1$.

Then, $\boldsymbol{S = U\Lambda U^{-1}} =$ $\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{bmatrix}$

Example continued

Let's divide ***U*** (and multiply ***U***$^{-1}$) by $\sqrt{2}$

Then, ***S=*** $\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$

Q Λ ***(***$Q^{-1} = Q^T$***)***

Symmetric Eigen Decomposition

- If $\mathbf{S} \in \mathbb{R}^{m \times m}$ is a **symmetric** matrix:
- **Theorem**: Exists a (unique) **eigen decomposition**

$$S = Q \Lambda Q^T$$

- where ***Q*** is **orthogonal:**
  - $\boldsymbol{Q^{-1} = Q^T}$
  - Columns of ***Q*** are normalized eigenvectors
  - Columns are orthogonal.
  - (everything is real)

Singular Value Decomposition

The SVD is a factorization of a $m \times n$ matrix into

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

where $\boldsymbol{U}$ is a $m \times m$ orthogonal matrix, $\boldsymbol{V}^T$ is a $n \times n$ orthogonal matrix and $\boldsymbol{\Sigma}$ is a $m \times n$ diagonal matrix.

For a square matrix ($\boldsymbol{m} = \boldsymbol{n}$):

$$\sigma_1 \geq \sigma_2 \geq \sigma_3 \ldots$$

$$\boldsymbol{A} = \begin{pmatrix} \vdots & \ldots & \vdots \\ \boldsymbol{u}_1 & \ldots & \boldsymbol{u}_n \\ \vdots & \ldots & \vdots \end{pmatrix} \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{pmatrix} \begin{pmatrix} \ldots & \mathbf{v}_1^T & \ldots \\ \vdots & \vdots & \vdots \\ \ldots & \mathbf{v}_n^T & \ldots \end{pmatrix}$$

$$\boldsymbol{A} = \begin{pmatrix} \vdots & \ldots & \vdots \\ \boldsymbol{u}_1 & \ldots & \boldsymbol{u}_n \\ \vdots & \ldots & \vdots \end{pmatrix} \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{pmatrix} \begin{pmatrix} \vdots & \ldots & \vdots \\ \boldsymbol{v}_1 & \ldots & \boldsymbol{v}_n \\ \vdots & \ldots & \vdots \end{pmatrix}^T$$

Singular Value Decomposition

What happens when $\boldsymbol{A}$ is not a square matrix?

1) $\boldsymbol{m > n}$

$$\boldsymbol{A = U\,\Sigma\,V^T} = \underbrace{\begin{pmatrix} \boxed{\begin{matrix} \vdots & \cdots & \vdots \\ \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_n \\ \vdots & \cdots & \vdots \end{matrix}} & \begin{matrix} \cdots & \vdots \\ \cdots & \boldsymbol{u}_m \\ \cdots & \vdots \end{matrix} \end{pmatrix}}_{m\times m} \underbrace{\begin{pmatrix} \boxed{\begin{matrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{matrix}} \\ \begin{matrix} & & 0 \\ & & \vdots \\ & & 0 \end{matrix} \end{pmatrix}}_{m\times n} \underbrace{\begin{pmatrix} \cdots & \mathbf{v}_1^T & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & \mathbf{v}_n^T & \cdots \end{pmatrix}}_{n\times n}$$

We can instead re-write the above as:

$$\boldsymbol{A = U_R\,\Sigma_R V^T}$$

Where $\boldsymbol{U_R}$ is a $m\times n$ matrix and $\boldsymbol{\Sigma_R}$ is a $n\times n$ matrix

Singular Value Decomposition

2) $n > m$

$$
A = U\,\Sigma\,V^T = \begin{pmatrix} \vdots & \cdots & \vdots \\ u_1 & \cdots & u_m \\ \vdots & \cdots & \vdots \end{pmatrix} \begin{pmatrix} \sigma_1 & & & 0 & & \\ & \ddots & & & \ddots & \\ & & \sigma_m & & & 0 \end{pmatrix} \begin{pmatrix} \cdots & \mathbf{v}_1^T & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & \mathbf{v}_m^T & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & \mathbf{v}_n^T & \cdots \end{pmatrix}
$$

$$
\qquad m{\times}m \qquad\qquad m{\times}n \qquad\qquad n{\times}n
$$

We can instead re-write the above as:

$$A = U\,\Sigma_R V_R{}^T$$

where V_R is a $n{\times}m$ matrix and Σ_R is a $m{\times}m$ matrix

In general:

$$A = U_R \Sigma_R V_R{}^T$$

U_R is a $m{\times}k$ matrix
Σ_R is a $k\times k$ matrix
V_R is a $n{\times}k$ matrix

$k = \min(m, n)$

Singular Value Decomposition

Let's take a look at the product $\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ has the singular values of a $\boldsymbol{A}$, a $m\times n$ matrix.

$$\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} = \underset{n\times m}{\begin{pmatrix} \sigma_1 & & & 0 & & \\ & \ddots & & & \ddots & \\ & & \sigma_n & & & 0 \end{pmatrix}} \underset{m\times n}{\begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ & & 0 \\ & & \vdots \\ & & 0 \end{pmatrix}} = \underset{n\times n}{\boxed{\begin{pmatrix} {\sigma_1}^2 & & \\ & \ddots & \\ & & {\sigma_n}^2 \end{pmatrix}}}$$

$m > n$

$$\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} = \underset{n\times m}{\begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_m \\ & & 0 \\ & & \vdots \\ & & 0 \end{pmatrix}} \underset{m\times n}{\begin{pmatrix} \sigma_1 & & & 0 & & \\ & \ddots & & & \ddots & \\ & & \sigma_m & & & 0 \end{pmatrix}} = \underset{n\times n}{\begin{pmatrix} {\sigma_1}^2 & & & 0 & & \\ & \ddots & & & \ddots & \\ & & {\sigma_m}^2 & & & 0 \\ 0 & & & 0 & & \\ & \ddots & & & \ddots & \\ & & 0 & & & 0 \end{pmatrix}}$$

$n > m$

Singular Value Decomposition

Assume $\boldsymbol{A}$ with the singular value decomposition $\boldsymbol{A} = \boldsymbol{U}\,\boldsymbol{\Sigma}\,\boldsymbol{V}^{\boldsymbol{T}}$. Let's take a look at the eigenpairs corresponding to $\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A}$:

$$\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A} = \left(\boldsymbol{U}\,\boldsymbol{\Sigma}\,\boldsymbol{V}^{\boldsymbol{T}}\right)^{\boldsymbol{T}}\left(\boldsymbol{U}\,\boldsymbol{\Sigma}\,\boldsymbol{V}^{\boldsymbol{T}}\right)$$

$$\left(\boldsymbol{V}^{\boldsymbol{T}}\right)^{\boldsymbol{T}}(\,\boldsymbol{\Sigma}\,)^{\boldsymbol{T}}\boldsymbol{U}^{\boldsymbol{T}}\left(\boldsymbol{U}\,\boldsymbol{\Sigma}\,\boldsymbol{V}^{\boldsymbol{T}}\right) = \boldsymbol{V}\boldsymbol{\Sigma}^{\boldsymbol{T}}\boldsymbol{U}^{\boldsymbol{T}}\,\boldsymbol{U}\,\boldsymbol{\Sigma}\,\boldsymbol{V}^{\boldsymbol{T}} = \boldsymbol{V}\,\boldsymbol{\Sigma}^{\boldsymbol{T}}\boldsymbol{\Sigma}\,\boldsymbol{V}^{\boldsymbol{T}}$$

Hence $\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A} = \boldsymbol{V}\,\boldsymbol{\Sigma}^{2}\,\boldsymbol{V}^{\boldsymbol{T}}$

Recall that columns of $\boldsymbol{V}$ are all linear independent (orthogonal matrix), then from diagonalization ($\boldsymbol{B} = \boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{-\boldsymbol{1}}$), we get:

- the columns of $\boldsymbol{V}$ are the eigenvectors of the matrix $\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A}$
- The diagonal entries of $\boldsymbol{\Sigma}^{\boldsymbol{2}}$ are the eigenvalues of $\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A}$

