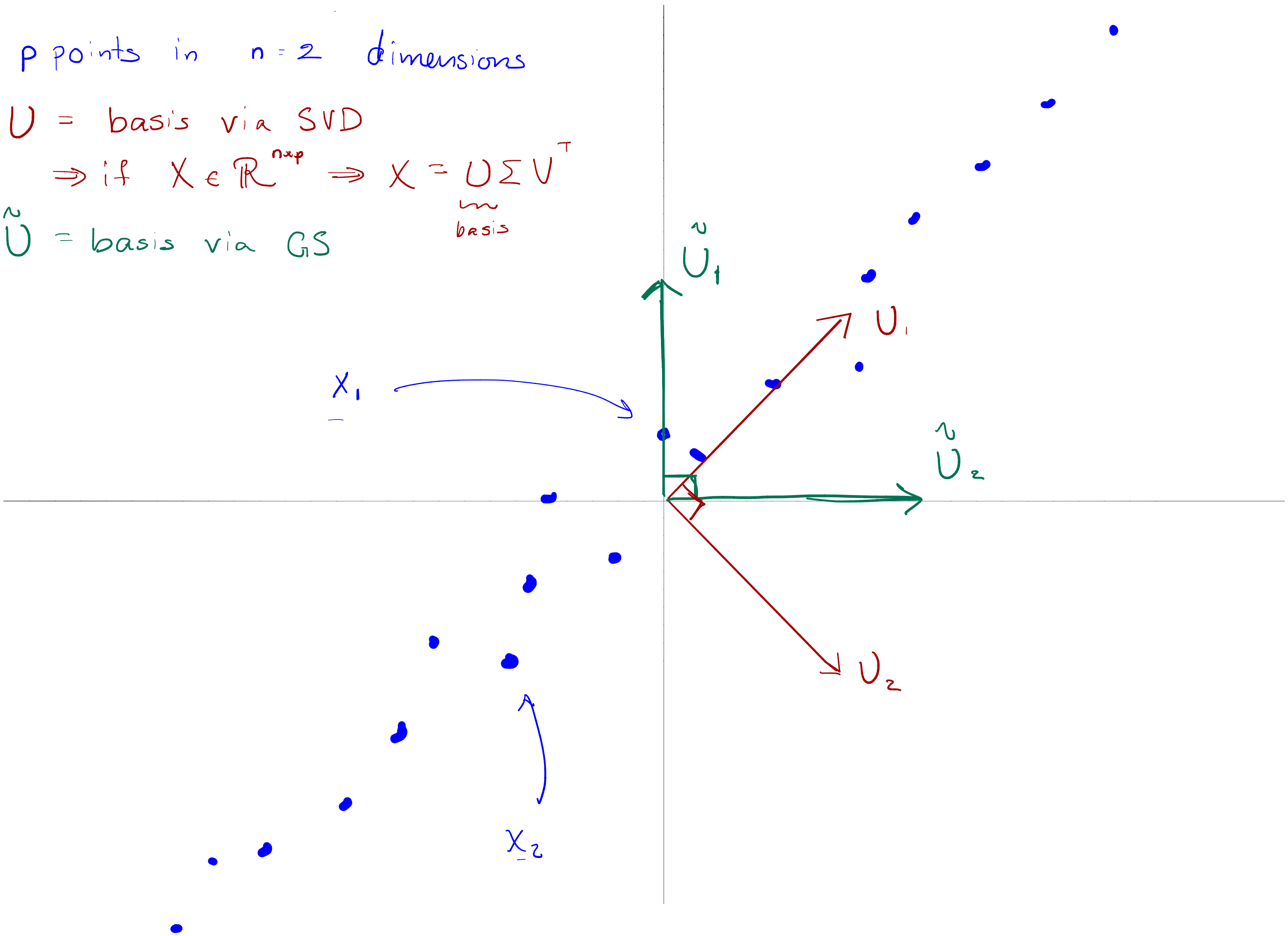
Lecture 7 : Introduction to the Singular Value Decomposition

p points in $n = 2$ dimensions

U = basis via SVD

$\Rightarrow$ if $X \in \mathbb{R}^{n \times p} \Rightarrow X = \underbrace{U}_{\text{basis}} \Sigma V^T$

