Lecture 8 :

Singular Value Decomposition

The Singular Value Decomposition

Consider a matrix $X \in \mathbb{R}^{n \times p}$. There always exists matrices U, Σ, V such that

$$X = U \Sigma V^T$$

$U \in \mathbb{R}^{n \times n}$ is orthogonal ($U^TU = UU^T = I$), called left singular vectors – basis for cols of X
$V \in \mathbb{R}^{p \times p}$ is orthogonal ($V^TV = VV^T = I$), called right singular vectors
$\Sigma \in \mathbb{R}^{n \times p}$ is diagonal; diagonal elements called singular values – basis for rows of X

$\underline{n = p}$

$$\Sigma = \begin{bmatrix} \sigma_1 & & & 0 \\ & \sigma_2 & & \\ & & \ddots & \\ 0 & & & \sigma_n \end{bmatrix}$$

$n > p$

$$\begin{bmatrix} \begin{matrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \end{matrix} \\ \hline 0 \end{bmatrix} \begin{matrix} \Big\} \ p \times p \\ \Big\} \ (n-p) \times p \end{matrix}$$

$n < p$

$$\left[\begin{array}{c|c} \underbrace{\begin{matrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{matrix}}_{n \times n} & \underbrace{0}_{n \times (p-n)} \end{array} \right]$$

Let $r = \min(n, p)$. Then $X = \sum_{i=1}^{r} u_i \sigma_i v_i^T$, $u_i = i^{th}$ col of U, $v_i = i^{th}$ column of V

= sum of rank-1 matrices

If X is square and has $\sigma_j = 0$ for any j, then X is not invertible, or "singular"

If X is square and not singular, then $X^{-1} = V \Sigma^{-1} U^T$

The SVD of $X^T = (U \Sigma V^T)^T = V \Sigma^T U^T$

$\Rightarrow$ columns of U are basis for the <u>columns</u> of X

and columns of V are basis for the <u>rows</u> of X

a note on multiplying by diagonal matrices:

$$\begin{bmatrix} | & | & | \\ A_1 & A_2 & A_3 \\ | & | & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \sigma_2 & \\ & & \sigma_3 \end{bmatrix} = \begin{bmatrix} | & | & | \\ \sigma_1 A_1 & \sigma_2 A_2 & \sigma_3 A_3 \\ | & | & | \end{bmatrix}$$

diagonal matrix on right $\Rightarrow$ reweigh columns

$$\begin{bmatrix} \sigma_1 & & \\ & \sigma_2 & \\ & & \sigma_3 \end{bmatrix} \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ - & a_3^T & - \end{bmatrix} = \begin{bmatrix} - & \sigma_1 a_1^T & - \\ - & \sigma_2 a_2^T & - \\ - & \sigma_3 a_3^T & - \end{bmatrix}$$

diagonal matrix on left $\Rightarrow$ reweigh rows

- U gives orthobasis for all of $\mathbb{R}^p$.
- 1st r columns of U give basis of best r-dim subspace fit to columns of X
- 1st r columns of V give coordinates / locations of each $\underline{x}_i$ within the subspace spanned by $\underline{U}_1, \dots, \underline{U}_r$
- σ_i's indicate how important each subspace dimension is to representing / approximating data

recall $\sigma_1 = \|X^T U_1\|_2 = \left[\sum_i (\underline{U}_1^T \underline{x}_i)^2\right]^{1/2}$ where $X = \begin{bmatrix} - x_1^T - \\ - x_2^T - \\ \vdots \\ - x_p^T - \end{bmatrix}$

