Lecture #2: 9-6-18

- Suppose we have some data $X_1, \dots, X_n \subset \mathbb{R}^D$

- When D is large, hard to visualize and do statistics (curse of dimensionality).
- We can try to reduce the dimension of the data while retaining important properties of the data.
- Let $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^{n \times D}$ (n: # data points, D: # dimensions)
- Suppose U is a $D \times D$ matrix whose columns are orthogonal: $U = \begin{pmatrix} | & & | \\ u_1 & \cdots & u_D \\ | & & | \end{pmatrix}$

s.t. $u_i \cdot u_j = u_i^T u_j = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}$

($u_i \cdot u_j$: dot product; $u_i^T u_j$: matrix product)

Then we can write any $x \in \mathbb{R}^D$ as $x = \sum_{i=1}^{D} \alpha_i u_i$. We can even compute α_i via orthogonality:

$$x = \sum_{i=1}^{D} \alpha_i u_i$$

$$\Rightarrow x u_j^T = \alpha_j, \quad \forall j = 1, \dots, D$$

$$\Rightarrow \alpha = (\alpha_1, \dots, \alpha_D) \text{ may be computed as } \alpha = U^T x.$$

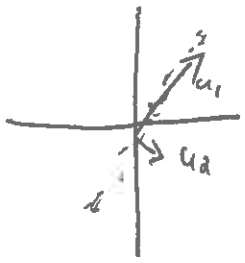
- This is valid for an orthogonal matrix U, or equivalently, for any orthonormal basis for $\mathbb{R}^D$ $u_1, \dots, u_D$.

Q: Which is "best"?

A: Depends on how you define "best"

- One method — principal component analysis — seeks a basis $u_1, \dots, u_D$ s.t.
  1) The projection of $X \in \mathbb{R}^{n \times D}$ onto u_1 is variance maximizing over all projections onto 1-dimensional spaces.
  2) The projection of X onto $\text{span}\{u_1, u_2\}$ is variance maximizing over all projections onto 2-dimensional spaces.
  etc.

ex:



$u_1 \perp u_2$ is required

Let's make this rigorous by recalling the definition of variance: for data $x_1, \dots, x_n \in \mathbb{R}$, the variance of the data is $\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$, where $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the mean.

- Let $u_* \in \mathbb{R}^{1 \times D}$ be the maximizer of

$$\frac{1}{n}\sum_{i=1}^{n}\left(u^T x_i - u^T \mu\right)^2$$

, where $x_i \in \mathbb{R}^{D \times 1}$, $\mu \in \mathbb{R}^{D \times 1}$ is the mean of $\{x_1, \dots, x_n\}$

- Suppose, without loss of generality, that $\mu = 0$, so

$$u_* = \arg\max_u \frac{1}{n}\sum_{i=1}^{n}\left(u^T x_i\right)^2.$$

- Let's analyze: $\left(u^T x_i\right)^2 = \left(u^T x_i\right) \cdot \left(u^T x_i\right)^T$, since these are just numbers

$= u^T x_i \cdot x_i^T \cdot u$, since $(AB)^T = B^T A^T$ and $(A^T)^T = A$

Then, $u_* = \arg\max_u \frac{1}{n}\sum_{i=1}^{n} u^T x_i x_i^T u$

$= \arg\max_u u^T \left[\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\right] u$ $\quad \in \mathbb{R}^{D\times D}$

$= \arg\max_u u^T \Sigma u$, where $\Sigma = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$ is the (empirical) covariance matrix of $x_1, \dots, x_n$.

Remark: $\Sigma = \frac{1}{n} X^T X$ (big!), where $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^{n\times D}$

- So, how to solve $\arg\max_u u^T \Sigma u$? Derivative! Indeed, the function $F: \mathbb{R}^D \to \mathbb{R}$, $u \mapsto u^T \Sigma u$ is smooth, so we can differentiate it; however, u is constrained since $u^T u = 1$.

? : How to constrained optimization? Lagrange multiplier:

$$\arg\max_u \underbrace{u^T \Sigma u}_{\text{to optimize}} - \underbrace{\lambda(u^T u - 1)}_{\text{constraint}}$$

- Differentiate and set equal to 0 vector: $\frac{\partial}{\partial u}\left[u^T \Sigma u - \lambda(u^T u - 1)\right] = 0$

$\Leftrightarrow$ (exercise) $\quad 2\Sigma u - 2\lambda u = 0$

$\Leftrightarrow \Sigma u = \lambda u$

$\Rightarrow$ 1) u is an eigenvector of Σ

2) λ is an eigenvalue of Σ

- Conclusion: the potential maximizers are the eigenvectors of Σ. Which is best? Well, $\Sigma u = \lambda u$

$$\Rightarrow u^T \Sigma u = u^T \lambda u$$

$$= \lambda \underbrace{u^T u}_{=1}$$

$$= \lambda$$

So, we should choose (u, λ) the eigenpair corresponding to the largest eigenvalue.

- In general, the principal components of mean-centered data $x_1, \dots, x_n \in \mathbb{R}^{D \times 1}$, i.e. $\frac{1}{n}\sum_{i=1}^{n} x_i = 0 \in \mathbb{R}^D$ are the eigenvectors of $\Sigma = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$

$$= \frac{1}{n} X^T X, \text{ where } X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^{n \times D}$$

- Call them $u_1, \dots, u_D$. - u_1 points in the direction of maximum variance of $x_1, \dots, x_n$

- u_2 points in the "direction" among those directions orthogonal to u_1

- u_3 " among those directions orthogonal to u_1 and u_2

- etc.

- The larger the eigenvalue associated to principal component a_i, the more significant the direction.

- Next time: - geometric considerations
  - computational complexity
  - numerical examples