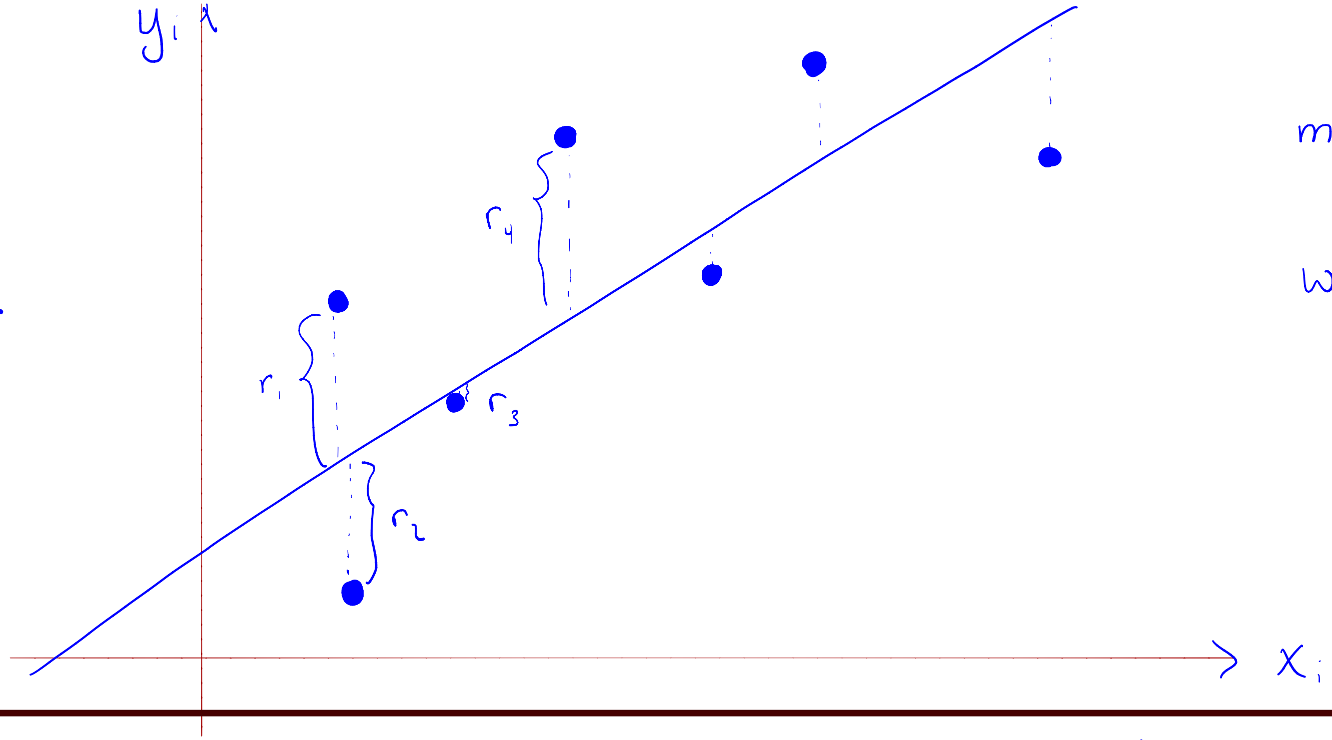
$\tilde{U}$ = basis via GS



$\underline{U}_1$ is the 1d subspace that is closest to all the $\underline{X}_i$'s (ie. best 1d subspace fit)

$x_i \in \mathbb{R}$

L.S.



