Lecture 7

Introduction to Singular Value Decomposition

SWCON253, Machine Learning

Won Hee Lee, PhD

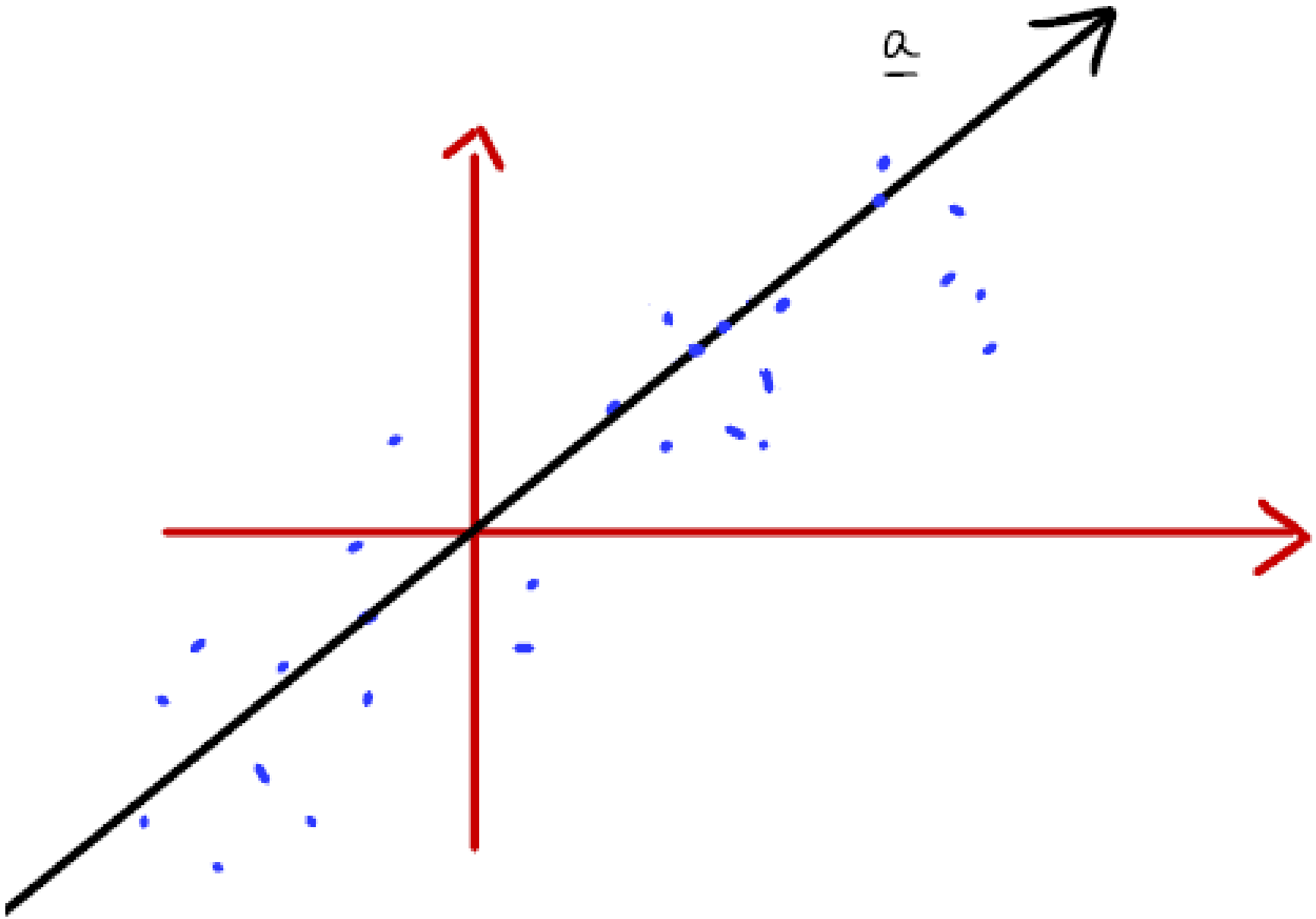
Learning Goals

- Understand the fundamental concepts of **singular value decomposition (SVD)** in machine learning

Announcements

- Get **bonus** points
  - E-campus Machine Learning - Q&A 문의게시판
  - Homework assignments (35%) + **alpha (bonus)**

- Lecture notes vs. Lecture slides


a

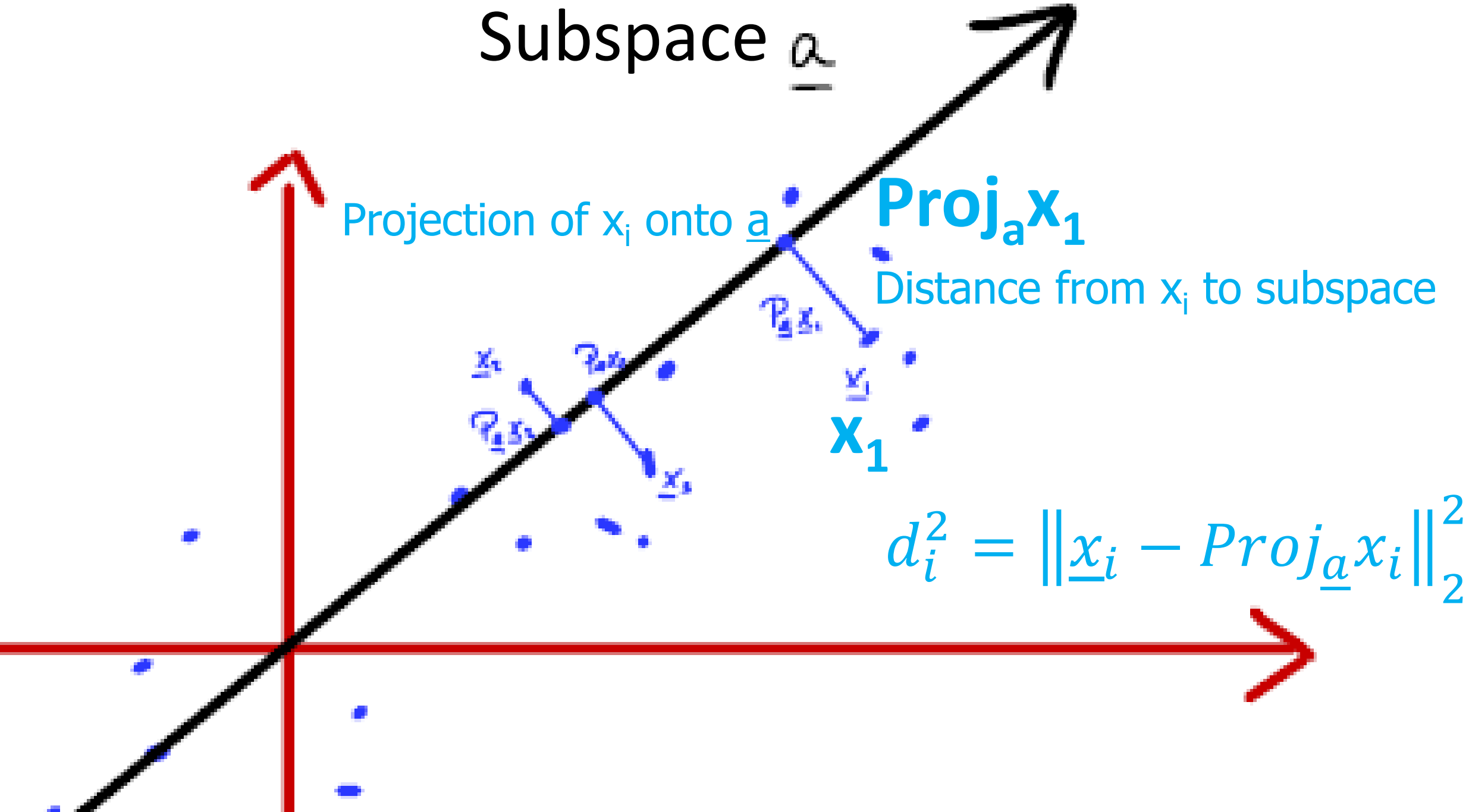
We observe $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_p \in \mathbb{R}^n$, equivalently, $X = [\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_p]$