Let's call λ the eigenvalues of $\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A}$, then ${\sigma_i}^2 = \lambda_i$

Singular Value Decomposition

In a similar way,

$$\boldsymbol{AA^T} = \left(\boldsymbol{U\,\Sigma\,V^T}\right)\left(\boldsymbol{U\,\Sigma\,V^T}\right)^{\boldsymbol{T}}$$

$$\left(\boldsymbol{U\,\Sigma\,V^T}\right)\left(\boldsymbol{V^T}\right)^{\boldsymbol{T}}(\,\boldsymbol{\Sigma}\,)^{\boldsymbol{T}}\boldsymbol{U^T} = \boldsymbol{U\,\Sigma\,V^TV\Sigma^TU^T} = \boldsymbol{U\Sigma\,\Sigma^TU^T}$$

Hence $\boldsymbol{AA^T} = \boldsymbol{U\,\Sigma^2\,U^T}$

Recall that columns of $\boldsymbol{U}$ are all linear independent (orthogonal matrices), then from diagonalization ($\boldsymbol{B} = \boldsymbol{XDX^{-1}}$), we get:

- The columns of $\boldsymbol{U}$ are the eigenvectors of the matrix $\boldsymbol{AA^T}$

How can we compute an SVD of a matrix A ?

1. Evaluate the n eigenvectors $\mathbf{v}_i$ and eigenvalues λ_i of $\boldsymbol{A}^T\boldsymbol{A}$
2. Make a matrix $\boldsymbol{V}$ from the normalized vectors $\mathbf{v}_i$. The columns are called "right singular vectors".

$$\boldsymbol{V} = \begin{pmatrix} \vdots & \dots & \vdots \\ \mathbf{v}_1 & \dots & \mathbf{v}_n \\ \vdots & \dots & \vdots \end{pmatrix}$$

3. Make a diagonal matrix from the square roots of the eigenvalues.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{pmatrix} \quad \sigma_i = \sqrt{\lambda_i} \quad \text{and} \quad \sigma_1 \geq \sigma_2 \geq \sigma_3 \dots$$

4. Find $\boldsymbol{U}$: $\boldsymbol{A} = \boldsymbol{U}\,\boldsymbol{\Sigma}\,\boldsymbol{V}^T \Longrightarrow \boldsymbol{U}\,\boldsymbol{\Sigma} = \boldsymbol{A}\,\boldsymbol{V} \Longrightarrow \boldsymbol{U} = \boldsymbol{A}\,\boldsymbol{V}\,\boldsymbol{\Sigma}^{-1}$. The columns are called the "left singular vectors".

Singular Value Decomposition

Singular values cannot be negative since $\boldsymbol{A}^T\boldsymbol{A}$ **is a positive semi-definite matrix** (for real matrices $\boldsymbol{A}$)

- A matrix is positive definite if $\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} > \boldsymbol{0}$ for $\forall \boldsymbol{x} \neq \boldsymbol{0}$
- A matrix is positive semi-definite if $\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} \geq \boldsymbol{0}$ for $\forall \boldsymbol{x} \neq \boldsymbol{0}$

- What do we know about the matrix $\boldsymbol{A}^T\boldsymbol{A}$?

$$\boldsymbol{x}^T\ \boldsymbol{A}^T\boldsymbol{A}\ \boldsymbol{x} = (\boldsymbol{A}\boldsymbol{x})^T\boldsymbol{A}\boldsymbol{x} = \|\boldsymbol{A}\boldsymbol{x}\|_2^2 \geq 0$$

- Hence we know that $\boldsymbol{A}^T\boldsymbol{A}$ is a positive semi-definite matrix

- A positive semi-definite matrix has non-negative eigenvalues

$$\boldsymbol{B}\boldsymbol{x} = \lambda\boldsymbol{x} \Longrightarrow \boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} = \boldsymbol{x}^T\ \lambda\ \boldsymbol{x} = \lambda\ \|\boldsymbol{x}\|_2^2 \geq 0 \Longrightarrow \lambda \geq 0$$

Singular Value Decomposition

- The SVD is a factorization of a $m \times n$ matrix into $\boldsymbol{A} = \boldsymbol{U}\,\boldsymbol{\Sigma}\,\boldsymbol{V}^{\boldsymbol{T}}$ where $\boldsymbol{U}$ is a $m \times m$ orthogonal matrix, $\boldsymbol{V}^{\boldsymbol{T}}$ is a $n \times n$ orthogonal matrix and $\boldsymbol{\Sigma}$ is a $m \times n$ diagonal matrix.

- In reduced form: $\boldsymbol{A} = \boldsymbol{U_R \Sigma_R V_R}^{\boldsymbol{T}}$, where $\boldsymbol{U_R}$ is a $m \times k$ matrix, $\boldsymbol{\Sigma_R}$ is a $k \times k$ matrix, and $\boldsymbol{V_R}$ is a $n \times k$ matrix, and $k = \min(m, n)$.

- The columns of $\boldsymbol{V}$ are the eigenvectors of the matrix $\boldsymbol{A^T A}$, denoted the right singular vectors.

- The columns of $\boldsymbol{U}$ are the eigenvectors of the matrix $\boldsymbol{AA^T}$, denoted the left singular vectors.

- The diagonal entries of $\boldsymbol{\Sigma}^{\boldsymbol{2}}$ are the eigenvalues of $\boldsymbol{A^T A}$. $\sigma_i = \sqrt{\lambda_i}$ are called the singular values.

- The singular values are always non-negative (since $\boldsymbol{A^T A}$ is a positive semi-definite matrix, the eigenvalues are always $\lambda \geq 0$)

Low-Rank Approximation

Another way to write the SVD (assuming for now $m > n$ for simplicity)

$$\boldsymbol{A} = \begin{pmatrix} \boxed{\begin{matrix} \vdots & \dots \\ \boldsymbol{u}_1 & \dots \\ \vdots & \dots \end{matrix}} & \begin{matrix} \vdots \\ \boldsymbol{u}_m \\ \vdots \end{matrix} \end{pmatrix} \begin{pmatrix} \boxed{\begin{matrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{matrix}} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{pmatrix} \dots & \mathbf{v}_1^T & \dots \\ \vdots & \vdots & \vdots \\ \dots & \mathbf{v}_n^T & \dots \end{pmatrix}$$

$$= \begin{pmatrix} \vdots & \dots & \vdots \\ \boldsymbol{u}_1 & \dots & \boldsymbol{u}_n \\ \vdots & \dots & \vdots \end{pmatrix} \begin{pmatrix} \dots & \sigma_1 \mathbf{v}_1^T & \dots \\ \vdots & \vdots & \vdots \\ \dots & \sigma_n \mathbf{v}_n^T & \dots \end{pmatrix}$$

$$= \sigma_1 \boldsymbol{u}_1 \mathbf{v}_1^T + \sigma_2 \boldsymbol{u}_2 \mathbf{v}_2^T + \cdots + \sigma_n \boldsymbol{u}_n \mathbf{v}_n^T$$

The SVD writes the matrix A as a sum of outer products (of left and right singular vectors).

