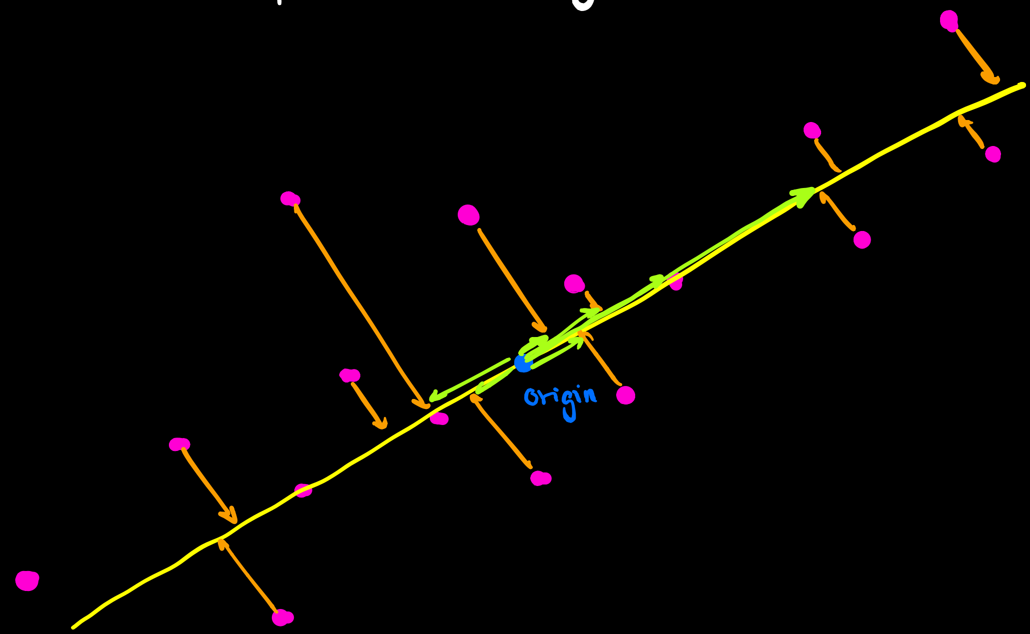
PCA (principle component analysis)



Q: What is the $\underbrace{\text{line of best fit}}$?

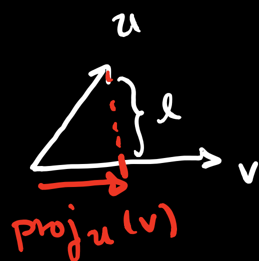
Want to minimize sum of the orange stuff.

$\iff$ maximize sum of the green stuff.

What is going on mathematically?

Preliminary math:

① The dot product has a geometric interpretation as a projection: If v is a unit vector, $(u \cdot v)v = \text{proj}_v u$



Note: $l = \| u - (u \cdot v) v \|$.

② Frobenius norm of a matrix has an alternative formulation:

$$A = \begin{bmatrix} \vec{a}_1 \\ \vdots \\ \vec{a}_n \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix}$$

$$\|A\|_F = \sqrt{\underbrace{(a_{11}^2 + \cdots + a_{1m}^2)}_{\vec{a}_1 \cdot \vec{a}_1} + \cdots\cdots + \underbrace{(a_{n1}^2 + \cdots + a_{nm}^2)}_{\vec{a}_n \cdot \vec{a}_n}}$$

$$= \sqrt{\underbrace{\sum \vec{a}_i \cdot \vec{a}_i}} \quad = \text{ sum of the diagonal entries of } AA^T.$$

Consider

$$AA^T = \begin{array}{c|c} & \begin{bmatrix} \vec{a}_1 & \cdots & \vec{a}_n \end{bmatrix} \\ \hline \begin{bmatrix} \vec{a}_1 \\ \vdots \\ \vec{a}_n \end{bmatrix} & \begin{bmatrix} \vec{a}_1 \cdot \vec{a}_1 & \cdots & \vec{a}_1 \cdot \vec{a}_n \\ \vdots & & \\ \vec{a}_n \cdot \vec{a}_1 & \cdots & \vec{a}_n \cdot \vec{a}_n \end{bmatrix} \end{array}$$

Define the trace of A to be $\operatorname{tr}(A) =$ sum of diagonal of A.

Then: $\|A\|_F = \sqrt{tr(AA^T)}$

(3) $tr(A+B) = tr(A) + tr(B)$

$tr(AB) = tr(BA)$

If I have data points $\vec{a}_1, \dots, \vec{a}_n$

Just consider $A = \begin{bmatrix} \vec{a}_1 \\ \vdots \\ \vec{a}_n \end{bmatrix}$.

I'm trying to minimize $\sum_{i=1}^{n} \|\vec{a}_i - \mathrm{proj}_w \vec{a}_i\|^2$, for

w a unit vector.