Goal

Find the 1D subspace that is closest to a set of points ("best" fits data)

i.e., distance from x_i to subspace to be as small as possible

Subspace $\underline{a}$



Observe $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_p \in \mathbb{R}^n$ $X = [\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_p]$

Goal

find the 1D subspace that is closest to a set of points ("best" fits data)

i.e., distance from x_i to subspace to be as small as possible

$$d_i^2 = \left\| \underline{x}_i - Proj_{\underline{a}} x_i \right\|_2^2$$

We want to minimize $\sum_{i=1}^{p} d_i^2$

Projection Matrices

If $A \in \mathbb{R}^{nxp}$ spans a subspace, then projection of $\underline{X}$ onto span(cols(A)) = $Proj_A \underline{X}$

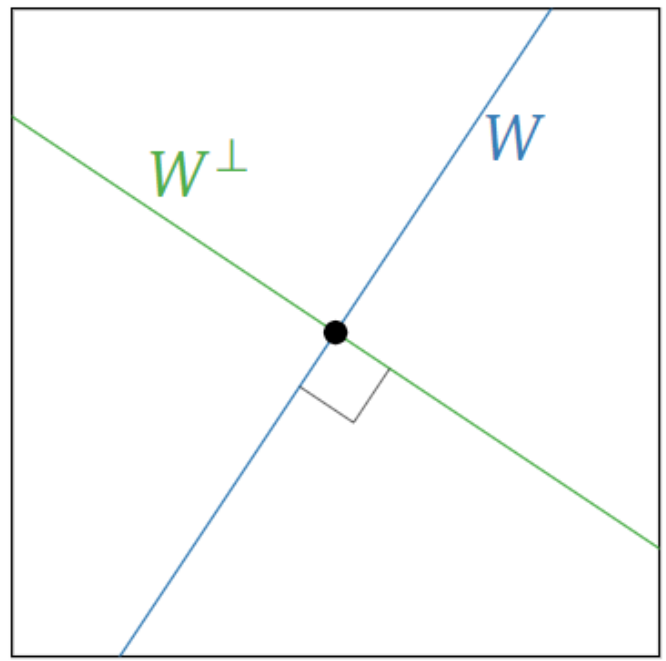
If columns of A are linearly independent,

$Proj_A \underline{X}$ = $A(A^TA)^{-1}A^TX$ (Projection matrix P_A from lecture 6)

Properties of projection matrices

- $\boldsymbol{P_A = P_A^2 = P_A P_A}$
- $\boldsymbol{P_A = P_A^T = P_A^T P_A}$ (Projection matrix is always symmetric)
- If A = $\underline{a}$ (single column vector), then $P_{\underline{a}} = a(a^Ta)^{-1}a^T = \frac{aa^T}{a^Ta}$ (as a^Ta is scaler)
- The orthogonal complement of a subspace is the orthogonal to the subspace.

Pictures of orthogonal complements. The orthogonal complement of a line W through the origin in $\mathbf{R}^2$ is the perpendicular line $W^\perp$.



Let B be a basis for orthogonal complement.