Low-Rank Approximation

The best **rank-$\boldsymbol{k}$** approximation for a $m \times n$ matrix $\boldsymbol{A}$, (where $k \leq min(m, n)$) is the one that minimizes the following problem:

$$\min_{A_k} \|A - A_k\|$$
$$\text{such that} \quad \text{rank}(A_k) \leq k.$$

When using the induced 2-norm, the best **rank-$\boldsymbol{k}$** approximation is given by:

$$\boldsymbol{A}_k = \sigma_1 \boldsymbol{u}_1 \mathbf{v}_1^T + \sigma_2 \boldsymbol{u}_2 \mathbf{v}_2^T + \cdots + \sigma_k \boldsymbol{u}_k \mathbf{v}_k^T$$

$$\sigma_1 \geq \sigma_2 \geq \sigma_3 \ldots \geq 0$$

Note that $rank(\boldsymbol{A}) = n$ and $rank(\boldsymbol{A}_k) = k$ and the norm of the difference between the matrix and its approximation is

$$\|\boldsymbol{A} - \boldsymbol{A}_k\|_2 = \left\|\sigma_{k+1} \boldsymbol{u}_{k+1} \mathbf{v}_{k+1}^T + \sigma_{k+2} \boldsymbol{u}_{k+2} \mathbf{v}_{k+2}^T + \cdots + \sigma_n \boldsymbol{u}_n \mathbf{v}_n^T\right\|_2$$

Image compression

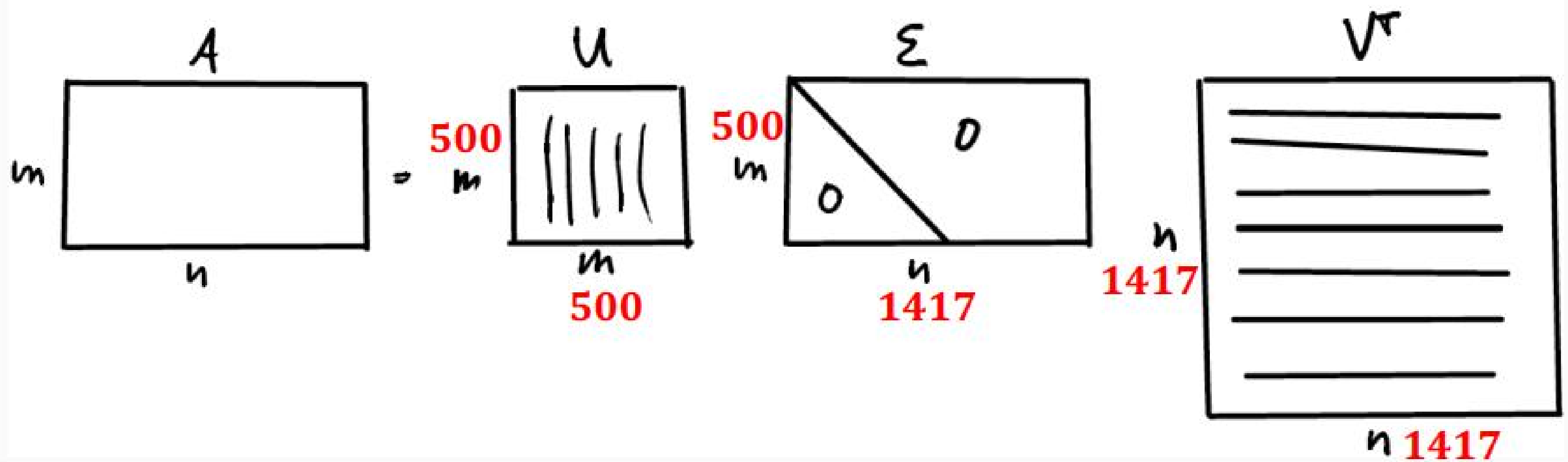
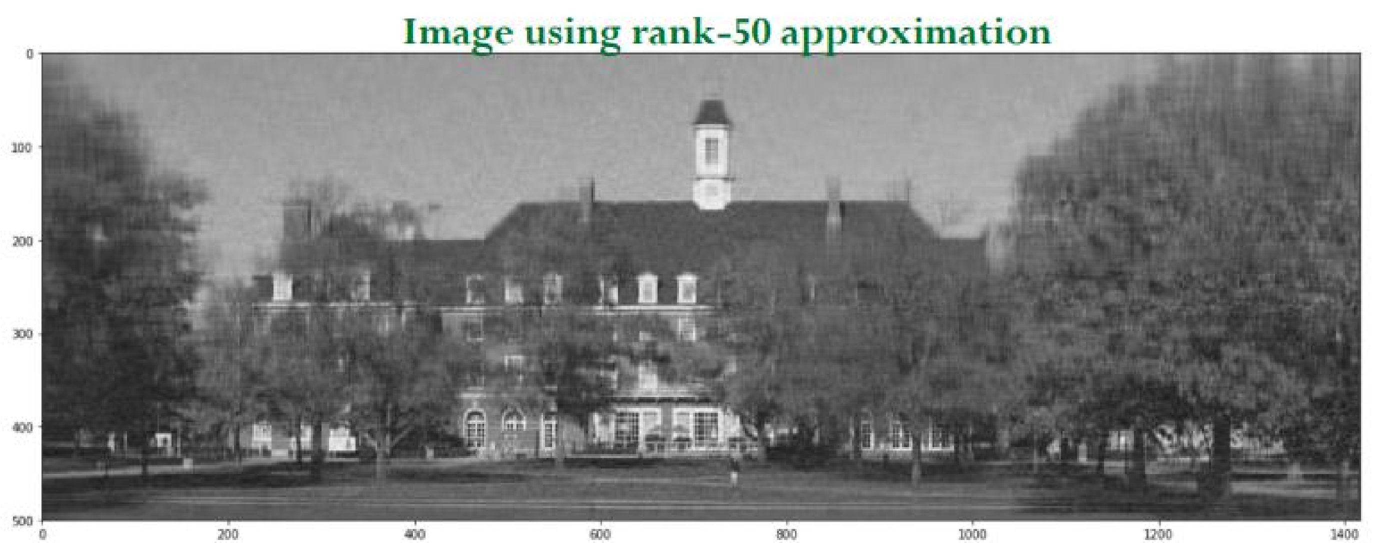


Image compression

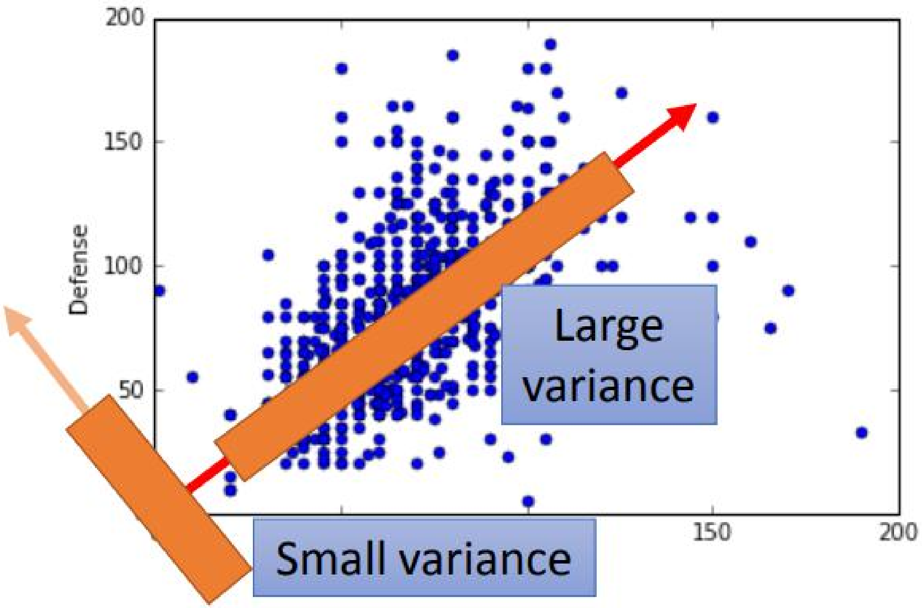


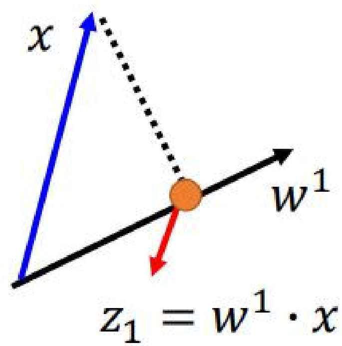
Principal Component Analysis

$$z = Wx$$

Reduce to 1-D:

$$z_1 = w^1 \cdot x$$





Project all the data points x onto w^1, and obtain a set of z_1

We want the variance of z_1 as large as possible

$$Var(z_1) = \sum_{z_1} (z_1 - \bar{z_1})^2 \quad \|w^1\|_2 = 1$$

Principal Component Analysis

$$z = Wx$$

Reduce to 1-D:

