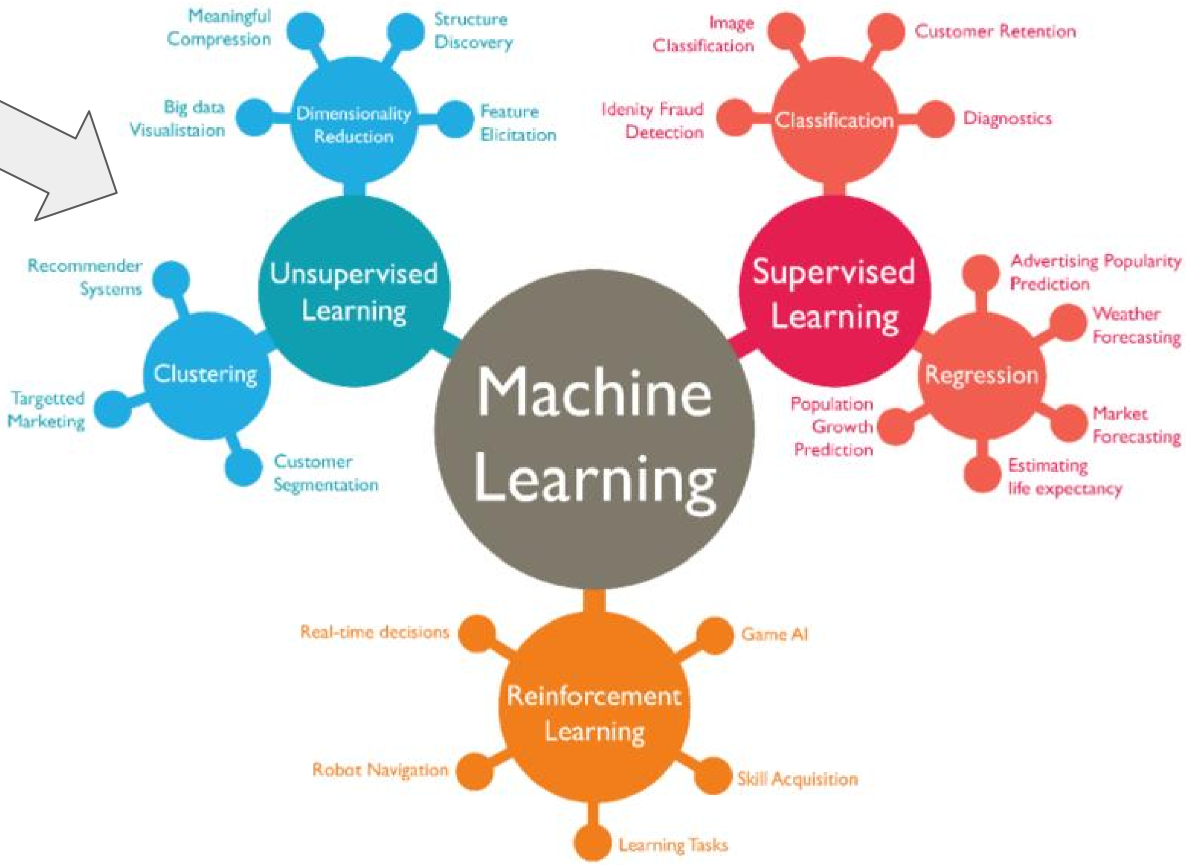
CS 182
Discussion 6

12/14/2023

Yuanming Shao

shaoym1@shanghaitech.edu.cn

Recall: Taxonomy of Machine Learning



Unsupervised Learning

The goal of **unsupervised learning** is to learn structure from **unlabeled** data.

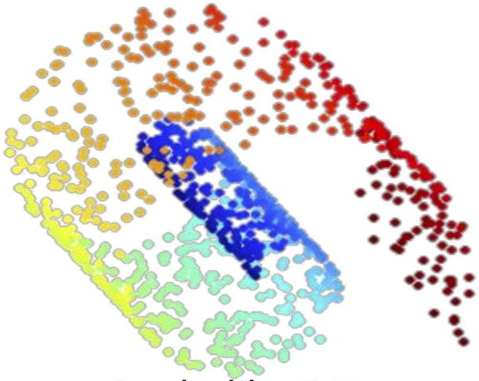
Tasks:

- **Clustering** = partition data into similar points.
  - Find a k-partition of a set of points $X = X_1 \cup X_1 \cup \dots X_k$ such that points within a subset (cluster) of points are "close" to each other.
- **Dimensionality reduction** = project high-dimensional points into a low-dimensional subspace.
  - Find a mapping $f : R^n \rightarrow R^d$ with $n > d$, such that if x_i is close to x_j, then $f(x_i)$ is close to $f(x_j)$.
- **Density estimation** = fit a continuous distribution to discrete data.
  - E.g. fitting a Gaussian to data via MLE.

