SWCON253 Machine Learning

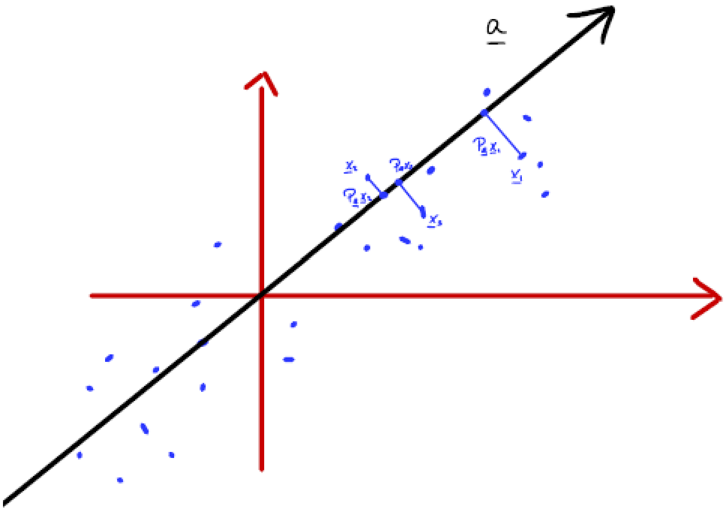
Lecture 7. Introduction to Singular Value Decomposition

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,

$Proj_A \underline{X}$ = $A(A^TA)^{-1}A^TX$

- $\boldsymbol{P_A = P_A^2 = P_A P_A = P_A^T = P_A^T P_A}$
- If A = $\underline{a}$, then $P_a = a(a^Ta)^{-1}a^T = \frac{\boldsymbol{aa^T}}{\boldsymbol{a^Ta}}$
- The orthogonal complement of a subspace is the orthogonal to the subspace.

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

$I\underline{x} = P_A\underline{x} + P_B\underline{x} = (P_A + P_B)\underline{x}$
$\rightarrow \mathrm{I} = P_A + P_B$
$\rightarrow \boldsymbol{P_B} = \mathbf{I} - \boldsymbol{P_A}$

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

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

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

Projection on $\underline{a}$ (orthogonal complement)

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

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.

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

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

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

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

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

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

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

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 = UΣV^T

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

$U \in \mathbb{R}^{n\times n}$ is orthogonal $(U^TU = UU^T = I)$, called left singular vectors
$V \in \mathbb{R}^{p\times p}$ 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

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 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) need to represent each column of X.

Σ is diagonal with non-negative diagonal elements.

$n = p$

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

$n > p$

$$\begin{bmatrix} \begin{matrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \end{matrix} \\ \hline 0 \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} p \times p \\ \left.\vphantom{a}\right\} (n-p) \times p \end{matrix}$$

$n < p$

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

$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p$

Let U = [u_1, u_2, …, u_n]

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

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

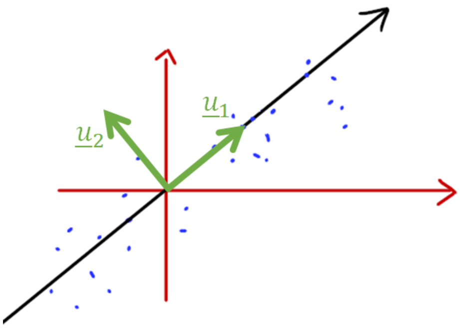
u_2 is the best 1D subspace fit to $\tilde{x}_i^{(1)}$'s.

$\tilde{x}_i^{(2)} = \underline{x_i} - Proj_{u_1}\underline{x_1} - Proj_{u_2}\underline{x_1}$.

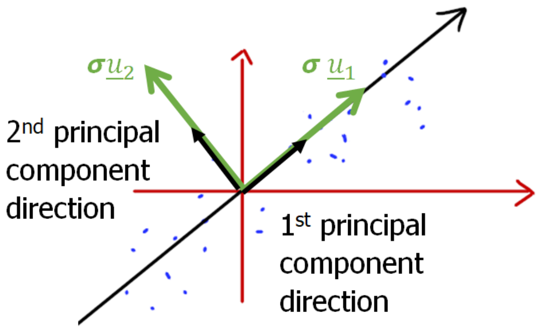
u_3 is the best 1D subspace fit to $\tilde{x}_i^{(2)}$'s.

…

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



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