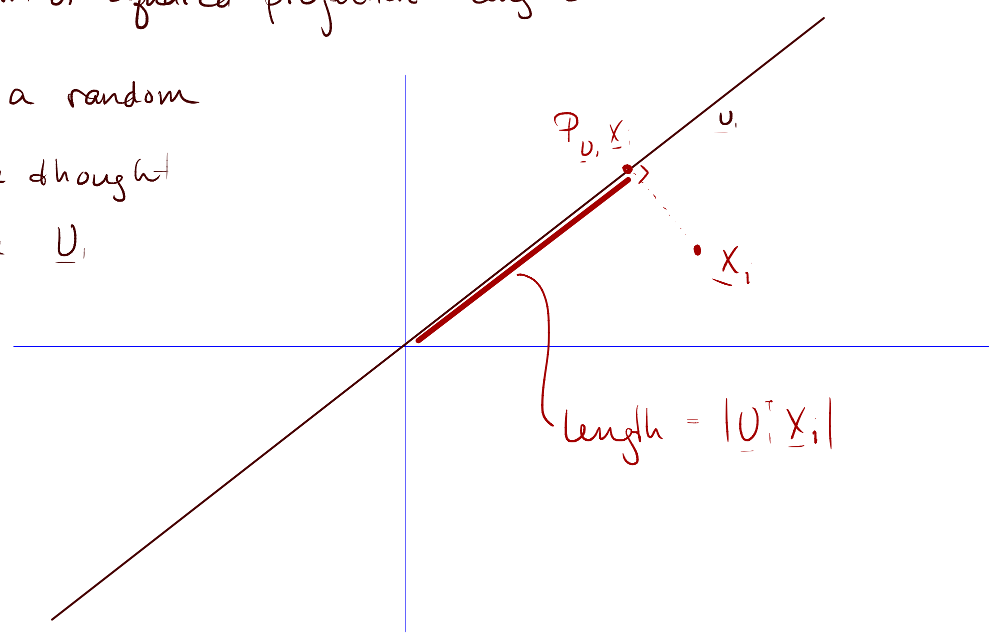
also recall $P_{\underline{U}_1} \underline{x}_i = \underline{U}_1 \underline{U}_1^T \underline{x}_i$ and that,

since $\underline{U}_1$ is an orthonormal basis, it is length-preserving: $\|U \underline{v}\|_2 = \|\underline{v}\|_2$ for any $\underline{v}$.

This means the length[2] of the projection of $\underline{x}_1$ onto $\underline{U}_1$ is $\|U_1 U_1^T x_1\|_2^2 = (U_1^T x_1)^2$

In other words, $\sigma_1^2 = \sum_i (U_1^T x_1)^2$ is the sum of squared projection lengths

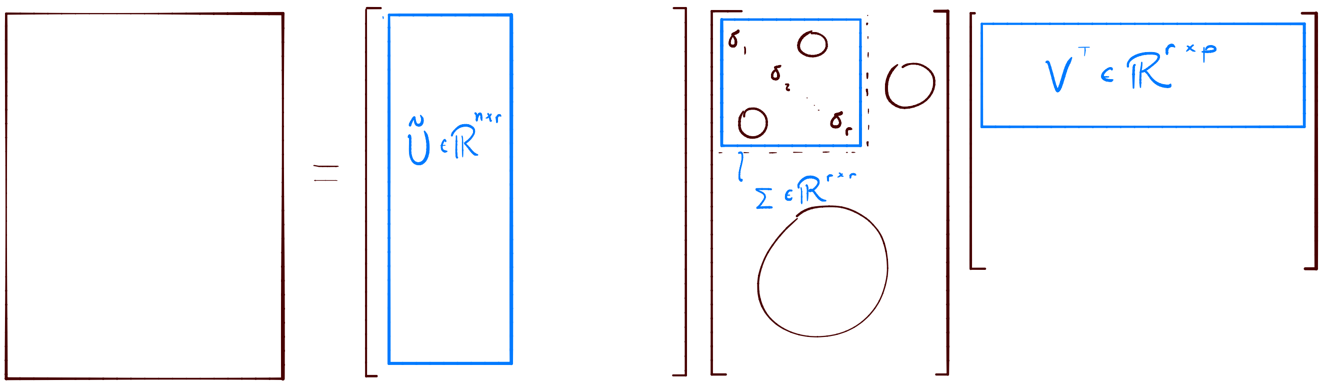
If we think of the x_1's as realizations of a random variable with mean $= 0$, then σ_1^2 can also be thought of as the variance of the $\underline{x}_i$'s in the $\underline{U}_1$ direction.