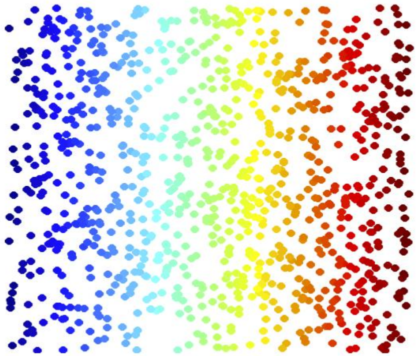
- Let V be vector space over the same field F, $T: V \to V$ be a linear mapping. We call $\boldsymbol{v} \in V$ an ***eigenvector*** of T if $\boldsymbol{v} \neq \mathbf{0}$ and
$$T(\boldsymbol{v}) = \lambda \boldsymbol{v}$$
where $\lambda \in F$. We call λ the ***eigenvalue*** associated with $\boldsymbol{v}$.
- Let F be a field ($\mathbb{R}$ or $\mathbb{C}$). A matrix $A \in F^{n \times n}$ can be viewed as a linear mapping $T: F^n \to F^n$, $T(\boldsymbol{x}) = A\boldsymbol{x}$. Hence we call λ the eigenvalue (of A) associated with eigenvector (of A) $\boldsymbol{v} \in F^n$ if $\boldsymbol{v} \neq \mathbf{0}$ and
$$A\boldsymbol{v} = \lambda \boldsymbol{v}$$
- Exercise: Eigenvector and eigenvalue of
$$A = \begin{bmatrix} -1 & 6 \\ 3 & 2 \end{bmatrix}$$

Dimension Reduction

vector x (High Dim) → function → vector z (Low Dim)



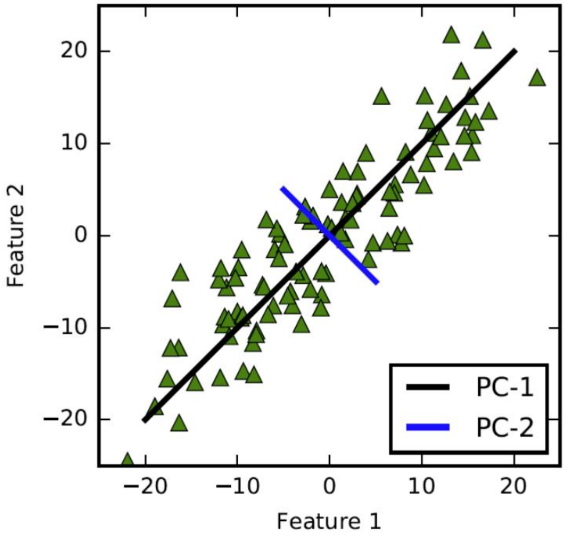
Looks like 3-D



Actually, 2-D

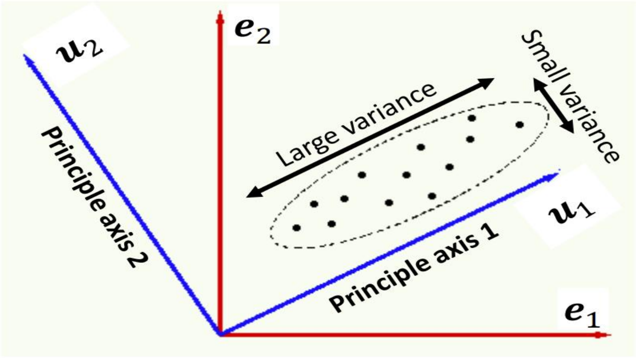
Principal Component Analysis (PCA)

- PCA's target: finding the best lower dimensional sub-space that conveys most of the variance in the original data
- Example: If we were to compress 2-D data to 1-D subspace, then PCA prefers projecting to the ***black*** line, since it preserves more variance comparing to ***blue*** line.