minimize $\sum_i r_i^2$

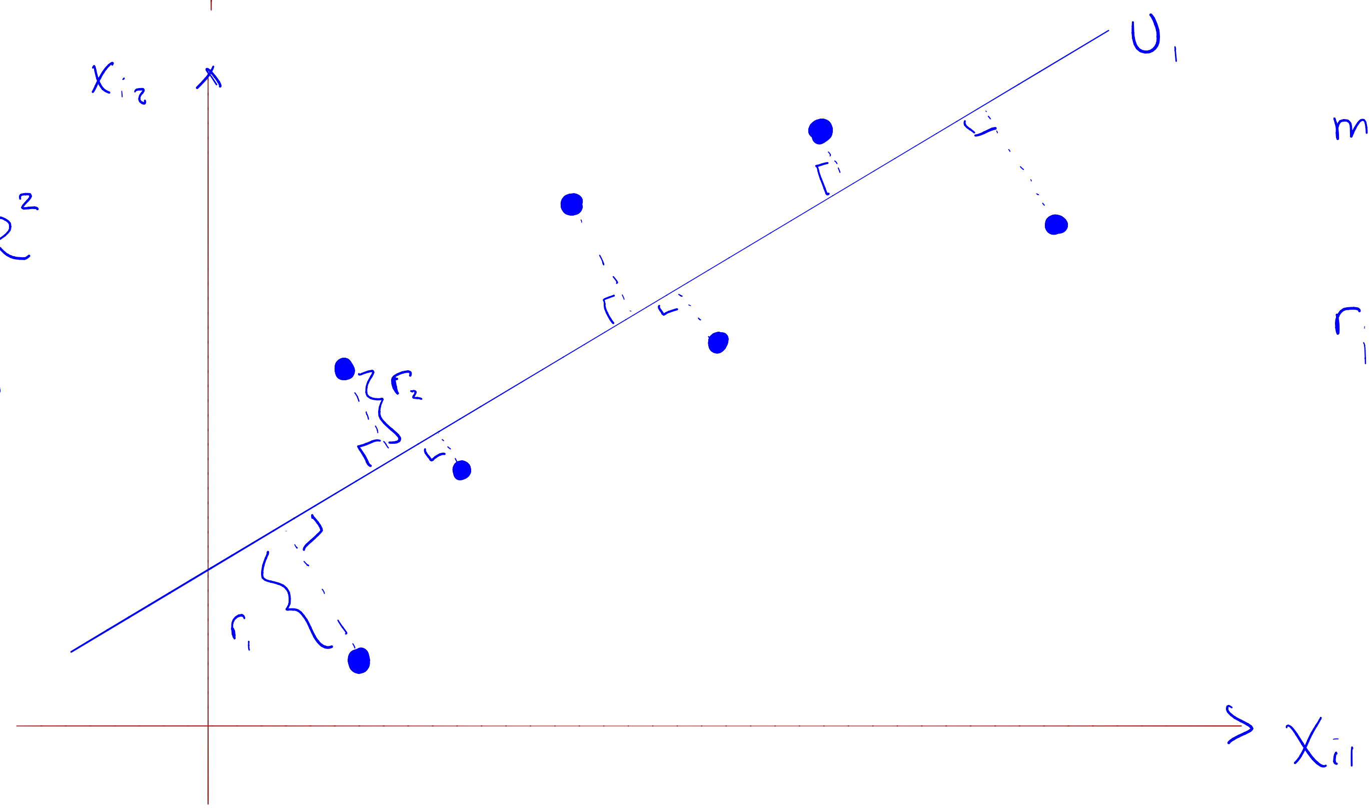
write

$y_i \approx \underbrace{a\, x_i + b}_{\underline{w}}$

$r_i = y_i - (a x_i + b)$

$x_i \in \mathbb{R}^2$

SVD

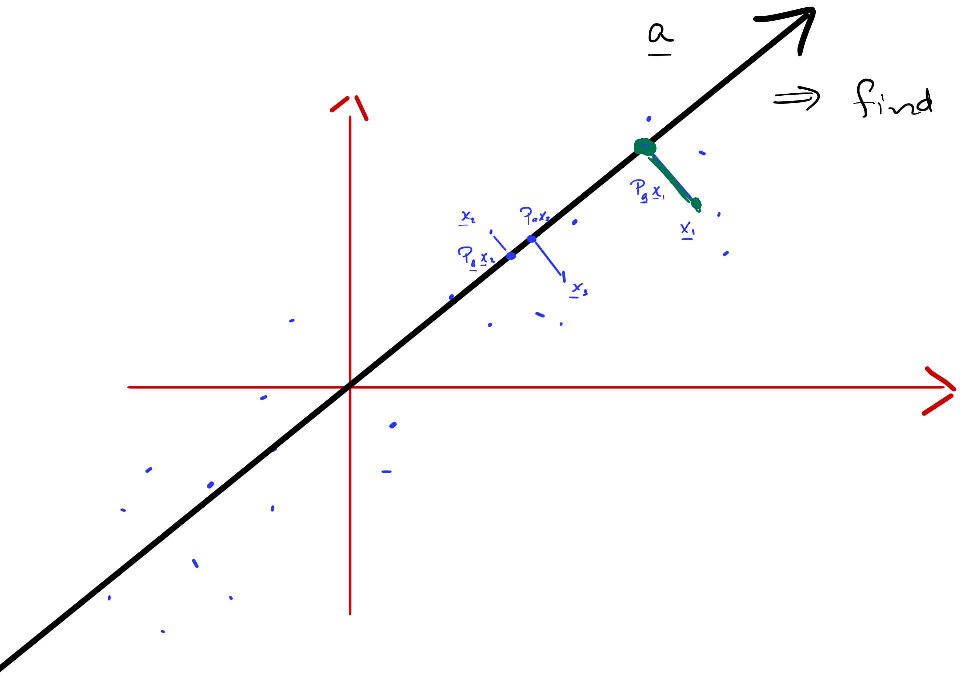


minimize $\sum_i \|r_i\|_2^2$

$r_i = \underline{x}_i - P_{U_1} \underline{x}_i$

Introduction to the Singular Value Decomposition (SVD)

Ex. find a 1d-subspace (line through origin) that is closest to a set of points $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n \in \mathbb{R}^p$



$\Rightarrow$ find $\underline{a}$ to minimize sum of squared distances

Projection Matrices

$$P_A = A(A^TA)^{-1}A^T \Rightarrow P_{\underline{a}} = \underline{a}(\underline{a}^T\underline{a})^{-1}\underline{a}^T$$

$$P_A^2 = A(A^TA)^{-1}\cancel{A^TA(A^TA)^{-1}}A^T$$

$$= A(A^TA)^{-1}A^T = P_A$$

given a subspace S spanned by columns of A, the orthogonal complement of S (aka orthogonal complement of A) is the set of all vectors b orthogonal to all columns of A (i.e. orthogonal to every vector in S)

let $A \in \mathbb{R}^{p\times r}$, $B \in \mathbb{R}^{p\times(p-r)}$ be orthogonal complements

$\Rightarrow$ $A^TB = 0$ and any $x \in \mathbb{R}^p$ can be written as $\underline{x} = \underbrace{A\underline{u}}_{P_A\underline{x}} + \underbrace{B\underline{v}}_{P_B\underline{x}}$ for some $\underline{u} \in \mathbb{R}^r$, $\underline{v} \in \mathbb{R}^{p-r}$

$\Rightarrow P_A + P_B = I$

$\Rightarrow I - P_A = P_A + P_B - P_A = P_B$

$\Rightarrow I - P_A$ is also a projection matrix!

distance from $\underline{x}_i$ to line $\underline{a}$:

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

$$= \| \underline{x}_i - \underline{a}(\underline{a}^T\underline{a})^{-1} \underline{a}^T \underline{x}_i \|_2^2$$

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

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

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

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

$$= \underline{x}_i^T \underline{x}_i - \frac{(\underline{a}^T \underline{x}_i)^2}{\underline{a}^T \underline{a}}$$