The "Economy SVD"

Let $X \in \mathbb{R}^{n \times p}$, with rank $r \ll \min(n,p)$

Then $X = U \in \mathbb{R}^{n \times n} \quad \Sigma \in \mathbb{R}^{n \times p} \quad V^T \in \mathbb{R}^{p \times p}$



$$\boxed{X = U \Sigma V^T = \tilde{U} \tilde{\Sigma} \tilde{V}^T}$$

this is the economy SVD

Netflix example

$n \approx 5k$ movies
$p \approx 100m$ customers

$\Longrightarrow$ storage $= 5k \cdot 100m \cdot 4$ bytes ≈ 2 TB $\approx$ okay to store, difficult to use in learning algorithms.

Now let's say we find a rank-r approximation to X using the subspace approximation theorem, and use its economy SVD. if $r = 10$,

$\tilde{U}$ takes $5k \cdot 10 \cdot 4$ bytes $= 200$ kB, $\tilde{V}$ takes $100m \cdot 10 \cdot 4$ bytes $= 4$ GB, $\tilde{\Sigma}$ takes $10 \cdot 4$ bytes

MUCH SMALLER!

Subspace Approximation Theorem

If $X \in \mathbb{R}^{p\times n}$ has rank $r > k$, then

$$\underset{Z:\, \text{rank}(Z)=k}{\text{argmin}} \; \| X - Z \|_F^2 = \underbrace{U_k}_{\substack{1^{st}\ k \text{ cols} \\ \text{of } U}} \underbrace{\Sigma_k}_{\substack{1^{st}\ k\times k \\ \text{block of} \\ \Sigma}} \underbrace{V_k^T}_{\substack{1^{st}\ k \\ \text{cols of} \\ V}}$$

and $\| X - X_k \|_F^2 = \sum_{i=k+1}^{r} \sigma_i^2$

Frobenius matrix norm

$$\| A \|_F = \left(\sum_{i,j} A_{ij}^2 \right)^{1/2}$$

if $A = [A_1 \; A_2 \; \cdots \; A_p]$,

then $\| A \|_F^2 = \sum_{i=1}^{p} \| A_i \|_2^2$

Ex.
$$\Sigma = \begin{bmatrix} 100 & & & & & \\ & 90 & & & 0 & \\ & & 80 & & & \\ & & & 3 & & \\ & 0 & & & 2 & \\ & & & & & 1 \end{bmatrix} \approx \begin{bmatrix} 100 & & & & & \\ & 90 & & & 0 & \\ & & 80 & & & \\ & & & 0 & & \\ & 0 & & & 0 & \\ & & & & & 0 \end{bmatrix}$$

$\Rightarrow$ X is "almost" low rank.

error of rank-3 approximation is $3^2 + 2^2 + 1^2 = 14$

Principal Components Analysis

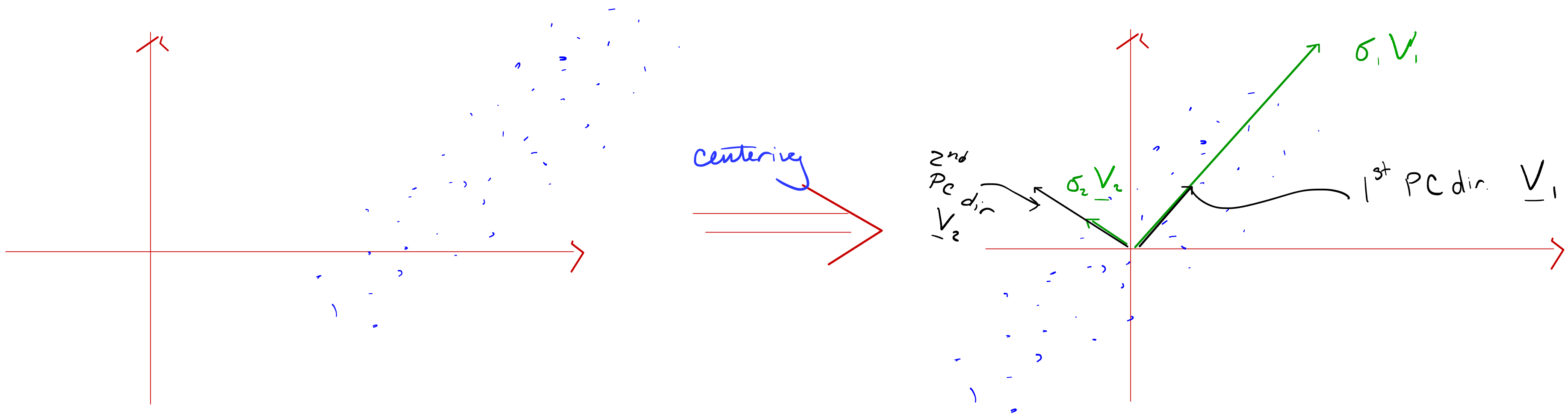
Let $X \in \mathbb{R}^{n \times p}$ be a data matrix with rows "centered" to have average value of 0.