Principle Axes

- Objective of PCA: Given data in $\mathbb{R}^M$, want to *rigidly rotate* the axes to new positions (principle axes) with the following properties:
  - ➢ *Ordered such that principle axis 1 has the highest variance, axis 2 has the next highest variance, ..., and axis M has the lowest variance.*
  - ➢ *Covariance among each pair of the principal axes is zero.*
- The k'th ***principle component*** is the projection to the k'th principle axis.
- Keep the first $m < M$ principle components for dimensionality reduction.



- Given N data $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in \mathbb{R}^M$, PCA first computes the covariance matrix for the data

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T = \boldsymbol{U\Lambda U}^T$$

where $\boldsymbol{\mu} \in \mathbb{R}^M$ is the data mean.

- Since $\boldsymbol{\Sigma}$ is symmetric, $\boldsymbol{\Sigma}$ can be written as $\boldsymbol{\Sigma} = \boldsymbol{U\Lambda U}^T$, where $\boldsymbol{U} = [\boldsymbol{u}_1 \dots \boldsymbol{u}_{\mathrm{M}}]$ is ***orthogonal*** matrix of eigenvectors (of $\boldsymbol{\Sigma}$), $\boldsymbol{\Lambda} = diag(\lambda_1, \dots, \lambda_M)$ is diagonal matrix of the associated eigenvalues arranged in non-ascending order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M \geq 0$. (Note that all eigenvalues are non-negative real scalars since $\boldsymbol{\Sigma}$ is ***semi-positive definite***.)

- For data $\boldsymbol{x} \in \mathbb{R}^M$, compute its 1st principle component as $\boldsymbol{u}_1^T\boldsymbol{x}$, 2nd principle component as $\boldsymbol{u}_2^T\boldsymbol{x}$,…, M’th principle component as $\boldsymbol{u}_M^T\boldsymbol{x}$

Given N data points $X_1, X_2 \ldots X_N \in \mathbb{R}^m$

Compute the covariance matrix of data

$$\Sigma = \frac{1}{N} \sum_{i=1}^{N} (X_i - \mu)(X_i - \mu)^T \in \mathbb{R}^{m \times m}$$

$$\mu = \frac{1}{N} \sum_{i=1}^{N} X_i$$

$\because \Sigma$ is symmetric, it can be diagnalized as.

$$\Sigma = U \Lambda U^T = \lambda_i u_i v_i^T$$

where $U = [u_1, u_2, \quad u_m] \in R^{m \times m}$

$$\Lambda = \begin{bmatrix} \lambda_1 & & \\ & & \\ & & \lambda_m \end{bmatrix} \in \mathbb{R}^{m \times m}$$

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0$$

Orthogonal matrix:

$\boldsymbol{U} = [\boldsymbol{u}_1 \ldots \boldsymbol{u}_\mathrm{M}] \in \mathbb{R}^{M \times M}$ is an orthogonal matrix if $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_\mathrm{M}$ are orthogonal and have unit length

$$\boldsymbol{u}_\mathrm{i}^T \boldsymbol{u}_\mathrm{j} = \begin{cases} 1 & \text{if i} = j \\ 0 & \text{if } i \neq \text{j} \end{cases}$$

That is, $\boldsymbol{U}^T \boldsymbol{U} = \boldsymbol{I}$, namely, $\boldsymbol{U}^{-1} = \boldsymbol{U}^T$.

Positive definite:

$\boldsymbol{\Sigma} \in \mathbb{R}^{M \times M}$ is positive semi-definite if $\boldsymbol{x}^T \boldsymbol{\Sigma} \boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^M$. If the equality holds only when $\boldsymbol{x} = \boldsymbol{0}$, then $\boldsymbol{\Sigma}$ is positive definite.

$\Sigma = U\Lambda U^T$

$\underline{\Sigma v_i} = \underline{U\Lambda U^T v_i} = U\Lambda e_i = \underbrace{U e_i}_{\lambda_i \times}$

v_i是什么？ $\underline{= \lambda_i v_i}$

$$\begin{bmatrix} v_1^T \\ v_2^T \\ \\ v_m^T \end{bmatrix} v_i = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} = e_i$$

Σ ?