Rewrite this sum: $\underbrace{\sum_i \left(a_i - (w\cdot a_i)w\right)\cdot\left(a_i - (w\cdot a_i)w\right)}$

$$= \| A - \underbrace{Aw}_{\text{rows of this are } a_i\cdot w} w^T \|_F^2$$

$$\|A - Aww^T\|_F = \mathrm{tr}\left[(A - Aww^T)(A - Aww^T)^T\right]$$

$$= \mathrm{tr}\left[(A - Aww^T)(A^T - \underbrace{(Aww^T)^T}_{ww^TA^T})\right]$$

$$= \mathrm{tr}\left[AA^T \underline{- Aww^TA^T} - \underline{Aww^TA^T} + \underline{Aw\overbrace{\cancel{w^Tw}}^{1.}w^TA^T} \right.$$

w is assumed to be a unit vector.

$$= \operatorname{tr}\left[AA^T - Aww^TA^T\right]$$

$$= \operatorname{tr}(AA^T) - \underbrace{\operatorname{tr}(\overbrace{Aw}\overbrace{w^TA^T})}_{\operatorname{tr}(w^TA^TAw)}$$

$$= \operatorname{tr}(AA^T) - \underbrace{w^TA^T}\underbrace{Aw}$$

$$= \underbrace{\operatorname{tr}(AA^T)}_{\text{const}} - (Aw)\cdot(Aw)$$

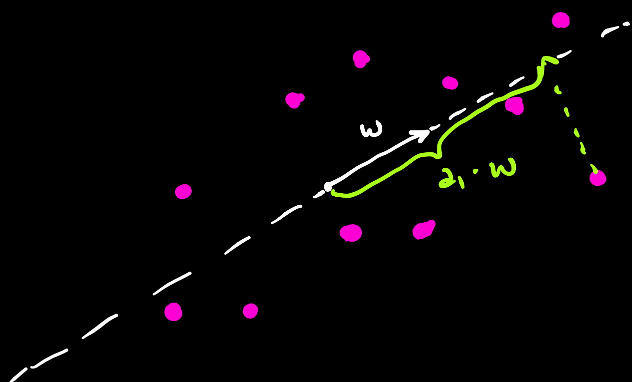
can I see this geometrically? Email me if you know!

Conclusion: $\|A - Aww^T\|_F = \text{const} - (Aw)\cdot(Aw)$

$\Rightarrow$ $\underbrace{\text{minimizing } \|A - Aww^T\|_F}_{\text{minimizes orange stuff}} \iff \underbrace{\text{maximizing } (Aw)\cdot(Aw)}$

$$A\vec{w} = \begin{bmatrix} \vec{a}_1 \\ \vdots \\ \vec{a}_n \end{bmatrix} w = \begin{bmatrix} \vec{a}_1 \cdot w \\ \vdots \\ \vec{a}_n \cdot w \end{bmatrix} \rightsquigarrow Aw \cdot Aw =$$

$$\sum_{i=1}^{n} (a_i \cdot w)(a_i \cdot w)$$



This is called <u>maximizing the varience</u>.

$w^T A^T A w$ is the key to why we care.

Last lecture: we saw that T^*T is self-adjoint.

$\Rightarrow$ A^TA is self-adjoint

$\Rightarrow$ Spectral thm $\Rightarrow$ A^TA is diagonalizable w.r.t. an orthonormal basis.

Max problem: given our orth. basis $w_1, \dots, w_m$

$$(b_1 w_1 + \dots + b_n w_n)^T A^T A (b_1 w_1 + \dots + b_n w_n) =$$

$$b_1^2 \lambda_1 + \dots + b_n^2 \lambda_n .$$

$\rightsquigarrow$ max. $b_1^2 \lambda_1 + \dots + b_n^2 \lambda_n$ subject to $\|b_1 w_1 + \dots + b_n w_n\|^2 = 1$

$$\underbrace{\qquad\qquad\qquad}_{b_1^2 + \dots + b_n^2 \;=\; 1}$$

$\rightsquigarrow$ Easy Lagrange multiplier problem!

Result: $\max\limits_{\|w\|=1} w^T A^T A w$ is realized by the eigenvector corresponding to the largest eigenvalue.

ZIPSHOT: To perform PCA, find the biggest eigenvalue and "its" unit eigenvector.

We saw: $\operatorname{span}(w)^{\perp} = \mathcal{U}$ is invariant under $A^T A$

$\rightsquigarrow$ Restrict $\boxed{A^T A}$ to $\mathcal{U}$, and perform PCA again.

→ the covariance matrix Σ of the data.

$\boxed{2}$ SVD. Why?

In matrix terms: SVD says $\boxed{A = \mathcal{U} D V}$, where $\mathcal{U}$ and V have orthonormal columns and D is diagonal.

$$\Sigma = A^T A = (\mathcal{U} D V)^T (\mathcal{U} D V)$$

$$= V^T D^T \overset{I}{\cancel{U^T U}} D V$$

$$= V^T D^T D V.$$

$$= V^T D^2 \underbrace{V}.$$

The columns of V are (the) eigenvectors of Σ.

Zipshot: performing SVD on A yields the eigenvectors desired as the output of PCA.