(n points in p dimensions)

1st "center" data — if $X = \begin{bmatrix} | & | & & | \\ X_1 & X_2 & \cdots & X_p \\ | & | & & | \end{bmatrix}$, for $i = 1, 2, \ldots, p$, $\underline{X}_i \leftarrow X_i - \text{mean}(X_i)$

$$= X_i - \left(\sum_{j=1}^{n} \frac{X_{ij}}{n} \right) \mathbb{1}_{n \times 1}$$

$X = U \Sigma V^T \Rightarrow V$ is basis matrix for $\mathbb{R}^p$



If $X = U \Sigma V^T$, then right singular vectors of X are called the Principal Component Directions

Equivalently let $C = X^T X = (U \Sigma V^T)^T (U \Sigma V^T) = V \Sigma U^T U \Sigma V^T = \underbrace{V \Sigma^2 V^T}_{\text{called "eigenvalue decomposition"}}$

right singular vectors of X = eigenvectors of $X^T X$

Dimensionality Reduction

We have n points $\underline{x}_i \in \mathbb{R}^p$, $i=1,\dots,n$. Dimensionality reduction means defining new points $\underline{z}_i \in \mathbb{R}^k$, $i=1,\dots,n$ for $k<p$ that preserve important properties of the $\underline{x}_i$'s.

(e.g. $\|\underline{x}_i - \underline{x}_j\| \approx \|\underline{z}_i - \underline{z}_j\|$)

Let $X = \begin{bmatrix} — \underline{x}_1^T — \\ — \underline{x}_2^T — \\ \vdots \\ — \underline{x}_n^T — \end{bmatrix} = U\Sigma V^T \in \mathbb{R}^{n\times p} \quad \Rightarrow \quad X^T = \begin{bmatrix} | & | & & | \\ \underline{x}_1 & \underline{x}_2 & \cdots & \underline{x}_n \\ | & | & & | \end{bmatrix} = V\Sigma U^T \in \mathbb{R}^{p\times n}$

Consider $X_k^T = V_k \Sigma_k U_k^T$; let i^{th} column of X_k^T be $\tilde{\underline{x}}_i$

by subspace approximation theorem, we know $\|\underline{x}_i - \tilde{\underline{x}}_i\|_2^2 \leq \sum_{j=k+1}^{r} \sigma_j^2$ — if σ_j's small for $j>k$, then $\underline{x}_i$ is close to $\tilde{\underline{x}}_i$!

but $\tilde{\underline{x}}_i$'s are all in a k-dimensional subspace, so we can represent them in terms of their coordinates in the subspace !

$$\Rightarrow \underline{z}_i = V_k^T \tilde{\underline{x}}_i = i^{th} \text{ col of } (\Sigma_k U_k^T) \in \mathbb{R}^k$$

Ex: $\underline{x}_i \in S = \{\underline{x} \in \mathbb{R}^3 : x_3 = 0\}$ = horizontal plane

instead of representing each $\underline{x}_i$ using 3d coordinates, we only need to represent where it is in the plane