$\Lambda e_i = \lambda_i e_i$

$y^T \Sigma y$

$= y^T \frac{1}{N}\sum_{i=1}^{N}(x_i - m)(x_i - m)^T y$

$= \frac{1}{N}\sum_{i=1}^{N} y^T (x_i - m)(x_i - m)^T y$

$= \frac{1}{N}\sum_{i=1}^{N} (y^T(x_i - m))^2 \geq 0$

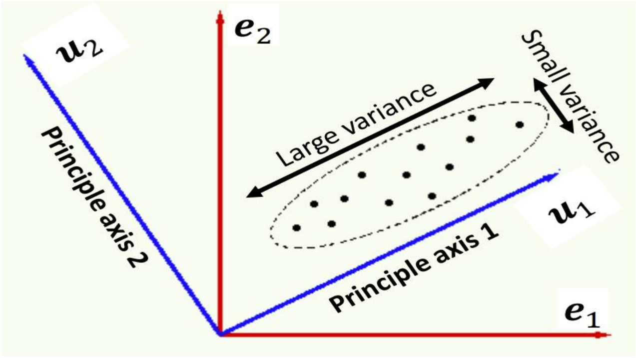
$0 \leq v_i^T \Sigma u_i = v_i^T \lambda_i u_i$

$= \lambda_i \|u_i\|^2$

$= \lambda_i$

Principle Axes

- Objective of PCA: Given data in $\mathbb{R}^M$, want to *rigidly rotate* the axes to new positions (principle axes) with the following properties:
  - ➢ *Ordered such that principle axis 1 has the highest variance, axis 2 has the next highest variance, ..., and axis M has the lowest variance.*
  - ➢ *Covariance among each pair of the principal axes is zero.*
- The k'th ***principle component*** is the projection to the k'th principle axis.
- Keep the first $m < M$ principle components for dimensionality reduction.



$$\Sigma = U \Lambda U^T = \begin{bmatrix} u_1 & & u_m \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & & \\ & & \lambda_m \end{bmatrix} \begin{bmatrix} u_1^T \\ \\ u_m^T \end{bmatrix}$$

$$= \sum_{i=1}^{m} \lambda u_i u_i^T$$

from M to m

$$\Rightarrow T: x \in \mathbb{R}^M \mapsto \begin{bmatrix} u_1^T x \\ u_2^T x \\ \\ u_m^T x \end{bmatrix} \in \mathbb{R}^m$$

$$\hat{x}_i = T(x_i) \in \mathbb{R}^{\boxed{m}}$$

$$T(x_i) = \begin{bmatrix} u_1^T x_i \\ \\ \\ u_m^T x_i \end{bmatrix}$$

$$\hat{x}_i \to \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N$$

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} \hat{x}_i = \frac{1}{N}\sum_{i=1}^{n} \begin{bmatrix} u_1^T \\ u_m^T \end{bmatrix} x_i = \begin{bmatrix} u_1^T \\ u_m^T \end{bmatrix} \mu.$$

$$(AB)^T = B^T A^T.$$

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{n} (\hat{x}_i - \hat{\mu})(\hat{x}_i - \hat{\mu})^T$$

$$= \frac{1}{N}\sum_{i=1}^{n} \left(\begin{bmatrix} u_1^T \\ u_m^T \end{bmatrix} (x_i - \mu) \right) \left(\begin{bmatrix} u_1^T \\ u_m^T \end{bmatrix} (x_i - \mu) \right)^T.$$

$$= \qquad (x_i - \mu)^T \begin{bmatrix} u_1 & u_2 & \ldots & u_m \end{bmatrix}$$

$$= \begin{bmatrix} u_1^T \\ u_m^T \end{bmatrix} \underbrace{\frac{1}{N}\sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T}_{\Sigma} \begin{bmatrix} u_1 & \ldots & u_m \end{bmatrix}$$

$$\hat{\Sigma} = \underset{m \times m}{\begin{bmatrix} u_1^T \\ \vdots \\ u_m^T \end{bmatrix}} \underline{\underset{m \times m}{\Sigma} \; \underset{m \times m}{\begin{bmatrix} u_1 & \cdots & u_m \end{bmatrix}}} \qquad \boxed{m \times n}$$

$$= \begin{bmatrix} u_1^T \\ \vdots \\ u_m^T \end{bmatrix} \begin{bmatrix} \lambda_1 u_1 & \lambda_2 \bar{u}_2 & \cdots & \lambda_m u_m \end{bmatrix}$$

$$= \begin{bmatrix} u_1^T \lambda_1 u_1 & u_1^T \lambda u_2 & \\ u_2^T \lambda_1 u_1 & & \\ & & \end{bmatrix} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_m \end{bmatrix}$$