$$z_1 = w^1 \cdot x$$

$$z_2 = w^2 \cdot x$$

$$W = \begin{bmatrix} (w^1)^T \\ (w^2)^T \\ \vdots \end{bmatrix}$$

Orthogonal matrix

Project all the data points x onto w^1, and obtain a set of z_1

We want the variance of z_1 as large as possible

$$Var(z_1) = \sum_{z_1} (z_1 - \bar{z_1})^2 \quad \|w^1\|_2 = 1$$

We want the variance of z_2 as large as possible

$$Var(z_2) = \sum_{z_2} (z_2 - \bar{z_2})^2 \quad \|w^2\|_2 = 1$$

$$w^1 \cdot w^2 = 0$$

Principal Component Analysis

$$z_1 = w^1 \cdot x$$

$$\bar{z_1} = \frac{1}{N}\sum z_1 = \frac{1}{N}\sum w^1 \cdot x = w^1 \cdot \frac{1}{N}\sum x = w^1 \cdot \bar{x}$$

$$Var(z_1) = \frac{1}{N}\sum_{z_1}(z_1 - \bar{z_1})^2$$

$$= \frac{1}{N}\sum_{x}(w^1 \cdot x - w^1 \cdot \bar{x})^2$$

$$= \frac{1}{N}\sum\left(w^1 \cdot (x - \bar{x})\right)^2$$

$$= \frac{1}{N}\sum (w^1)^T(x - \bar{x})(x - \bar{x})^T w^1$$

$$= (w^1)^T \boxed{\frac{1}{N}\sum (x - \bar{x})(x - \bar{x})^T} w^1$$

$$= (w^1)^T Cov(x) w^1$$

$S = Cov(x)$

$$(a \cdot b)^2 = (a^T b)^2 = a^T b a^T b$$

$$= a^T b (a^T b)^T = a^T b b^T a$$

Find w^1 maximizing

$$(w^1)^T S w^1$$

$$\|w^1\|_2 = (w^1)^T w^1 = 1$$

Principal Component Analysis

Find w^1 maximizing $(w^1)^T S w^1$ $\quad (w^1)^T w^1 = 1$

$S = Cov(x)$ Symmetric positive-semidefinite (non-negative eigenvalues)

Using Lagrange multiplier [Bishop, Appendix E]

$$g(w^1) = (w^1)^T S w^1 - \alpha\left((w^1)^T w^1 - 1\right)$$

$$\left.\begin{aligned} &\partial g(w^1)/\partial w_1^1 = 0 \\ &\partial g(w^1)/\partial w_2^1 = 0 \\ &\vdots \end{aligned}\right\}$$

$Sw^1 - \alpha w^1 = 0$

$Sw^1 = \alpha w^1$ $\quad w^1$: eigenvector

$(w^1)^T S w^1 = \alpha (w^1)^T w^1$

$= \alpha$ $\quad$ Choose the maximum one

w^1 is the eigenvector of the covariance matrix S

Corresponding to the largest eigenvalue λ_1

Principal Component Analysis

Find w^2 maximizing $(w^2)^T S w^2 \quad (w^2)^T w^2 = 1 \quad (w^2)^T w^1 = 0$

$$g(w^2) = (w^2)^T S w^2 - \alpha\left((w^2)^T w^2 - 1\right) - \beta\left((w^2)^T w^1 - 0\right)$$

$$\left.\begin{array}{l} \partial g(w^2)/\partial w_1^2 = 0 \\ \partial g(w^2)/\partial w_2^2 = 0 \\ \vdots \end{array}\right\} \quad Sw^2 - \alpha w^2 - \beta w^1 = 0$$

$$\boxed{0} - \alpha \boxed{0} - \beta \boxed{1} = 0$$

$$= \left((w^1)^T S w^2\right)^T = (w^2)^T S^T w^1$$

$$= (w^2)^T S w^1 = \lambda_1 (w^2)^T w^1 = 0$$

$$\boxed{Sw^1 = \lambda_1 w^1}$$

$\beta = 0$: $\quad Sw^2 - \alpha w^2 = 0 \quad Sw^2 = \alpha w^2$

w^2 is the eigenvector of the covariance matrix S

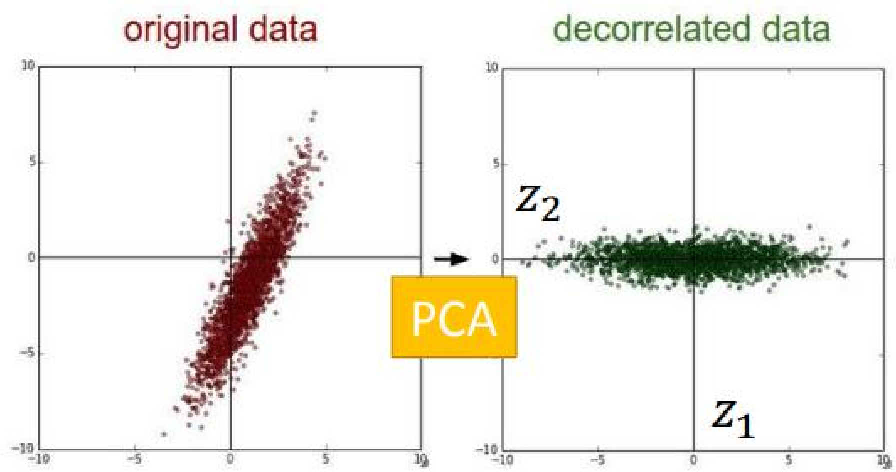
Corresponding to the 2nd largest eigenvalue λ_2

Principal Component Analysis

$z = Wx$

$Cov(z) = D$

Diagonal matrix



$$Cov(z) = \frac{1}{N}\sum (z-\bar{z})(z-\bar{z})^T = WSW^T \qquad S = Cov(x)$$

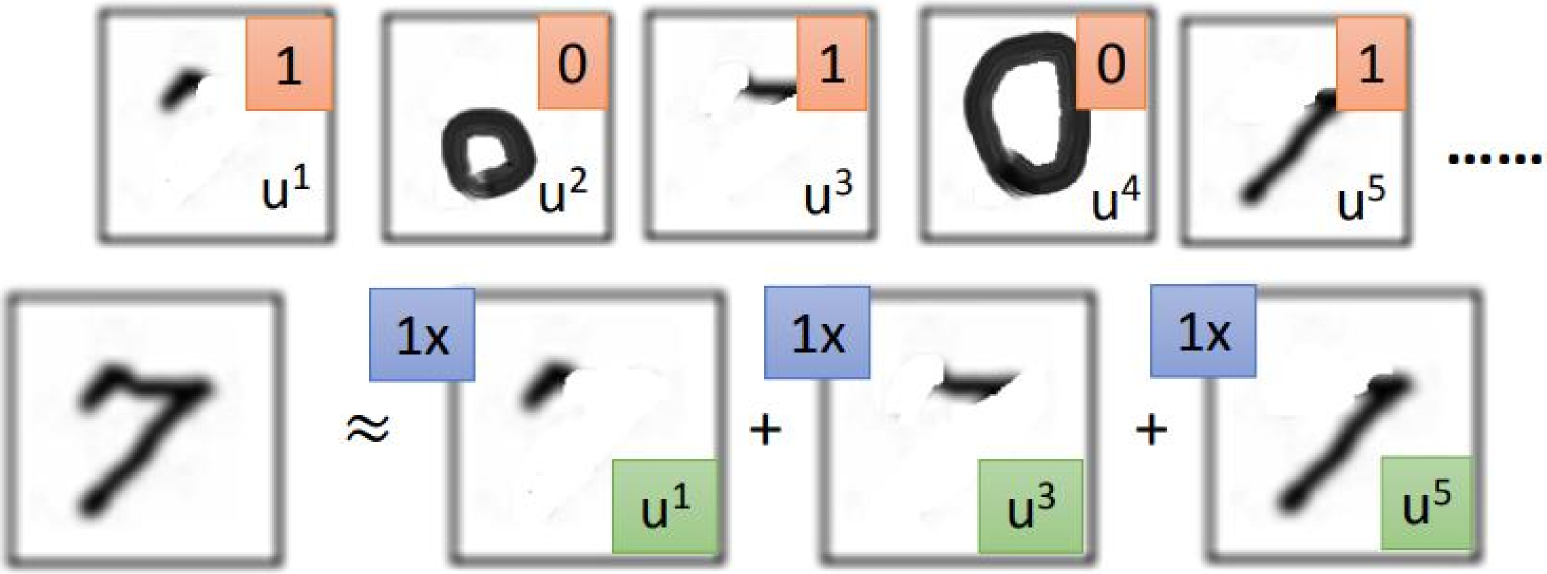
$$= WS\begin{bmatrix} w^1 & \cdots & w^K \end{bmatrix} = W\begin{bmatrix} Sw^1 & \cdots & Sw^K \end{bmatrix}$$

$$= W\begin{bmatrix} \lambda_1 w^1 & \cdots & \lambda_K w^K \end{bmatrix} = \begin{bmatrix} \lambda_1 W w^1 & \cdots & \lambda_K W w^K \end{bmatrix}$$

$$= \begin{bmatrix} \lambda_1 e_1 & \cdots & \lambda_K e_K \end{bmatrix} = D$$ Diagonal matrix