$A^T B = 0$

For any vector $\underline{x} \in \mathbb{R}^n$, it can be written as

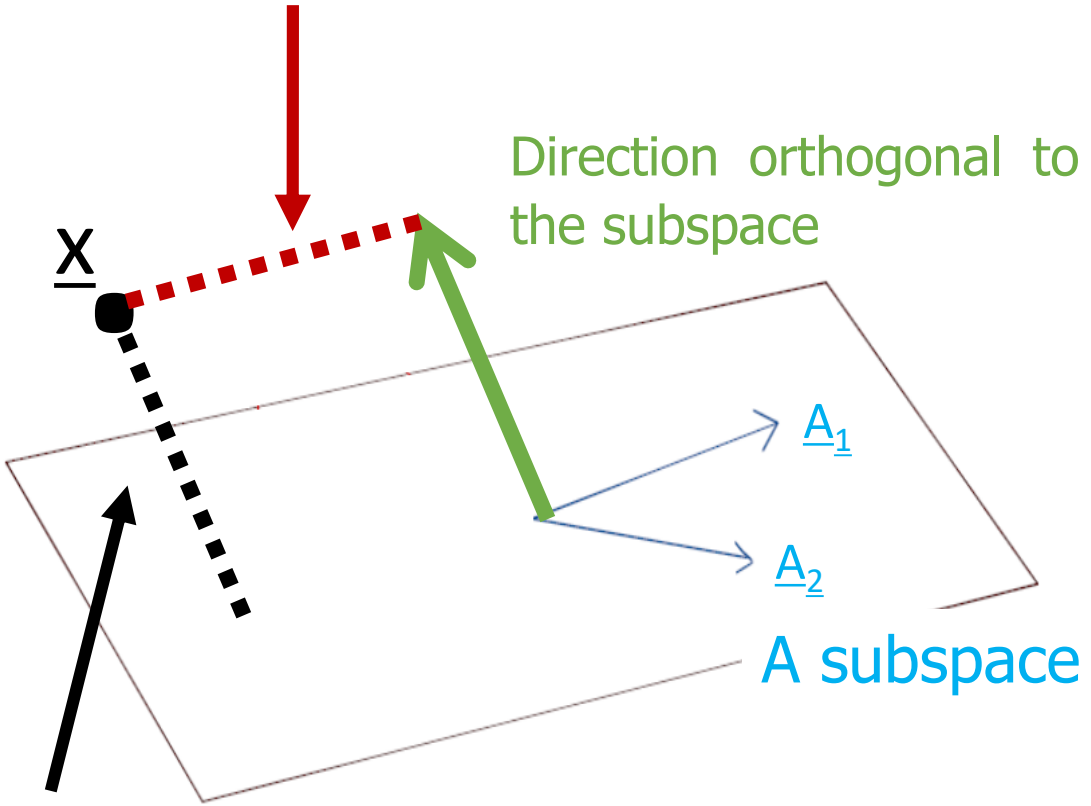
$$\underline{x} = P_A\underline{x} + P_B\underline{x} = (P_A + P_B)\underline{x}$$

$$I\underline{x} = P_A\underline{x} + P_B\underline{x} = (P_A + P_B)\underline{x}$$

$$\rightarrow \mathrm{I} = P_A + P_B$$

$$\rightarrow \boldsymbol{P_B = \mathrm{I} - P_A}$$

For new point $\underline{x}$ whose projection is relative to the orthogonal subspace



For new point $\underline{x}$ that is not in the subspace,
Projection of $\underline{x}$ is in the subspace

$$d_i^2 = \left\| \underline{x}_i - P_{\underline{a}} \underline{x}_i \right\|_2^2$$

$$= \left\| \underline{x}_i - \frac{aa^T}{a^T a} \underline{x}_i \right\|_2^2$$

$$= \left\| (I - \frac{aa^T}{a^T a}) \underline{x}_i \right\|_2^2$$

$$= x_i^T \left(I - \frac{aa^T}{a^T a} \right)^T (I - \frac{aa^T}{a^T a}) \underline{x}_i$$

$$= x_i^T (I - \frac{aa^T}{a^T a}) \underline{x}_i$$

$$= \underline{x}_i^T \underline{x}_i - \frac{\underline{x}_i^T \underline{a} \underline{a}^T \underline{x}_i}{\underline{a}^T \underline{a}}$$

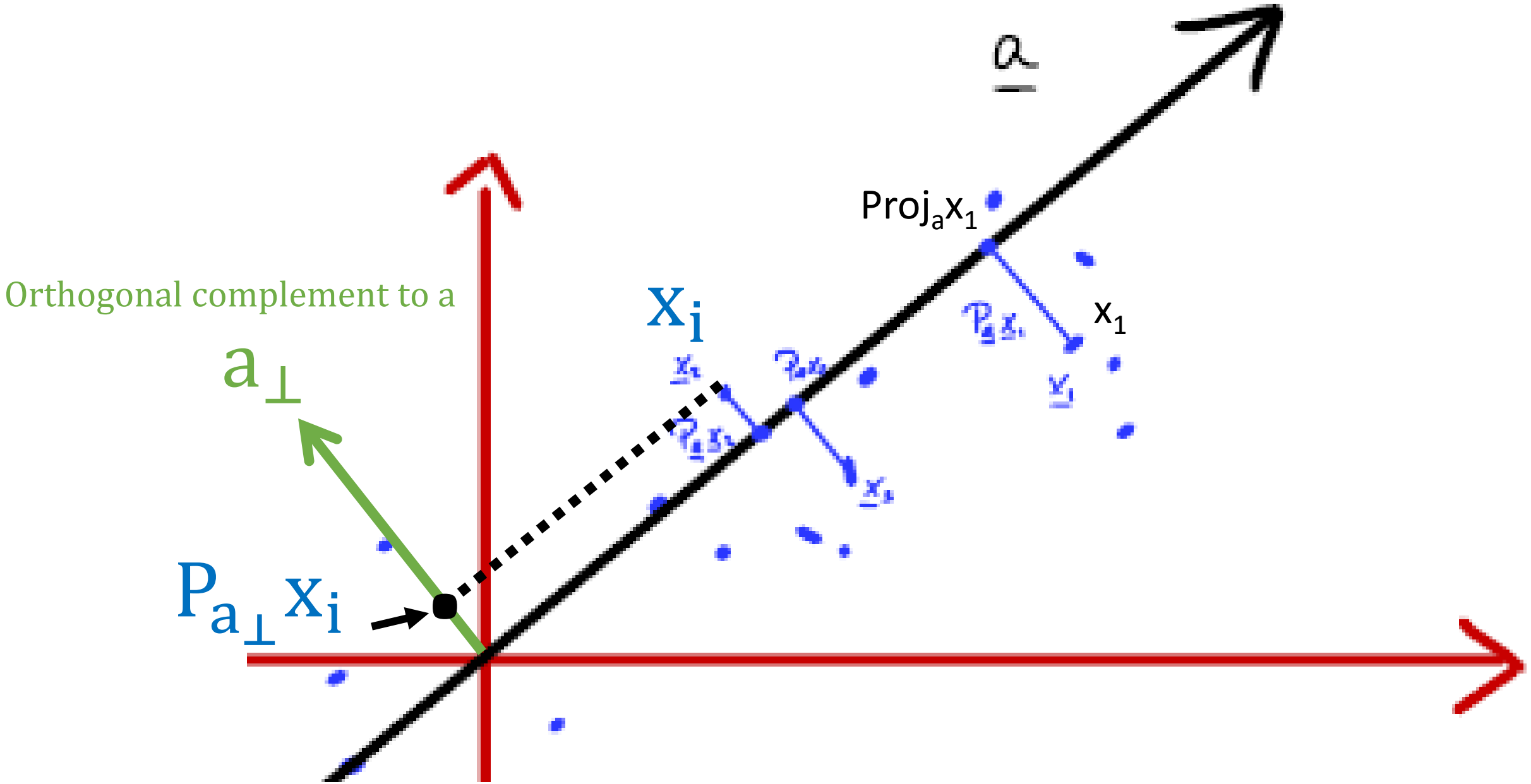
$$\mathrm{P_a} = \frac{aa^T}{a^T a}$$

Projection on $\mathrm{a}_\perp$ (orthogonal complement)