Principle Components are Uncorrelated

- The covariance of the k'th and ℓ'th principle components of data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ is

$$\frac{1}{N}\sum_{i=1}^{N}\left[\boldsymbol{u}_k^T(\boldsymbol{x}_i-\boldsymbol{\mu})\right]\left[\boldsymbol{u}_\ell^T(\boldsymbol{x}_i-\boldsymbol{\mu})\right] = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{u}_k^T(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^T\boldsymbol{u}_\ell$$

$$= \boldsymbol{u}_k^T\boldsymbol{\Sigma}\boldsymbol{u}_\ell = \boldsymbol{u}_k^T\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{u}_\ell = \boldsymbol{e}_k^T\boldsymbol{\Lambda}\boldsymbol{e}_\ell = \begin{cases}\lambda_k & \text{if k} = \ell \\ 0 & \text{if k} \neq \ell\end{cases}$$

Therefore

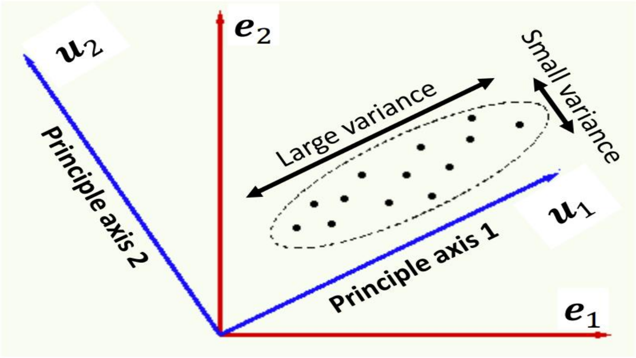
- ➢The variance of the k'th principle components is λ_k.

  ⇒ *principle axis 1 has the highest variance, axis 2 has the next highest variance, …, and axis M has the lowest variance.*

- ➢The covariance of different principle components is zero.
- ➢⇒ *Covariance among each pair of the principal axes is zero.*

Principle Axes

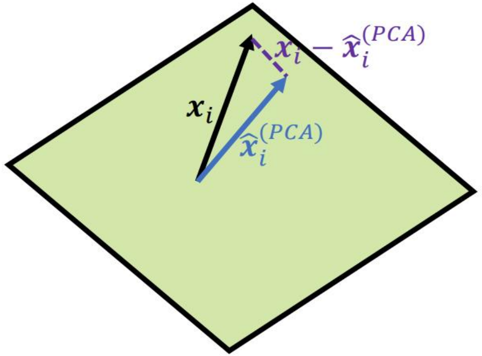
- Objective of PCA: Given data in $\mathbb{R}^M$, want to <u>*rigidly rotate*</u> the axes to new positions (principle axes) with the following properties:
  - *Ordered such that principle axis 1 has the highest variance, axis 2 has the next highest variance, ..., and axis M has the lowest variance.*
  - *Covariance among each pair of the principal axes is zero.*
- The k'th ***principle component*** is the projection to the k'th principle axis.
- Keep the first $m < M$ principle components for dimensionality reduction.