Principal Component Analysis

Basic Component:



$$\begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ \vdots \end{bmatrix}$$

$$x \approx c_1 u^1 + c_2 u^2 + \cdots + c_K u^K + \bar{x}$$

Pixels in a digit image

component

$$\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_K \end{bmatrix}$$

Represent a digit image

Principal Component Analysis

$$x - \bar{x} \approx c_1 u^1 + c_2 u^2 + \cdots + c_K u^K \ = \hat{x}$$

Reconstruction error:
$\| (x - \bar{x}) - \hat{x} \|_2$

Find $\{u^1, \dots, u^K\}$ minimizing the error

$$L = \min_{\{u^1,\dots,u^K\}} \sum \left\| (x - \bar{x}) - \underbrace{\left(\sum_{k=1}^{K} c_k u^k \right)}_{\hat{x}} \right\|_2$$

PCA: $z = Wx$

$$\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_K \end{bmatrix} = \begin{bmatrix} (w_1)^{\mathrm{T}} \\ (w_2)^{\mathrm{T}} \\ \vdots \\ (w_K)^{\mathrm{T}} \end{bmatrix} x$$

$\{w^1, w^2, \dots w^K\}$ is the component $\{u^1, u^2, \dots u^K\}$ minimizing L

Proof in [Bishop, Chapter 12.1.2]

Principal Component Analysis

$$x - \bar{x} \approx c_1 u^1 + c_2 u^2 + \cdots + c_K u^K = \hat{x}$$

Reconstruction error:
$\| (x - \bar{x}) - \hat{x} \|_2$

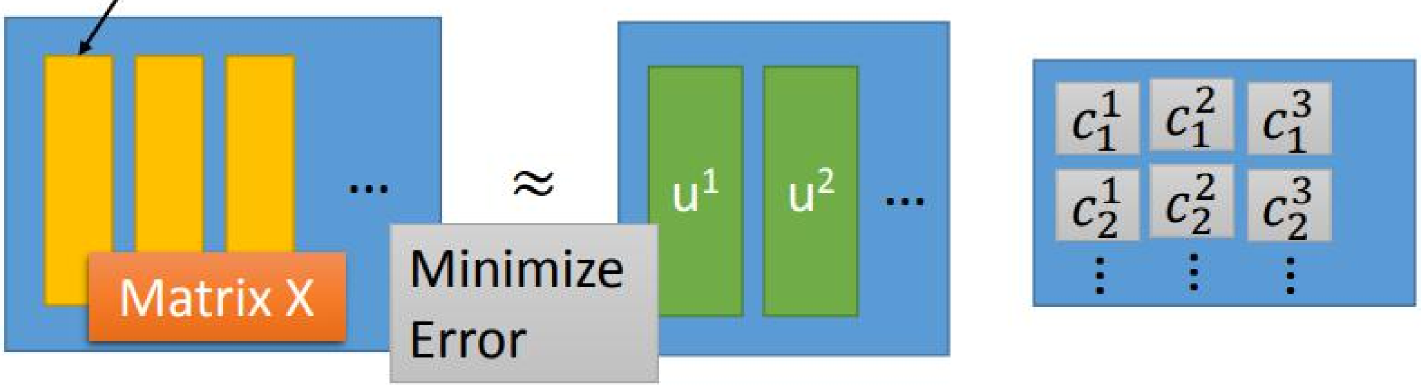
Find $\{u^1, \dots, u^K\}$ minimizing the error

$$\underline{x^1 - \bar{x}} \approx \underline{c_1^1}\,\underline{u^1} + \underline{c_2^1}\,\underline{u^2} + \cdots$$

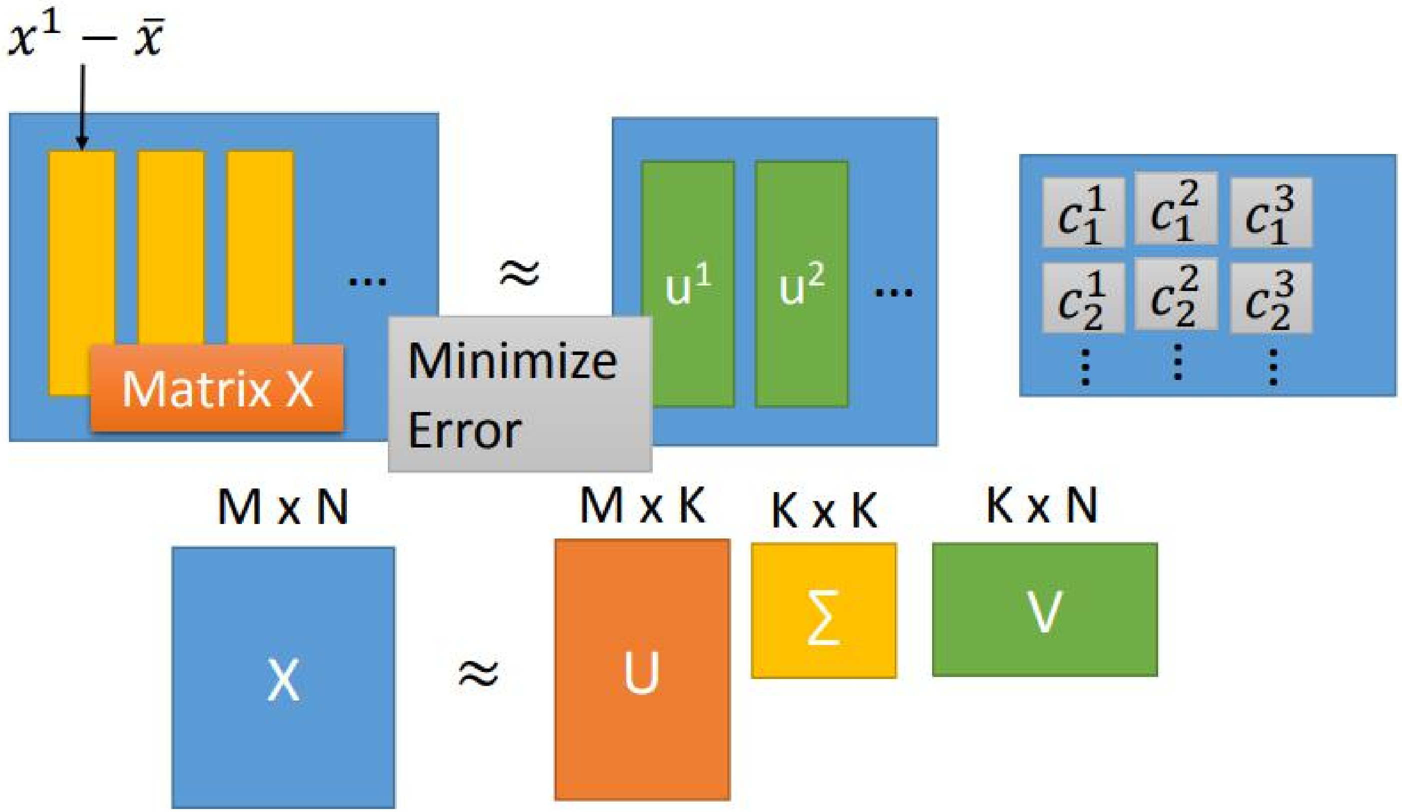
$$x^2 - \bar{x} \approx c_1^2 u^1 + c_2^2 u^2 + \cdots$$