Want to minimize $\sum_{i=1}^p d_i^2 = \sum_{i=1}^p \left(\underline{x}_i^T \underline{x}_i - \frac{(\underline{a}^T\underline{x}_i)^2}{\underline{a}^T\underline{a}} \right)$

constant with respect to $\underline{a}$

$$\Rightarrow \arg\min_{\underline{a}} \sum_{i=1}^p d_i^2(\underline{a}) = \arg\max_{\underline{a}} \sum_{i=1}^p \frac{\underline{a}^T \underline{x}_i \underline{x}_i^T \underline{a}}{\underline{a}^T \underline{a}} = \arg\max_{\underline{a}:\ \underline{a}^T\underline{a}=1} \sum_{i=1}^p \underline{a}^T \underline{x}_i \underline{x}_i^T \underline{a}$$

Let $X = [\underline{x}_1, \underline{x}_2, \dots, \underline{x}_p] \in \mathbb{R}^{n \times p}$

$$\Rightarrow \underline{U}_1 = \arg\max_{\underline{a}:\ \underline{a}^T\underline{a}=1} \sum_i \underline{a}^T \underline{x}_i \underline{x}_i^T \underline{a} = \arg\max_{\underline{a}:\ \underline{a}^T\underline{a}=1} \underline{a}^T XX^T \underline{a}$$

value of $\underline{a}$ that achieves maximum is 1st left singular vector of X, denoted $\underline{U}_1$

also, $\sigma_1 = \max_{\underline{a}:\ \underline{a}^T\underline{a}=1} \| X^T \underline{a} \|_2 = \| X^T \underline{U}_1 \|_2$

is called the 1st singular value of X

(also called the "operator norm" of X: $\|X\|_{op} = \|X\|_2$)

not like 2-norm of vector!

The bigger σ_1 is, the smaller $\sum_i d_i^2$ is

$\rightarrow$ the better the $\underline{x}_i$'s are aligned with a 1-d subspace

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 ($U^T U = U U^T = I$), called left singular vectors
$V \in \mathbb{R}^{p \times p}$ is orthogonal ($V^T V = V V^T = I$), called right singular vectors
$\Sigma \in \mathbb{R}^{n \times p}$ is diagonal; diagonal elements called singular values

$n = p$

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

$n > p$

$$\begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \\ \hline & 0 & \end{bmatrix} \begin{array}{l} \} \; p \times p \\ \\ \\ \} \; (n-p) \times p \end{array}$$

$n < p$

$$\begin{bmatrix} \underbrace{\begin{matrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{matrix}}_{n \times n} & \Bigg| & \underbrace{\begin{matrix} 0 \end{matrix}}_{n \times (p-n)} \end{bmatrix}$$

The columns of U form an orthonormal basis for the columns of X.

The singular values weigh (scale the length) of the corresponding singular vectors.

- The number of non-zero singular vectors is the RANK of X.

The columns of V^T (rows of V) are the basis coefficients (weights on the columns of $U\Sigma$) needed to represent each column of X.

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

  $\underline{\tilde{x}}_i^{(1)} = \underline{x}_i - P_{u_1} \underline{x}_i$ = i^{th} residual $\forall i$

  u_2 = best 1d subspace fit to all $\underline{\tilde{x}}_i^{(1)}$

  ⋮

  $\underline{\tilde{x}}_i^{(k)} = \underline{x}_i - P_{[u_1, \ldots, u_k]} \underline{x}_i = \underline{\tilde{x}}_i^{(k-1)} - P_{u_k} \underline{\tilde{x}}_i^{(k-1)} \quad \forall i$

  U_{k+1} = best 1d subspace fit to all $\underline{\tilde{x}}_i^{(k)}$

singular values σ_1, σ_2, etc. indicate how spread out points are in the subspace.

Recall Gram-Schmidt:

$$U_1 = x_1 / \|x_1\|_2$$

$$\underline{\tilde{x}}_2 = \underline{x}_2 - P_{u_1} \underline{x}_1$$

$$U_2 = \tilde{x}_2 / \|\tilde{x}_2\|_2$$

order of points matters!

basis vectors less interpretable