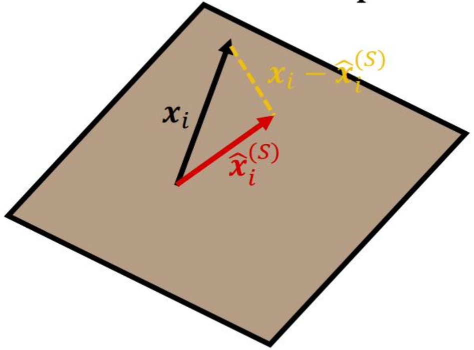
PCA and Reconstruction Error

WLOG assume zero mean $\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i = \boldsymbol{0}$

$S_{PCA} = Span(\boldsymbol{u}_1, \dots, \boldsymbol{u}_m)$



S: Arbitrary m-dimensional subspace



Variance after projection:

$$\sum_{i=1}^{N} \left\| \hat{\boldsymbol{x}}_i^{(PCA)} \right\|^2 \geq \sum_{i=1}^{N} \left\| \hat{\boldsymbol{x}}_i^{(S)} \right\|^2$$

Mean square error after projection:

$$\sum_{i=1}^{N} \left\| \boldsymbol{x}_i - \hat{\boldsymbol{x}}_i^{(PCA)} \right\|^2 \leq \sum_{i=1}^{N} \left\| \boldsymbol{x}_i - \hat{\boldsymbol{x}}_i^{(S)} \right\|^2$$

$$\hat{x}(PCA) = (u_1^T x) u_1 + (u_2^T x) u_2 + \dots + (u_m^T x) u_m$$

Low Rank Approximation

Eckart-Young-Mirsky Theorem:

Let $\boldsymbol{X} \in \mathbb{R}^{M\times N}$ be a matrix with singular value decomposition $\boldsymbol{X} = \boldsymbol{UDV}^T$, where $\boldsymbol{U} \in \mathbb{R}^{M\times M}, \boldsymbol{V} \in \mathbb{R}^{N\times N}$ are orthogonal matrices of left- and right-eigenvectors (of $\boldsymbol{X}$), and $\boldsymbol{D} \in \mathbb{R}^{M\times N}$ is a diagonal matrix of singular values $\sigma_i = D_{ii}$, arranged by their magnitude

$$|\sigma_1| \geq |\sigma_2| \geq \cdots \geq \left|\sigma_{\min(M,N)}\right|$$

Let $m \leq \min(\mathrm{M}, \mathrm{N})$, then both low rank approximation problems

$$\min_{\widehat{\boldsymbol{X}}} \left\|\boldsymbol{X} - \widehat{\boldsymbol{X}}\right\|_2 \qquad \text{subject to } rank(\widehat{\boldsymbol{X}}) \leq m$$

$$\min_{\widehat{\boldsymbol{X}}} \left\|\boldsymbol{X} - \widehat{\boldsymbol{X}}\right\|_F \qquad \text{subject to } rank(\widehat{\boldsymbol{X}}) \leq m$$

Has optimal solution $\widehat{\boldsymbol{X}} = \sum_{i=1}^{m} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$. Here $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ denotes the i’th column in matrices $\boldsymbol{U}, \boldsymbol{V}$, respectively.

Let $U = [u_1, u_2 \dots u_m]$

$V = [v_1, v_2 \dots v_N]$

For each $i \le \min\{m, N\}$, we have.

$$u_i^T X = (U e_i)^T X = e_i^T U^T U D V^T = e_i^T D V^T = \sigma_i e_i^T V^T = \sigma_i (V e_i)^T = \sigma_i v_i^T$$

$$X v_i = X V e_i = U D V^T V e_i = U D e_i = U \sigma_i e_i = \sigma_i u_i$$

We call u_i the left eigenvector of X

v_i the right eigenvector of X

WLOG assume zero mean $\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i = \boldsymbol{0}$

$$\boldsymbol{X} = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \dots \boldsymbol{x}_N] = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$$

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T = \frac{1}{N}\boldsymbol{X}\boldsymbol{X}^T = \frac{1}{N}\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{D}^T\boldsymbol{U}^T = \boldsymbol{U}\left(\frac{1}{N}\boldsymbol{D}\boldsymbol{D}^T\right)\boldsymbol{U}^T$$

$$\boldsymbol{\Lambda} = diag(\lambda_1, \dots, \lambda_M) = \frac{1}{N}\boldsymbol{D}\boldsymbol{D}^T = diag\left(\frac{\sigma_1^2}{N}, \dots, \frac{\sigma_{\min(M,N)}^2}{N}, 0, \dots, 0\right)$$

$|\sigma_1| \geq |\sigma_2| \geq \cdots$ implies $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M$