$$x^3 - \bar{x} \approx c_1^3 u^1 + c_2^3 u^2 + \cdots$$

$$\vdots$$



Principal Component Analysis

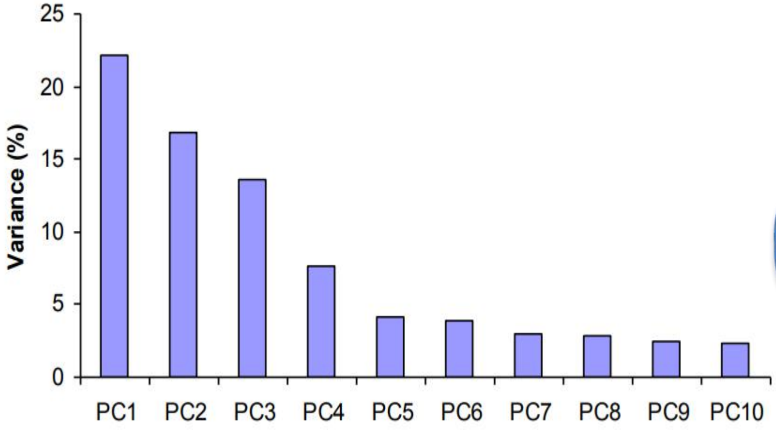


K columns of U: a set of orthonormal eigen vectors corresponding to the k largest eigenvalues of XX^T

This is the solution of PCA

Additional remarks

- How many PCs?
  - We want to retain as much information as possible using these components.
  - We can compute each PC explains how much variance and then makes decision (still a parameter)



$$\frac{\lambda_k}{\sum_{i=1}^{N} \lambda_i}$$ Propotion of variance

$$\frac{\sum_{k=1}^{d} \lambda_k}{\sum_{i=1}^{N} \lambda_i}$$ Cumulative propotion

Weakness of PCA

To be continued . . .

Kernel PCA、Probabilistic PCA

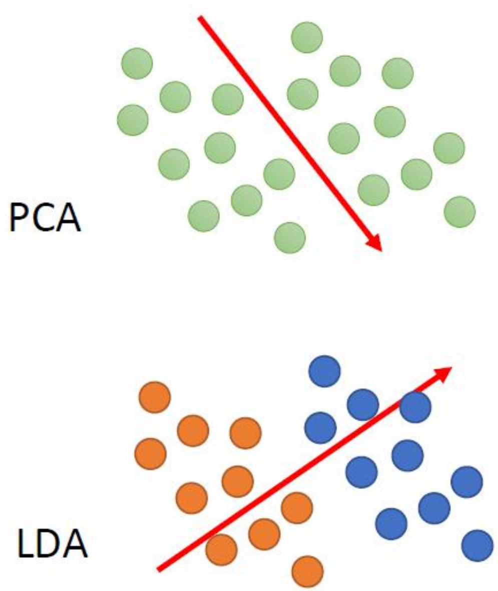
Linear Discriminant Analysis (LDA)

Matrix factorization

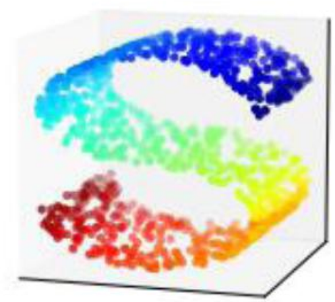
Canonical Correlation Analysis (CCA)

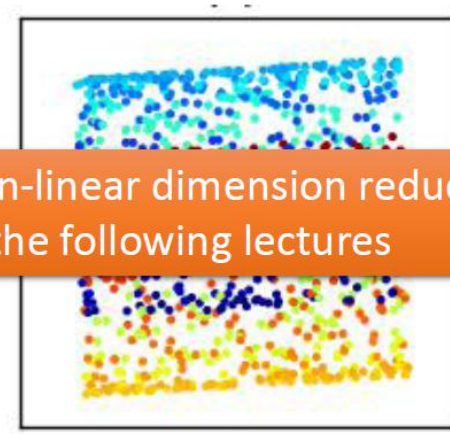
Deep Autoencoder . . .

- Unsupervised



- Linear



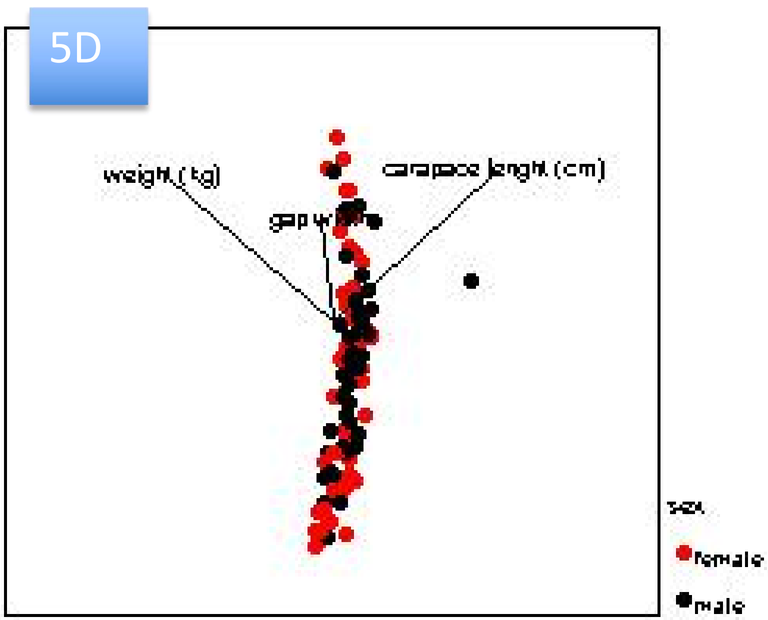


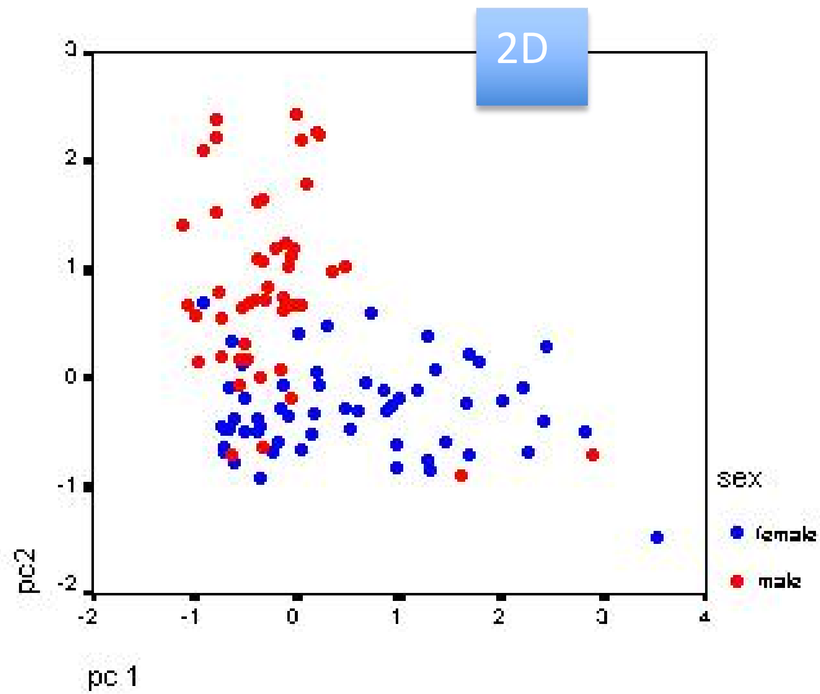
Non-linear dimension reduction in the following lectures

http://www.astroml.org/book_figures/chapter7/fig_S_manifold_PCA.html

Example: The data matrix

case	ht (x_1)	wt(x_2)	age(x_3)	sbp(x_4)	heart rate (x_5)
1	175	1225	25	117	56
2	156	1050	31	122	63
m	202	1350	58	154	67