$\mathrm{a}_\perp$ "a perp"

projection matrix x projection matrix = projection matrix

$$\boldsymbol{P_A = P_A^2 = P_A P_A = P_A^T = P_A^T P_A}$$



the distance of that point to the origin = the distance of x_i to the subspace

We want to minimize

$$\sum_{i=1}^{p} d_i^2 = \sum_{i=1}^{p} \underline{x}_i^T \underline{x}_i - \frac{\underline{x}_i^T \underline{a}\underline{a}^T \underline{x}_i}{\underline{a}^T \underline{a}}$$

$\underline{x}_i^T \underline{x}_i$ does not depend on a, i.e., constant with respect to a

$$\hat{\underline{a}} = \underset{\underline{a}}{\operatorname{argmin}} \sum_{i=1}^{p} d_i^2$$

$$\hat{\underline{a}} = \underset{\underline{a}}{\operatorname{argmin}} \sum_{i=1}^{p} -\frac{\underline{x}_i^T \underline{a}\, \underline{a}^T \underline{x}_i}{\underline{a}^T \underline{a}}$$

(Minimum of negative is the maximum of non-negative)

$$\hat{\underline{a}} = \underset{\underline{a}}{\operatorname{argmax}} \sum_{i=1}^{p} \frac{\underline{x}_i^T \underline{a}\, \underline{a}^T \underline{x}_i}{\underline{a}^T \underline{a}}$$

($\underline{x}_i^T \underline{a}$, $\underline{a}^T \underline{x}_i$ are scalers) ($\underline{x}_i^T \underline{a} = \underline{a}^T \underline{x}_i$)

$$\hat{\underline{a}} = \underset{\underline{a}}{\operatorname{argmax}} \sum_{i=1}^{p} \frac{\underline{a}^T \underline{x}_i \underline{x}_i^T \underline{a}}{\underline{a}^T \underline{a}}$$

$$\hat{\underline{a}} = \underset{\underline{\boldsymbol{a}}}{\mathbf{argmax}} \frac{\underline{\boldsymbol{a}}^{\boldsymbol{T}} \boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}} \underline{\boldsymbol{a}}}{\underline{\boldsymbol{a}}^{\boldsymbol{T}} \underline{\boldsymbol{a}}}$$

$$\underline{\hat{a}} = \underset{\underline{a}}{\mathbf{argmax}} \frac{\underline{a}^T \boldsymbol{X}\boldsymbol{X}^T \underline{a}}{\underline{a}^T \underline{a}}$$

- The vector $\underline{\hat{a}}$ that achieves the maximum is called the 1st left singular vector of X.
- The value of $\frac{\underline{\hat{a}} x x^T \underline{\hat{a}}}{\underline{\hat{a}}^T \underline{\hat{a}}} = \sigma_i^2$ is the squared 1st singular value of X.

The Singular Value Decomposition (SVD)

Consider a matrix $X \in \mathbb{R}^{nxp}$. There exist matrices U, Σ, V such that

$X = U\Sigma V^T$

- The columns of U are called the "left singular vectors".
- U is an orthogonal matrix ($U^TU = UU^T = I$).
- The columns of U give an orthonormal basis for the columns of X.

The 1st column is the best 1D line that fits to all data.

All the singular vectors are orthogonal to one another.

The Singular Value Decomposition (SVD)

Consider a matrix $X \in \mathbb{R}^{nxp}$. There exist matrices U, Σ, V such that

$X = U\Sigma V^T$

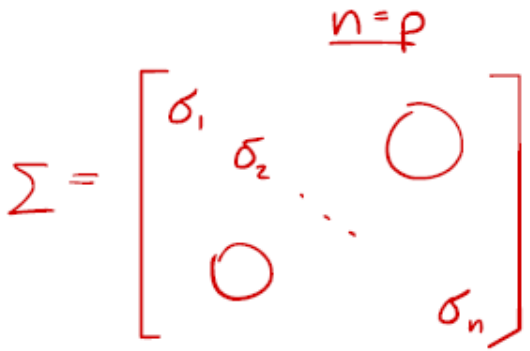
- The columns of V are called the "right singular vectors".
- V is an orthogonal matrix ($V^TV = VV^T = I$).
- The columns of V^T (rows of V) are the basis coefficients (weights on the column of U) needed to represent each column of X.

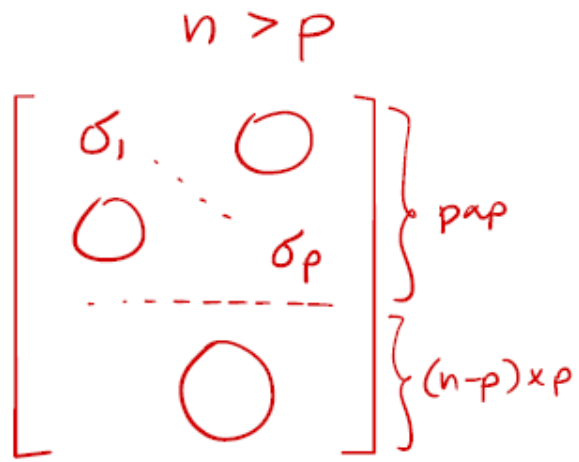
The Singular Value Decomposition (SVD)

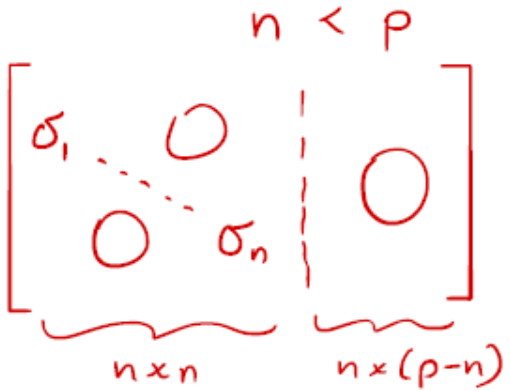
Consider a matrix $X \in \mathbb{R}^{nxp}$. There exist matrices U, Σ, V such that

X = UΣV^T

- Σ is diagonal with non-negative diagonal elements.