Projection by PCA: $\widehat{\boldsymbol{x}}_n^{(PCA)} = \sum_{i=1}^{m} \boldsymbol{u}_i \boldsymbol{u}_i^T \boldsymbol{x}_n$

$$\widehat{\boldsymbol{X}}^{(PCA)} = \left[\widehat{\boldsymbol{x}}_1^{(PCA)} \widehat{\boldsymbol{x}}_2^{(PCA)} \dots \widehat{\boldsymbol{x}}_N^{(PCA)}\right] = \sum_{i=1}^{m} \boldsymbol{u}_i\boldsymbol{u}_i^T\boldsymbol{X} = \sum_{i=1}^{m} \boldsymbol{u}_i\boldsymbol{u}_i^T\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T = \sum_{i=1}^{m} \sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^T$$

Projection to S: $\widehat{\boldsymbol{x}}_n^{(S)} \in \boldsymbol{S}$

$$\widehat{\boldsymbol{X}}^{(S)} = \left[\widehat{\boldsymbol{x}}_1^{(S)}\ \widehat{\boldsymbol{x}}_2^{(S)} \dots \widehat{\boldsymbol{x}}_N^{(S)}\right] \Rightarrow \boldsymbol{rank}(\widehat{\boldsymbol{X}}^{(S)}) \leq \boldsymbol{dim}(\boldsymbol{S}) = \boldsymbol{m}$$

Hence by Eckart-Young-Mirsky Theorem,

$$\left\|\boldsymbol{X} - \widehat{\boldsymbol{X}}^{(PCA)}\right\|_F \leq \left\|\boldsymbol{X} - \widehat{\boldsymbol{X}}^{(S)}\right\|_F, \quad \textit{for all m-dimensional subspace S}$$

That is,

$$\sum_{i=1}^{N}\left\|\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i^{(PCA)}\right\|^2 \leq \sum_{i=1}^{N}\left\|\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i^{(S)}\right\|^2, \quad \textit{for all m-dimensional subspace S}$$

Projection by PCA

$x \in \mathbb{R}^M \to z \in \mathbb{R}^m$

reconstruct x by z linearly.

say $\hat{x} = Wz$ $\quad W \in \mathbb{R}^{M \times m}$

$$\hat{x}_n^{(PCA)} = \sum_{i=1}^{m} u_i \underbrace{u_i^T x_n}_{z^{(n)}} \in \mathbb{R}^M$$

$$\hat{X}^{(PCA)} = [\hat{x}_1^{(PCA)} \cdots\cdots \hat{x}_N^{(PCA)}] \in \mathbb{R}^{M \times N}$$

$$= [\sum_{i=1}^{m} u_i u_i^T x_1 \; \sum_{i=1}^{m} u_i u_i^T x_2 \quad \sum_{i=1}^{m} u_i u_i^T x_m]$$

$$= \sum_{i=1}^{m} u_i u_i^T [x_1 \; x_2 \quad x_N]$$

$$= \sum_{i=1}^{m} u_i u_i^T X = \sum_{i=1}^{m} u_i \underbrace{\boxed{u_i^T U}}_{e_i^T} D V^T = \sum_{i=1}^{m} u_i \sigma_i e_i^T V^T = \sum_{i=1}^{m} \sigma_i u_i v_i^T$$

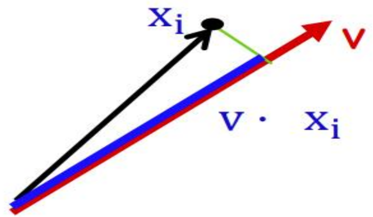
Two Interpretations

So far: Maximum Variance Subspace. PCA finds vectors v such that projections on to the vectors capture maximum variance in the data

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{v}^T\mathbf{x}_i)^2 = \mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v}$$

Alternative viewpoint: Minimum Reconstruction Error. PCA finds vectors v such that projection on to the vectors yields minimum MSE reconstruction

$$\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - (\mathbf{v}^T\mathbf{x}_i)\mathbf{v}\|^2$$



Why? Pythagorean Theorem

E.g., for the first component.

