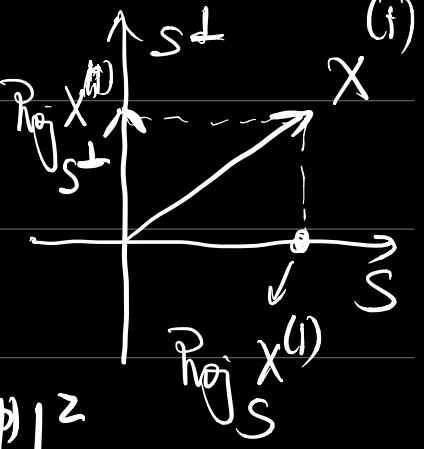
For a sample point $x^{(i)}$, we have

$$x^{(i)} = \underbrace{(x^{(i)T} u^{(1)})\, u^{(1)} + \cdots + (x^{(i)T} u^{(d)})\, u^{(d)}}_{= \text{Proj}_S x^{(i)} \in S} + \underbrace{\cdots + (x^{(i)T} u^{(p)}) u^{(p)}}_{= \text{Proj}_{S^\perp} x^{(i)} \in S^\perp}$$

squared distance from $x^{(i)}$ to S

$$= \| \text{Proj}_{S^\perp} x^{(i)} \|^2$$

$$= \| (x^{(i)T} u^{(d+1)}) u^{(d+1)} + \cdots + (x^{(i)T} u^{(p)}) u^{(p)} \|^2$$



Pythagorean

$$= |(x^{(i)T} u^{(d+1)})|^2 + \cdots + |u^{(i)T} u^{(p)}|^2$$

$$= \sum_{j=d+1}^{p} |x^{(i)T} u^{(j)}|^2$$

Therefore, the total squared distance from the sample points to S

$$= \sum_{i=1}^{n} \sum_{j=d+1}^{p} |x^{(i)T} u^{(j)}|^2 \qquad (*)$$

(This is what is to be minimized)

We compute this in another way:

For $x^{(i)}$, Pythagorean theorem gives:

$$\|x^{(i)}\|^2 \overset{\text{Pythagorean}}{=} \|\text{Proj}_S x^{(i)}\|^2 + \|\text{Proj}_{S^\perp} x^{(i)}\|^2$$

$$= \|(x^{(i)T} u^{(1)}) u^{(1)} + \dots + (x^{(i)T} u^{(d)}) u^{(d)}\|^2$$

$$+ \|(x^{(i)T} u^{(d+1)}) u^{(d+1)} + \dots + (x^{(i)T} u^{(p)}) u^{(p)}\|^2$$

$$\overset{\text{Pythagorean}}{=} |x^{(i)T} u^{(1)}|^2 + \dots + |x^{(i)T} u^{(d)}|^2$$

$$+ |x^{(i)T} u^{(d+1)}|^2 + \dots + |x^{(i)T} u^{(p)}|^2$$

$$= \underbrace{\sum_{j=1}^{d} |x^{(i)T} u^{(j)}|^2}_{= \|\text{Proj}_S x^{(i)}\|^2} + \underbrace{\sum_{j=d+1}^{p} |x^{(i)T} u^{(j)}|^2}_{= \|\text{Proj}_{S^\perp} x^{(i)}\|^2}$$

Insert $\sum_{j=d+1}^{p} |x^{(i)T} u^{(j)}|^2 = \|x^{(i)}\|^2 - \sum_{j=1}^{d} |x^{(i)T} u^{(j)}|^2$

into (*):

the total squared distance =

$$\sum_{i=1}^{n} \left(\|x^{(i)}\|^2 - \sum_{j=1}^{d} |x^{(i)T} u^{(j)}|^2 \right)$$

$$= \underbrace{\sum_{i=1}^{n} \|x^{(i)}\|^2}_{\text{indep of } u^{(1)} \ldots u^{(p)}} - \sum_{i=1}^{n} \sum_{j=1}^{d} |x^{(i)T} u^{(j)}|^2$$