$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$$

Let $U = [u_1, u_2, \ldots, u_n]$,

These vectors form the basis for all of the columns in the matrix X.

u_1 is the best 1D subspace fit to x_i's (all data).

$$u_1 = \underset{\underline{a}}{\mathbf{argmax}} \frac{\underline{a}^T \boldsymbol{X}\boldsymbol{X}^T \underline{a}}{\underline{a}^T \underline{a}}$$

$\tilde{x}_i^{(1)} = \underline{x}_i - Proj_{u_1}\underline{x}_i$ projection x_i onto u_1

u_2 is the best 1D subspace fit to $\tilde{x}_i^{(1)}$'s.

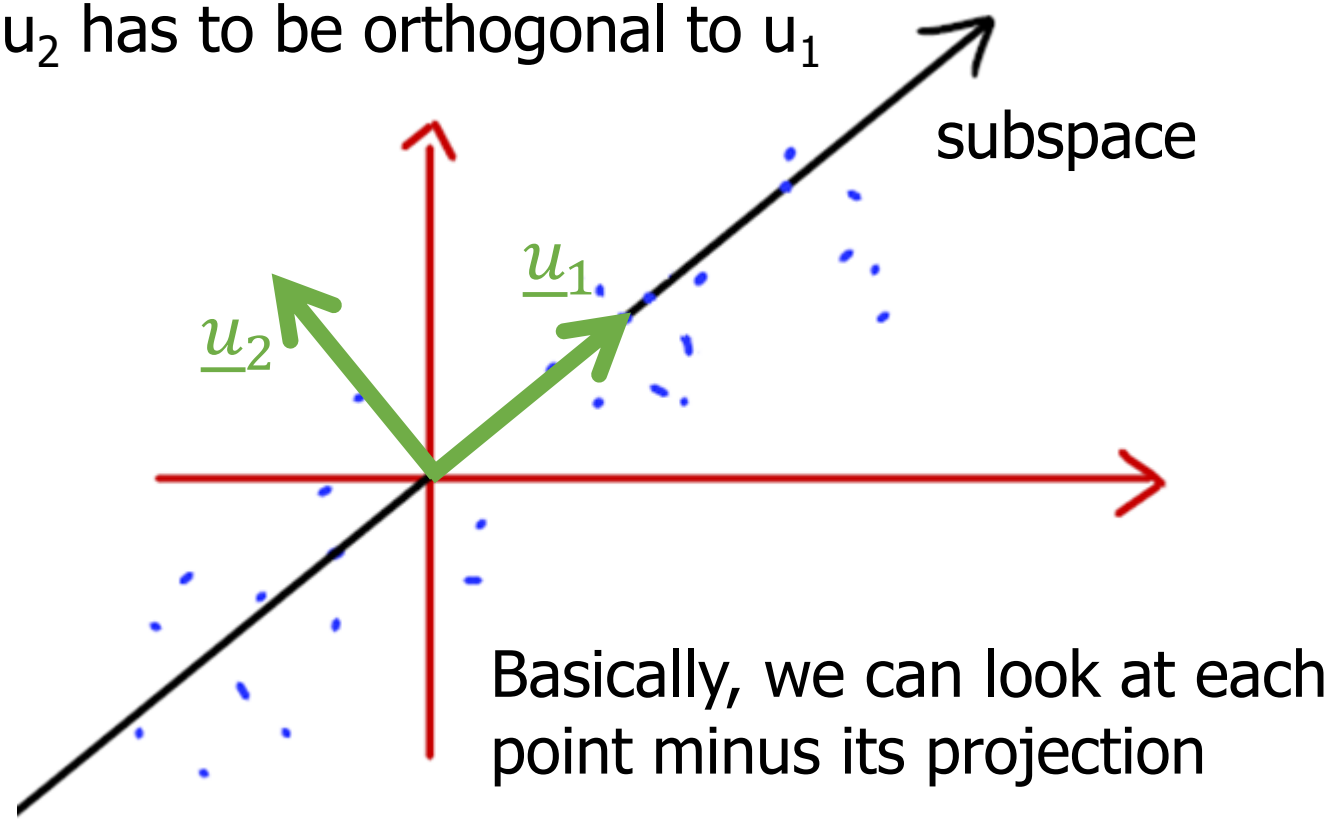
$\tilde{x}_i^{(2)} = \underline{x}_i - Proj_{u_1}\underline{x}_1 - Proj_{u_2}\underline{x}_1$.

u_3 is the best 1D subspace fit to $\tilde{x}_i^{(2)}$'s.