Allow us choose small number of uncorrelated varies to perform machine learning tasks

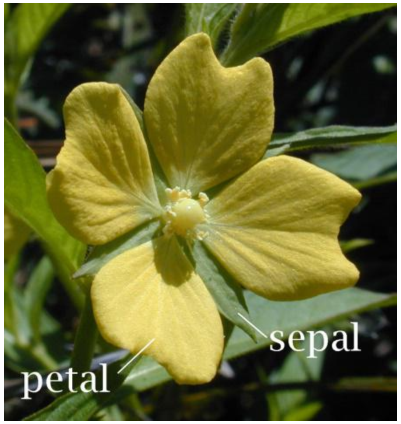
Dimensionality Reduction: PCA

Dimensions = 206

Example of a data: Iris Flower Data Set

- Many of the exploratory data techniques are illustrated with the famous ***Iris Flower*** data set (a.k.a. "**Iris**").
  - Available at the UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
  - From the statistician R.A. Fisher
  - Three flower types (classes):
    - Iris Setosa
    - Iris Versicolour
    - Iris Virginica
  - Four (non-class) attributes
    - Sepal width
    - Sepal length
    - Petal width
    - Petal length
  - Total number Instances = 150

https://en.wikipedia.org/wiki/Iris_flower_data_set




Iris setosa


Iris versicolor


Iris virginica

R Example using Iris data

- Iris

```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

- irisPCA<-**princomp**(iris[-5]) # Exclude Species and perform PCA
- summary(irisPCA)

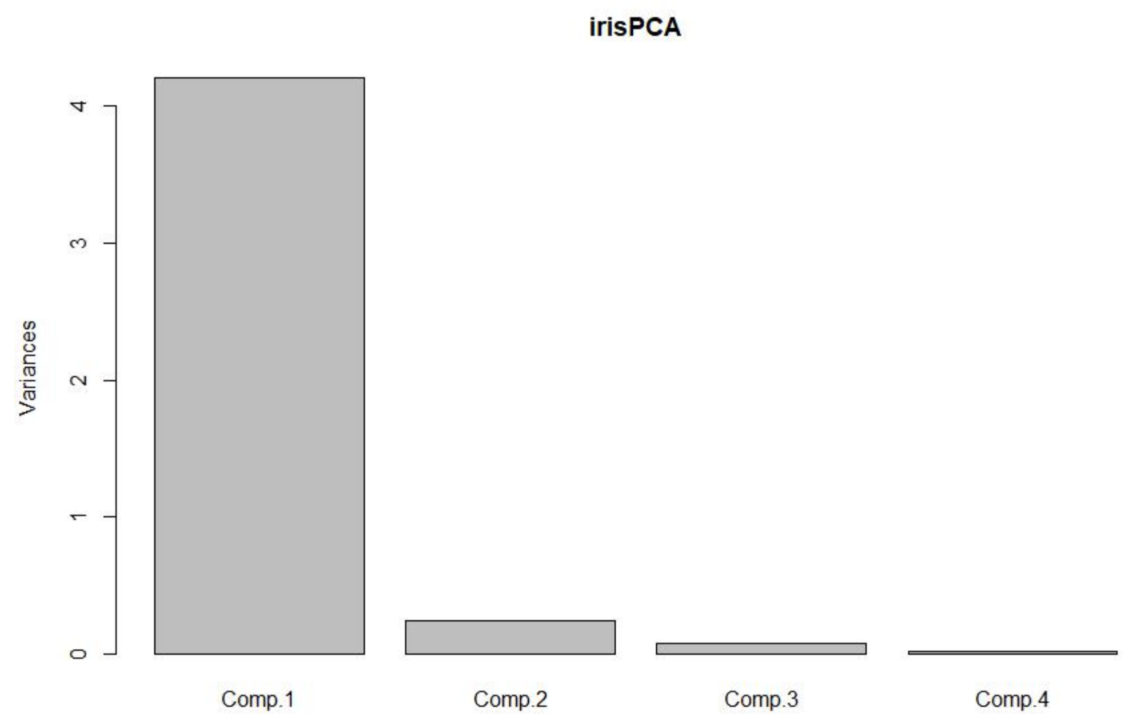
PC1 PC2 PC3 PC4