Example: $X = \begin{bmatrix} 1 & -1 & -1 & 1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & 0 & 0 \end{bmatrix} \Longrightarrow \underline{x}_i = (i^{th} \text{ row of } X)^T = i^{th} \text{ column of } X^T$, for $i = 1, \dots, 5$

$\hookrightarrow$ 4-dimensional

(e.g. $\underline{x}_1 = \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix}$)

$$= \begin{bmatrix} 1/\sqrt{5} & 1/2 \\ -1/\sqrt{5} & 1/2 \\ 1/\sqrt{5} & 1/2 \\ -1/\sqrt{5} & 1/2 \\ 1/\sqrt{5} & 0 \end{bmatrix} \underbrace{\begin{bmatrix} \sqrt{5}\cdot\sqrt{2} & 0 \\ 0 & 2\sqrt{2} \end{bmatrix}}_{\begin{bmatrix} \sqrt{5} & 0 \\ 0 & 2 \end{bmatrix} \cdot \begin{bmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{2} \end{bmatrix}} \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 & 0 \\ 0 & 0 & -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

economy SVD

$$= \begin{bmatrix} 1 & 1 \\ -1 & 1 \\ 1 & 1 \\ -1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

$$\Rightarrow X^T = \underbrace{\left(\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & -1 \\ 0 & 1 \end{bmatrix} \right)}_{\tilde{V}_k} \underbrace{\left(\begin{bmatrix} 1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix} \cdot \sqrt{2} \right)}_{\tilde{\Sigma}_k \tilde{U}_k^T} \approx> \underline{x}_i = \text{weighted sum of } \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix} \text{ \& } \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 0 \\ -1 \\ 1 \end{bmatrix}$$

$$\underline{z}_i = V_k^T \tilde{\underline{x}}_i \Rightarrow \underline{z}_1 = V_k^T \underline{x}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 2 \\ 2 \end{bmatrix} = \begin{bmatrix} \sqrt{2} \\ \sqrt{2} \end{bmatrix}$$

let $\underline{z}_1 = \sqrt{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $\underline{z}_2 = \sqrt{2} \begin{bmatrix} -1 \\ 1 \end{bmatrix}$, etc. each $\underline{z}_i \in \mathbb{R}^2$ $\Rightarrow$ reduced-dimension representation of $\underline{x}_i$'s

$$X = \begin{bmatrix} 1.1 & 2.0 & 3.4 & 4.05 \\ 2.01 & 4.2 & 6.1 & 8.05 \\ 3.2 & 6.0 & 9.05 & 12 \\ 4 & 8.1 & 12 & 16 \\ 5 & 10 & 15 & 20 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix}}_{\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 6 & 8 \\ 3 & 6 & 9 & 12 \\ 4 & 8 & 12 & 16 \\ 5 & 10 & 15 & 20 \end{bmatrix}} + \text{"errors"} \approx \underbrace{\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix}}_{=: \tilde{X}}$$

so $\tilde{X} \approx X$

can find $\tilde{X}$ by thresholding small singular values.

$\tilde{X}$ is the same size as X $\Rightarrow$ no dimensionality reduction has occured yet

(Let's think of $X \in \mathbb{R}^{n \times p}$ – each row is a point x_i^T, $x_i \in \mathbb{R}^p$, $i = 1, \dots, n$

to perform dim red., want n new points $z_i \in \mathbb{R}^r$ for $r < p$, $i = 1, \dots, n$

To do this, just use $k = 1$, $V_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} \dfrac{1}{\sqrt{55}}$ $\Rightarrow$ $z_1 = V_1^T x_1 = 7.51 \approx \sqrt{55}$

$z_2 = V_1^T x_2 = 14.94 \approx 2\sqrt{55}$

More generally:

Step 1: find $\tilde{X}$ $\Leftarrow$ see how entries of X are related / tied together

Step 2: exploit those ties to get simple representation of each x_i — set small singular values to zero, let $z_i = V_k^T x_i$ for $i = 1, \dots, n$.