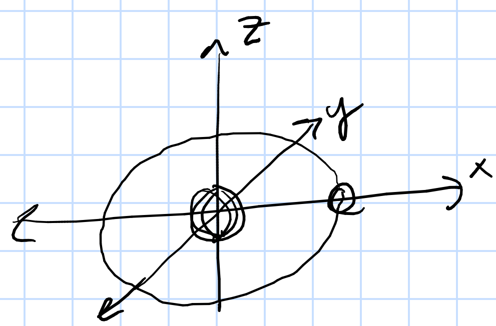
Principal Component Analysis

Data typically lie in a high dimensional space (ie one we cannot easily visualize). Further, some classifiers scale w/ the dimensionality (eg $\mathcal{O}(d^2 \log n)$). We want to look at way to reduce dimensionality

Why?

1. Reduce time complexity
2. Reduce space requirements
3. Occam's Razor
4. Data visualization



- HW # 2 : 02/17/2021
- Exam #1 : 02/26/2021 → 03/01/2021

- Re-opening?
  - Stage 2 course can meet on campus

02/22

Derivation of PCA

Let use frame the task of PCA as minimizing a distortion

$$J(x_0) = \sum_{i=1}^{n} \|x_0 - x_i\|_2^2$$

Solution

$$x_0 = \frac{1}{n} \sum_{i=1}^{n} x_i$$

A vector can be approximated

$$\boxed{x_k = m + \alpha_k w}$$

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$x_i \in \mathbb{R}^d$ $\quad \alpha \in \mathbb{R}$

Minimize $J(\alpha_1, \alpha_2, .., w)$ $\quad w = \in \mathbb{R}^d$

Assumption

$$\|w\|_2^2 = 1 \quad *$$

$$w^T(x_k - m) = \alpha_k w^T w$$

$$\boxed{\alpha_k = w^T(x_k - m)} \quad *$$

$$J(\alpha_1, \alpha_2, .., w) = \sum_{i=1}^{n} \|(m + \alpha_i w) - x_i\|_2^2$$

$$= \sum_{i=1}^{n} \|\alpha_i w - (x_i - m)\|_2^2$$

$$\|a - b\|_2^2 \in \mathbb{R}_+$$

$$= \sum_{i=1}^{n} \alpha_i^2 \underbrace{\|w\|_2^2}_{1} - 2\sum_{i=1}^{n} \alpha_i w^T(x_i - m) + \sum_{i=1}^{n} \|x_i - m\|_2^2$$

$\alpha_i = w^T(x_i - m)$

$$= \sum_{i=1}^{n} \left[w^T(x_i - m)\right]^2 - 2\sum_{i=1}^{n} \left[w^T(x_i - m)\right]^2 + \sum_{i=1}^{n} \|x_i - m\|_2^2$$

(1) (2)

$$= -\sum_{i=1}^{n} \left(w^T(x_i - m)\right)^2 + \sum_{i=1}^{n} \|x_i - m\|_2^2$$

$$\propto -\sum_{i=1}^{n} w^T(x_i - m)(x_i - m)^T w$$

$S \to$ scatter matrix

$$= -w^T S w$$

s.t. $\|w\|_2^2$