```
>
> summary(irisPCA)
Importance of components:
                          Comp.1     Comp.2     Comp.3      Comp.4
Standard deviation     2.0494032 0.49097143 0.27872586 0.153870700
Proportion of Variance 0.9246187 0.05306648 0.01710261 0.005212184
Cumulative Proportion  0.9246187 0.97768521 0.99478782 1.000000000
```

92.5% of variation is explained by PC1 alone; 97.8% is explained by PC1 and PC2

Screen plot

- It shows the proportion of the total variation that is explained by each of the components. Perhaps 1 or 2 PC2 will be sufficient
- screeplot(irisPCA)



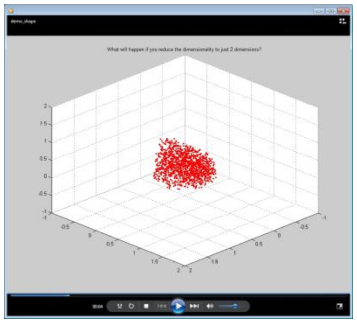
While **dimensionality reduction** is an important tool in machine learning/data mining, we must always be aware that it can *distort* the data in misleading ways.

Above is a two dimensional projection of an intrinsically three dimensional world….

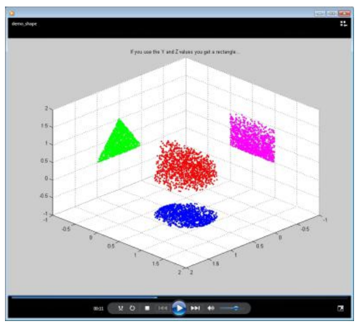
Original photographer unknown/
See also www.cs.gmu.edu/~jessica/DimReducDanger.htm

We may lose some important information when we perform feature selection

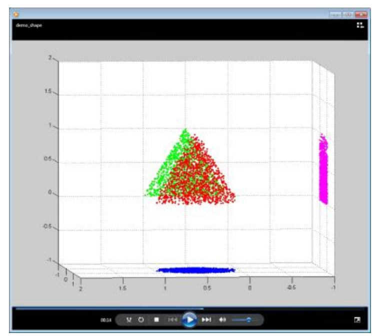
A cloud of points in 3D



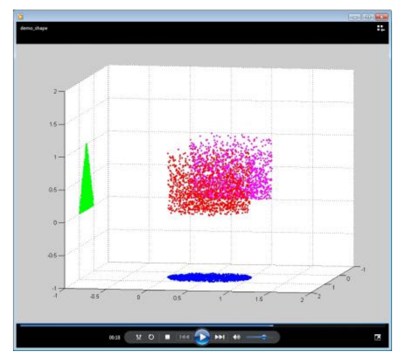
Can be projected into 2D
XY or XZ or YZ



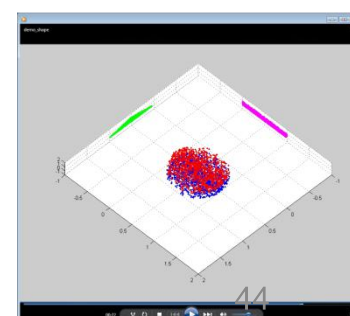
In 2D XZ we see a triangle



In 2D YZ we see a square



In 2D XY we see a circle



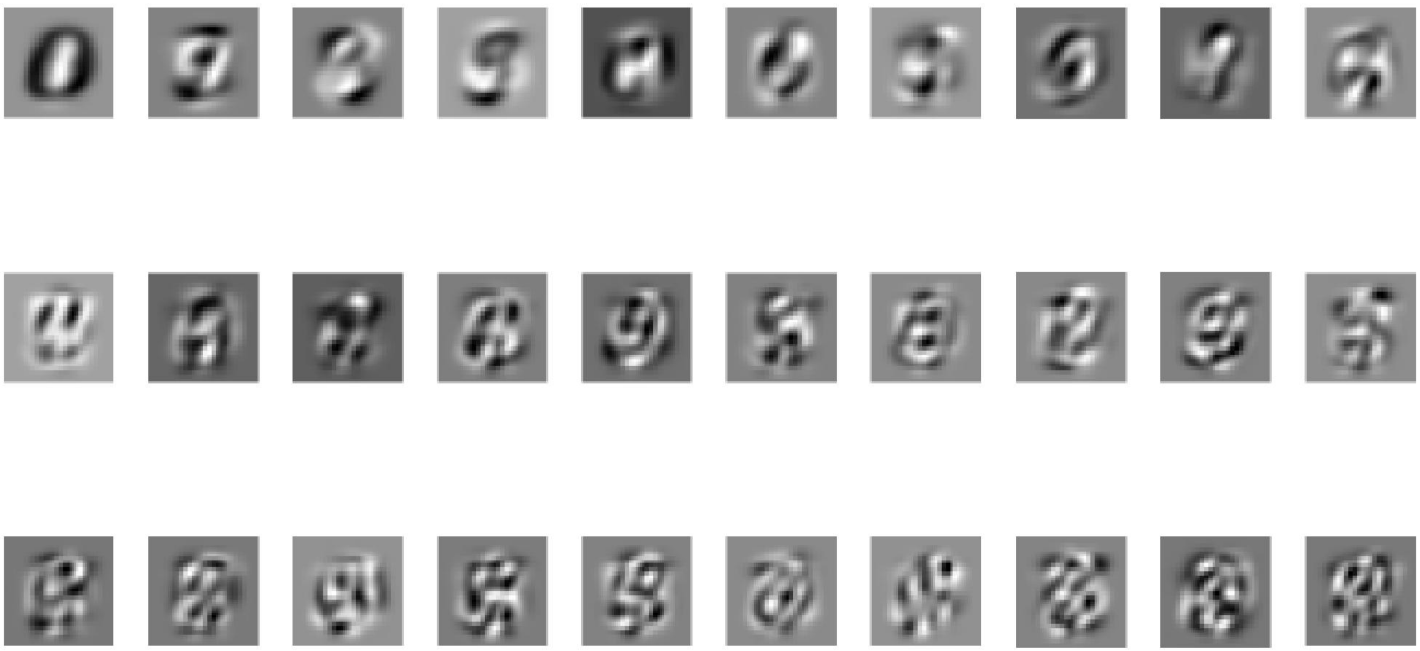
Screen dumps of a short video from
www.cs.gmu.edu/~jessica/DimReducDanger.htm

Principal Component Analysis

MNIST  $= a_1 w^1 + a_2 w^2 + \cdots$

images

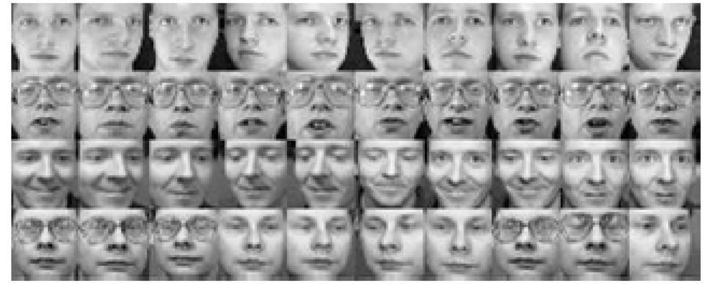
30 components:



Eigen-digits

Principal Component Analysis

Face



30 components:

http://www.cs.unc.edu/~lazebnik/research/spring08/assignment3.html

Eigen-face

Principal Component Analysis

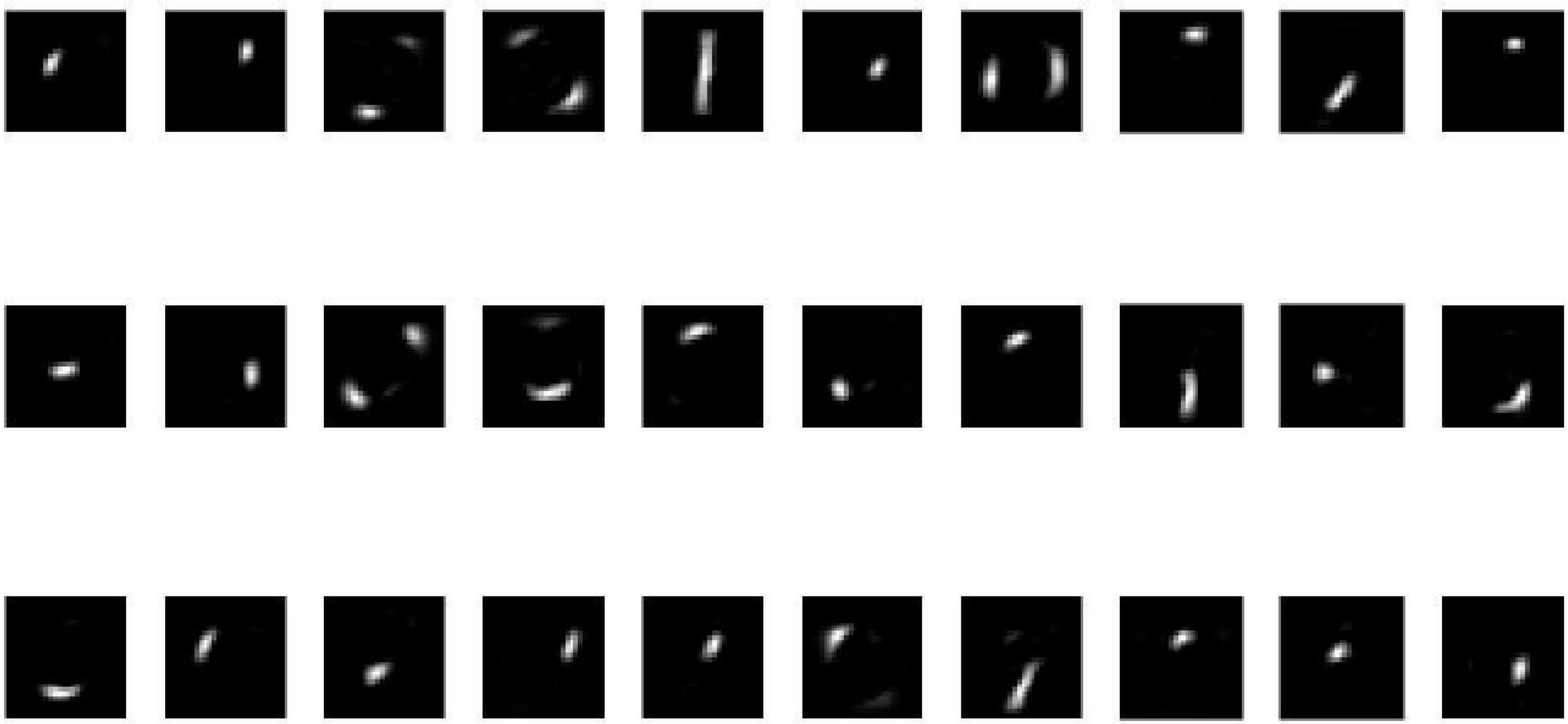
$$= a_1 w^1 + a_2 w^2 + \cdots$$

Can be any real number

- PCA involves adding up and subtracting some components (images)
  - Then the components may not be "parts of digits"
- Non-negative matrix factorization (NMF)
  - Forcing a_1, a_2 ...... be non-negative
    - additive combination
  - Forcing w^1, w^2...... be non-negative
    - More like "parts of digits"
- Ref: Daniel D. Lee and H. Sebastian Seung. "Algorithms for non-negative matrix factorization." *Advances in neural information processing systems*. 2001.

Principal Component Analysis

NMF on MNIST



Principal Component Analysis

NMF on Face