…

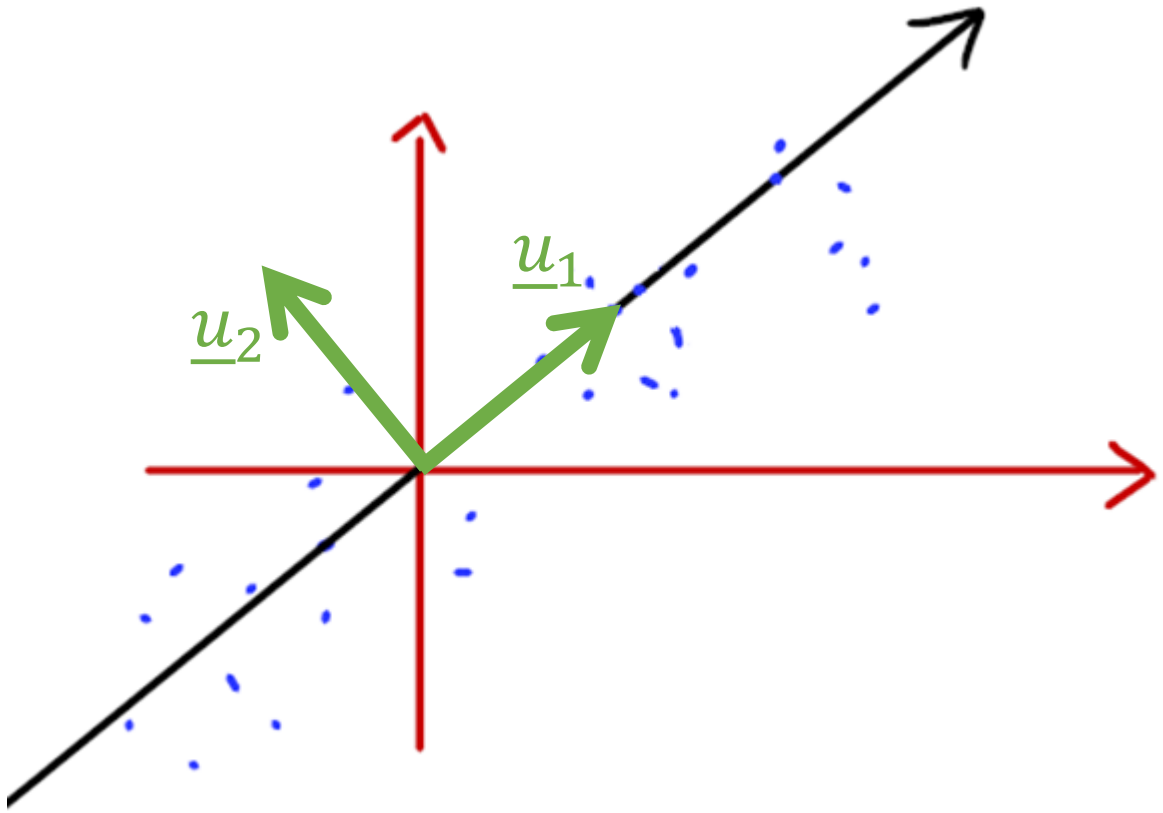
$[u_1, u_2, \ldots, u_k]$ is best k-dim subspace fit to x_i's.