It remains to maximize (by choosing $u^{(1)}, \ldots, u^{(d)}$).

$$\sum_{i=1}^{n} \sum_{j=1}^{d} |x^{(i)T} u^{(j)}|^2 = \sum_{j=1}^{d} \left(\sum_{i=1}^{n} |\underbrace{x^{(i)T} u^{(j)}}_{\text{scalar}}|^2 \right)$$

$$= \sum_{j=1}^{d} \left(\sum_{i=1}^{n} (x^{(i)T} u^{(j)})^T (x^{(i)T} u^{(j)}) \right)$$

$$= \sum_{j=1}^{d} \left(\sum_{i=1}^{n} u^{(j)T} (x^{(i)} x^{(i)T}) u^{(j)} \right)$$

$$= \sum_{j=1}^{d} \left(u^{(j)T} \sum_{i=1}^{n} (x^{(i)} x^{(i)T}) u^{(j)} \right)$$

computed this before

$$= \sum_{j=1}^{d} \left(u^{(j)T} \, XX^T \, u^{(j)} \right)$$

$$= u^{(1)T}(XX^T)u^{(1)} + \dots + u^{(d)T}(XX^T)u^{(d)}$$

This is the sum of d Rayleigh quotients.

To maximize $u^{(1)T}(XX^T)u^{(1)}$, we choose

$u^{(1)}$ to be a normalized eigenvector associated to the largest eigenvalue of XX^T.

To maximize $u^{(2)T}(XX^T)u^{(2)}$ with $u^{(2)} \perp u^{(1)}$ (because $\{u^{(1)}, \dots, u^{(d)}\}$ is an orthonormal basis of S), we choose $u^{(2)}$ to be a normalized eigenvector associated to the second largest eigenvalue of XX^T

⋮

Thm: The d-dimension subspace that best fits the sample points is spanned by the d eigenvectors associated to the d largest eigenvalues of XX^T.

3.3 PCA: Optimization Viewpoint:

Let $x^{(1)}, \dots, x^{(n)} \in \mathbb{R}^p$ be sample points and $X = \begin{bmatrix} | & & | \\ x^{(1)} & \cdots & x^{(n)} \\ | & & | \end{bmatrix}$ be the sample matrix.

Q: find a matrix $A \in \mathbb{R}^{p \times n}$ such that

$$\min_A \|X - A\|_F \quad \text{s.t. } \operatorname{rank}(A) = d \ (< p)$$

(i.e. find the optimal low-rank approximation of rank d)

3.3.1 Math Prep.

Recall: Given two matrices $A = (A_{ij})$, $B = (B_{ij})$ $\in \mathbb{R}^{m \times n}$, their <u>Frobenius inner product</u> is

$$(A, B)_F \overset{\text{def}}{=} \operatorname{tr}(A^T B) = \operatorname{tr}(B^T A) = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} B_{ij}$$

(because $A^T B = \begin{pmatrix} A_{11} & A_{21} & \cdots & A_{m1} \\ A_{12} & A_{22} & \cdots & A_{m2} \\ \vdots & \vdots & & \vdots \\ A_{1n} & A_{2n} & \cdots & A_{mn} \end{pmatrix} \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1n} \\ B_{21} & B_{22} & \cdots & B_{2n} \\ \vdots & \vdots & & \vdots \\ B_{m1} & B_{m2} & \cdots & B_{mn} \end{pmatrix}$

$= \begin{pmatrix} A_{11}B_{11} + A_{21}B_{21} + \cdots + A_{m1}B_{m1} & & \\ & A_{12}B_{12} + A_{22}B_{22} + \cdots + A_{m2}B_{m2} & * \\ & & \ddots \\ * & & A_{1n}B_{1n} + A_{2n}B_{2n} + \cdots + A_{mn}B_{mn} \end{pmatrix}$

The induced <u>Frobenius norm</u> of A is

$$\|A\|_F^2 \overset{\text{def}}{=} \operatorname{tr}(A^T A) = \sum_{i=1}^{m} \sum_{j=1}^{n} |A_{ij}|^2$$