Maximum Variance Direction: 1st PC a vector v such that projection on to this vector capture maximum variance in the data (out of all possible one dimensional projections)

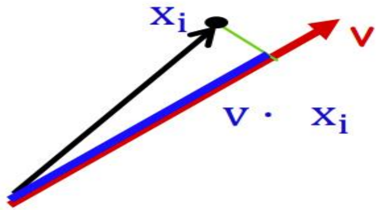
$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{v}^T\mathbf{x}_i)^2 = \mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v} \qquad \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - (\mathbf{v}^T\mathbf{x}_i)\mathbf{v}\|^2$$

Minimum Reconstruction Error: 1st PC a vector v such that projection on to this vector yields minimum MSE reconstruction

blue^2 + green^2 = black^2

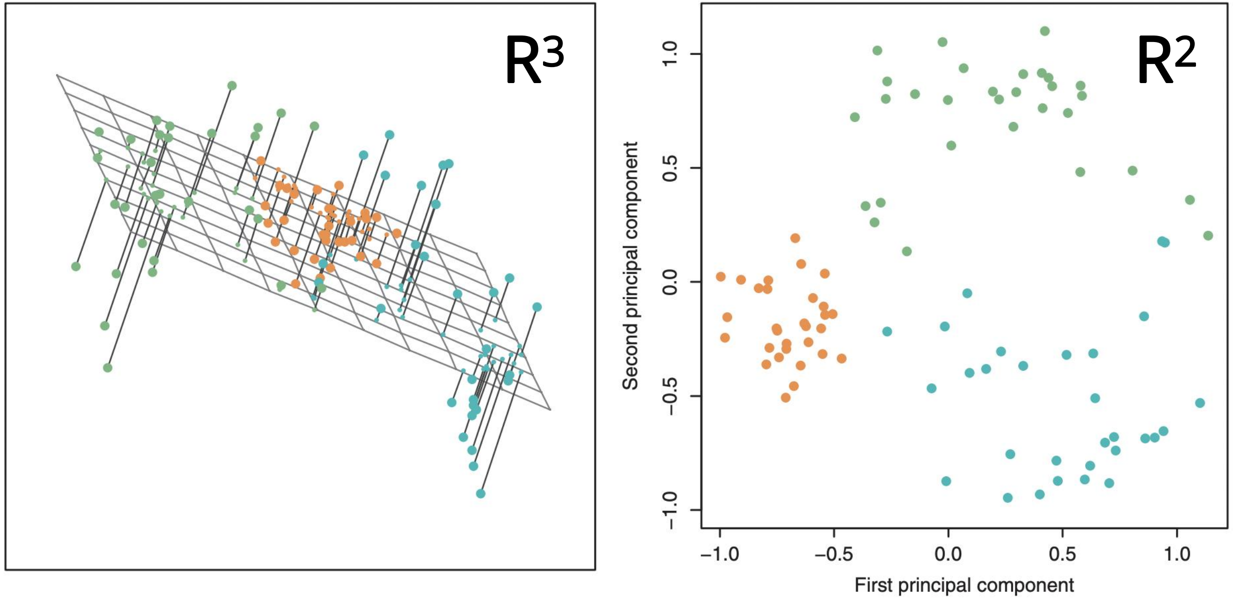
black^2 is fixed (it's just the data)

So, maximizing blue^2 is equivalent to minimizing green^2



Principal Component Analysis (PCA)

Goal: Given points in R^d, find the k directions of most variation, effectively performing dimensionality reduction.



Why is PCA used?

- **Lower latency inference** = embed high-dim points into a low-dim space, to be used for many downstream tasks
- **Reduce overfitting** = reducing irrelevant dimensions helps learning algorithms to generalize better
  - More powerful than just feature selection, because we're computing linear combinations of features!
- **Rich representation** = super high-dim points can be represented with a few dimensions, which helps to represent variation among complex things

Concept Check

1. What decomposition allows us to calculate the principal components?
    a. Eigendecomposition $\Sigma = V\Lambda V^T$.
2. Why are we always guaranteed a solution by eigendecomposition for any full-rank data matrix X?
    a. Because we perform eigendecomposition on the **covariance matrix**, which is $\Sigma = (1/n) X^TX$.
    b. This operation generates a square symmetric matrix, which is the criterion for performing an eigendecomposition.
3. Intuitively, why do we want to maximize sample variance in PCA?
    a. The goal of PCA is to find as information-rich of a representation as possible. When projecting down, we want to lose as little information as possible.
    b. When points are close together (low sample variance), we lose information.

Thanks for listening