u_2 has to be orthogonal to u_1

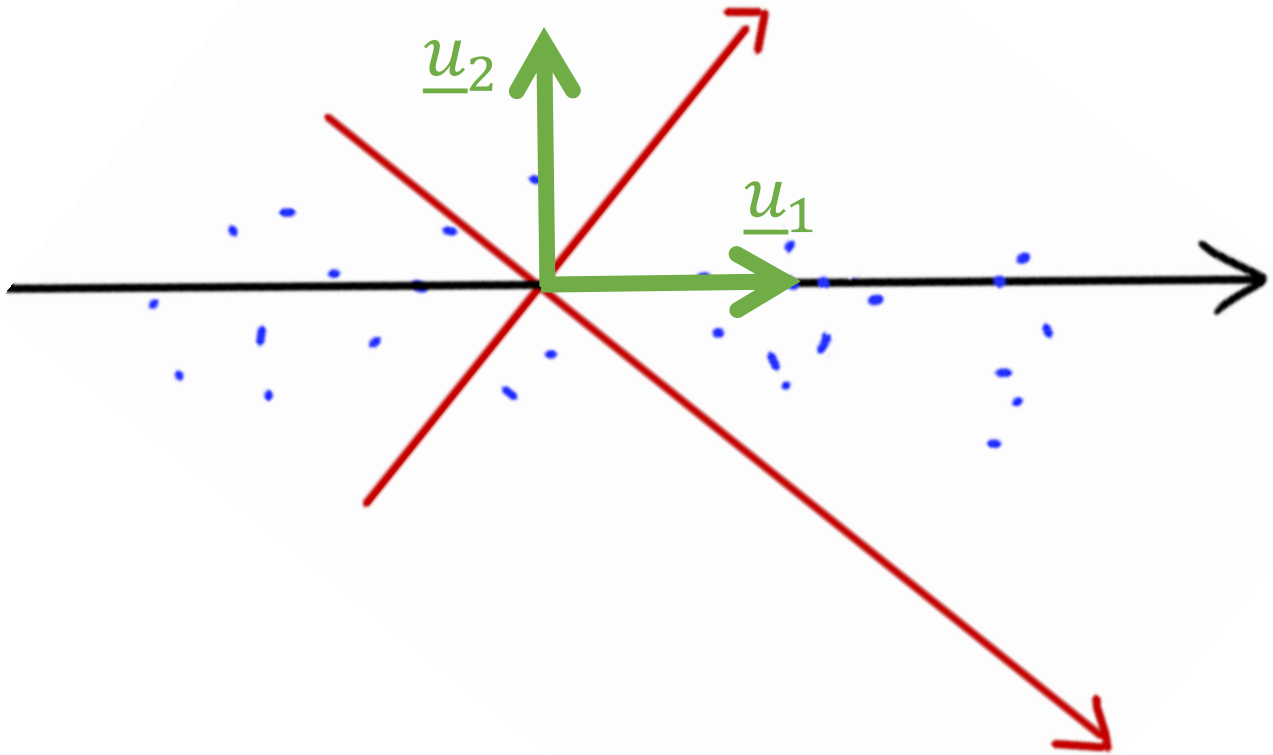


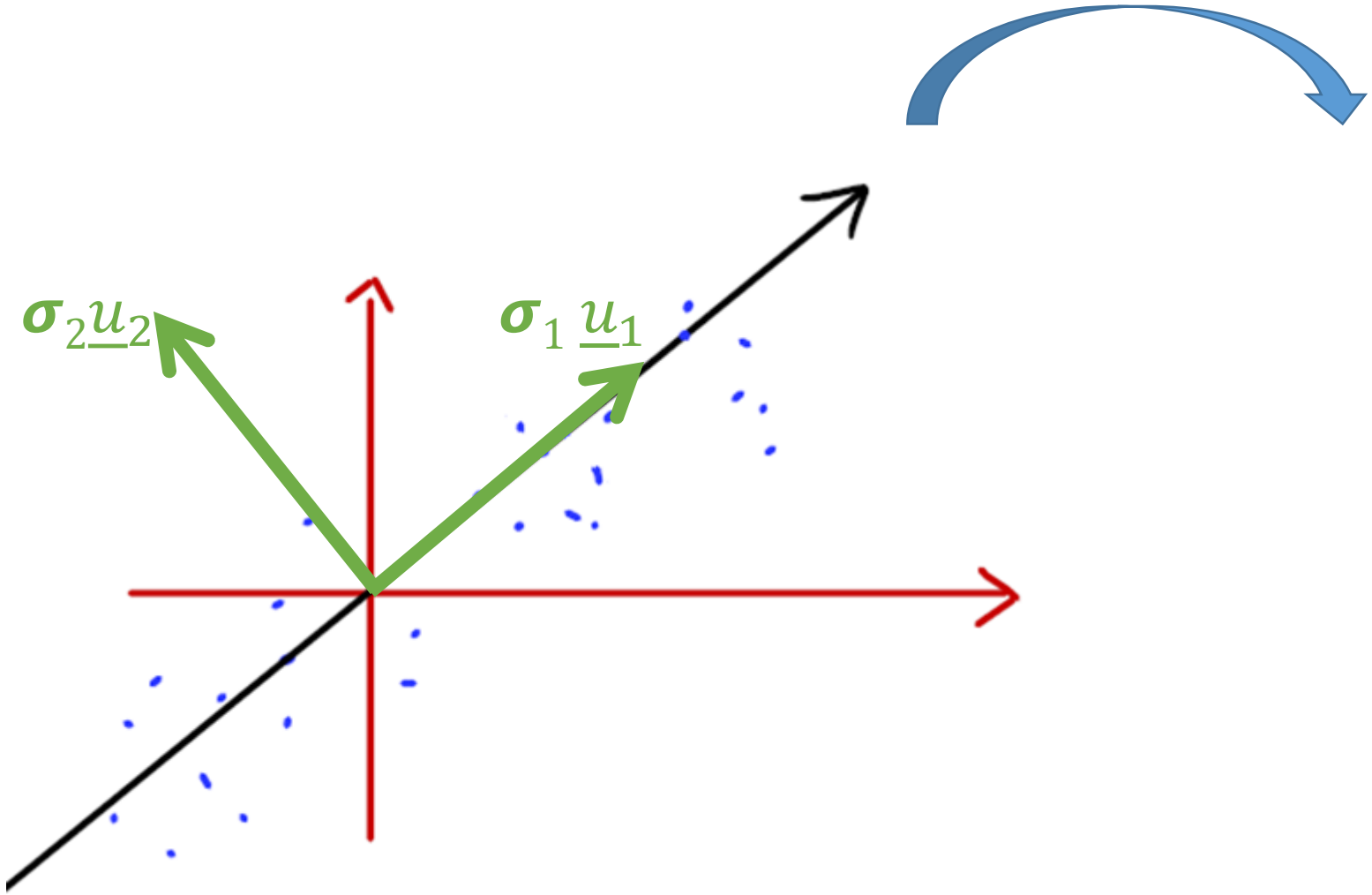
Basically, we can look at each point minus its projection

Cartesian Coordinates

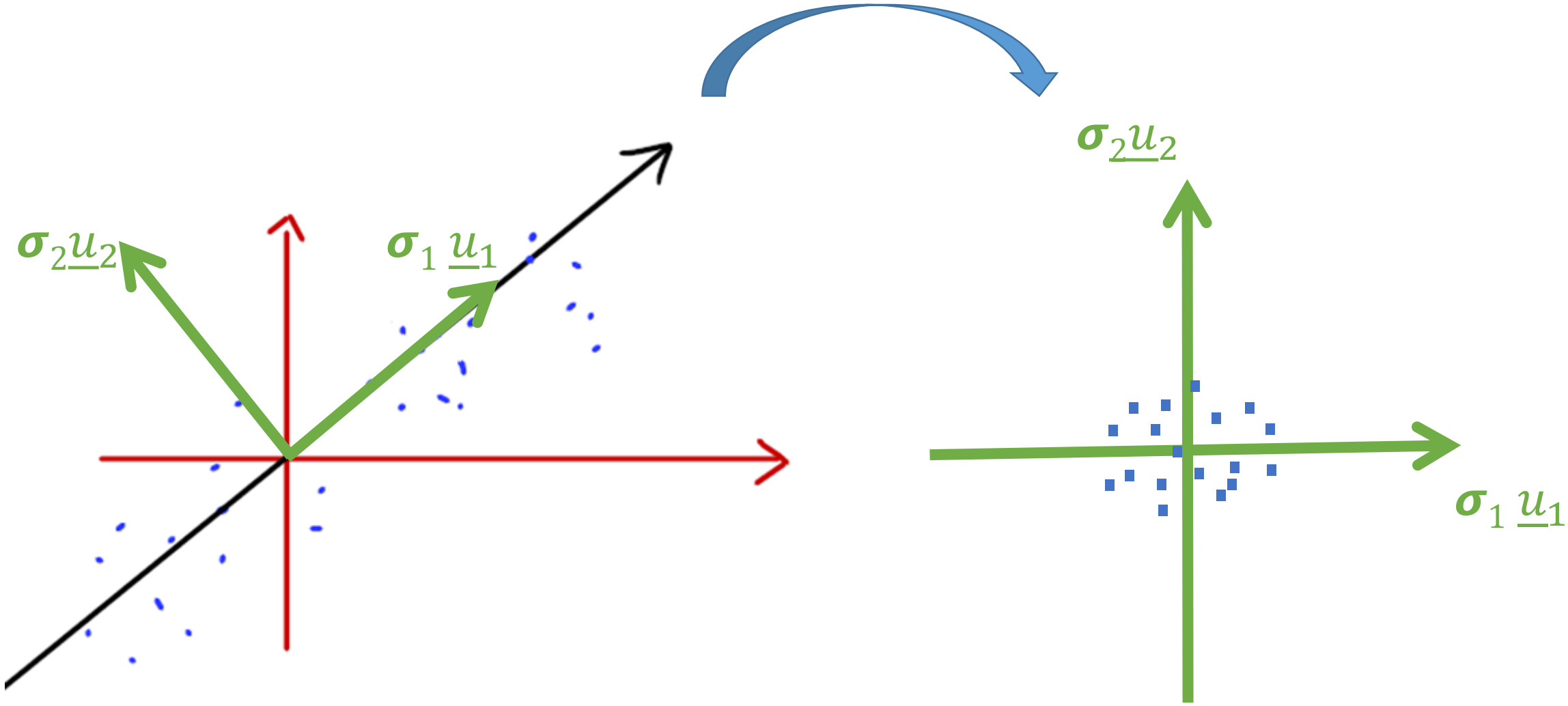


u_1 and u_2 directions

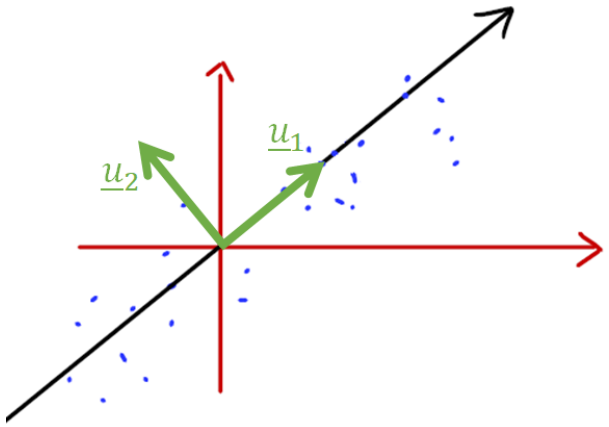




Singular values $\sigma_1, \sigma_2, \ldots$ indicate how spread out points are in the subspace



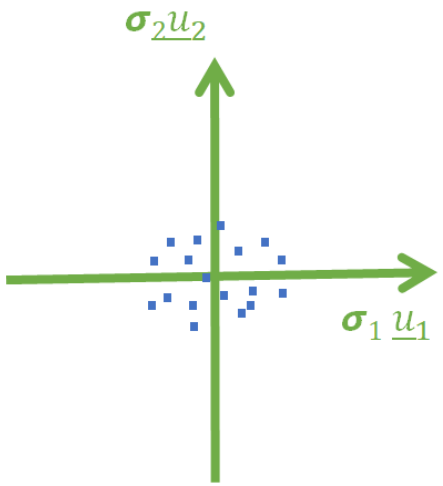
Any point x = $\sigma_1 u_1 v_1 + \sigma_2 u_2 v_2$



Using the u_i's, we **rotated** all of our data points

Using the σ_i, we **rescaled** the axes

and what v_i's do is, within the new coordinate system where the point was located.

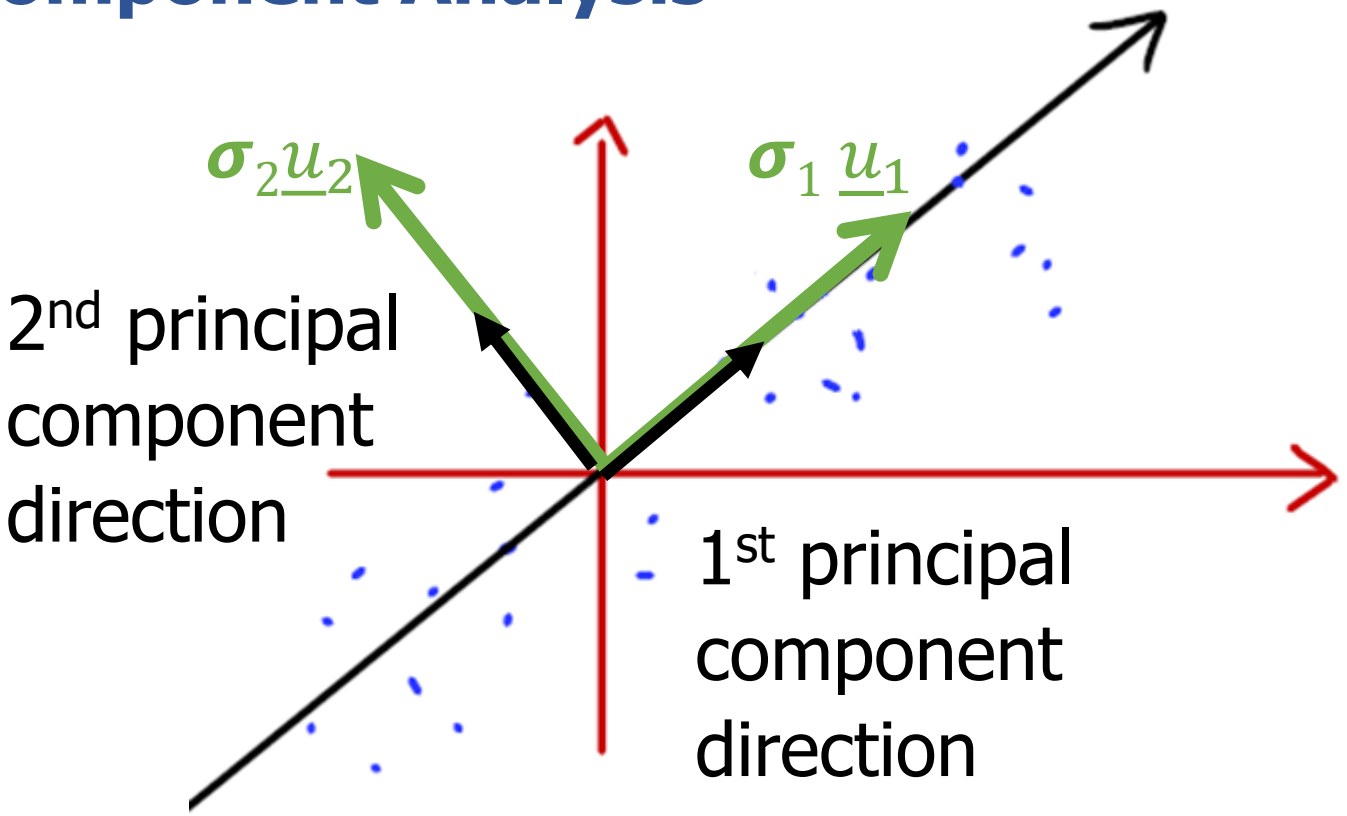


So instead of describing where our point is located in the Cartesian coordinate,

we describe where point is located in the new rescaled and rotated coordinate defined by u_i's and σ_i.

So, the v is sort of the analog of the Cartesian coordinates, but they corresponds to this new coordinate system

Principal Component Analysis



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

The 1st principal component directions = the 1st left singular vector of a matrix

The 2nd principal component directions = the 2nd left singular vector of a matrix

Announcements

- Homework #1 out
  - 스스로 복습한 내용을 A4용지에 손글씨로 작성/스캔하여 제출
  - 1-page 이상 per a lecture (lectures 3-5)
  - Any forms (pdf, jpeg, etc.) should be fine!
  - **Due Today March 23th Tuesday at 11:59 pm**