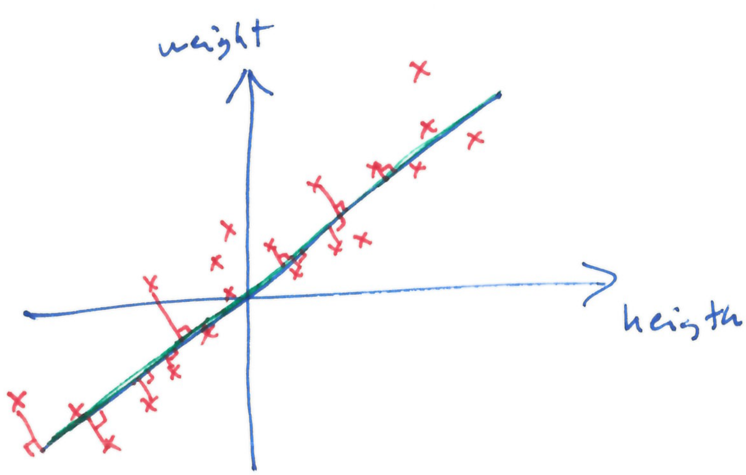
SVD & Best Least Square Fit

In the $\boxed{\text{first formulation}}$ of this problem, – i.e. "the best least square fit problem" – we want to find a subspace which minimizes the distance of a given collection of points from that subspace.



n samples (red points)

two variables (e.g. height and weight)

data already centered around means ⇝ negative h/w makes sense.

green line = the best one-dim subspace of $\mathbb{R}^2$ minimizing the sum of squared distances

Another name for this problem is "perpendicular least squares".

Also sometimes called "orthogonal regression".

Let's formulate this problem in more generality (higher dimension).

Suppose we have n points (data points, sample) in a d-dimensional linear space ($d = 2$ in the figure of last page).

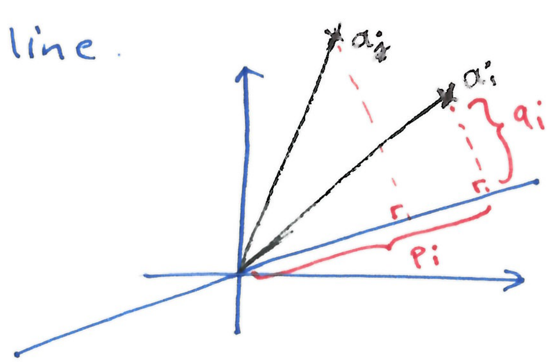
Geometric Understanding

We have n points $a_1, \dots, a_n$ in the d-dim space $\mathbb{R}^d$.

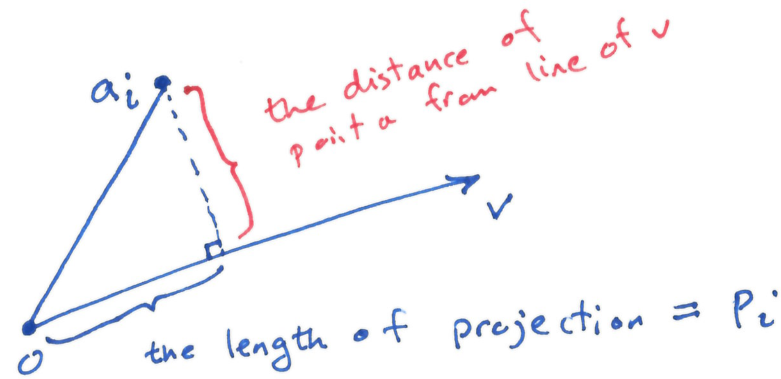
Arrange them in an $n \times d$ matrix A.

$$
\begin{array}{r c}
\text{point } a_1 \rightarrow & \\
\text{point } a_2 \rightarrow & \\
\vdots & \\
\text{point } a_n \rightarrow &
\end{array}
\begin{bmatrix}
a_{11} & a_{12} & \cdots & a_{1d} \\
a_{21} & a_{22} & \cdots & a_{2d} \\
\vdots & \vdots & & \vdots \\
a_{n1} & a_{n2} & \cdots & a_{nd}
\end{bmatrix}
$$

Consider a line through origin determined by a vector v, and project points $\{a_i\}_{i=1}^n$ onto this line.



(Pythagoras in higher dimensions)



P_i = the length of projection $= \dfrac{|\langle a_i, v\rangle|}{\|v\|}$

the vector of projection $= \dfrac{\langle a_i, v\rangle}{\|v\|^2}\, v$

Note that

$$\left\| \frac{\langle a, v\rangle}{\|v\|^2} v \right\| = \frac{|\langle a, v\rangle|}{\|v\|^2} \|v\| = \frac{|\langle a, v\rangle|}{\|v\|}$$

the distance of pt a from line of v $= \left\| a - \dfrac{\langle a, v\rangle}{\|v\|^2} v \right\|$

$\boxed{\text{Note}}$ $\quad \left\langle a - \dfrac{\langle a, v\rangle}{\|v\|^2} v \,,\, v \right\rangle =$

$$\langle a, v\rangle - \frac{\langle a, v\rangle}{\|v\|^2} \langle v, v\rangle = 0$$

So $a - \dfrac{\langle a, v\rangle}{\|v\|^2} v \perp v$

For each point $a_i \in \mathbb{R}^d$,

$$p_i^2 + q_i^2 = (\text{length of Projection})^2 + (\text{distance to line})^2$$

$$= \frac{|\langle a_i, v\rangle|^2}{\|v\|^2} + \left\| a_i - \frac{\langle a_i, v\rangle}{\|v\|^2} v \right\|^2$$

$$= \frac{|\langle a_i, v\rangle|^2}{\|v\|^2} + \left\langle a_i - \frac{\langle a_i, v\rangle}{\|v\|^2} v,\ a_i - \frac{\langle a_i, v\rangle}{\|v\|^2} v \right\rangle$$

$$= \frac{|\langle a_i, v\rangle|^2}{\|v\|^2} + \langle a_i, a_i\rangle - \frac{|\langle a_i, v\rangle|^2}{\|v\|^2} - \frac{\langle a_i, v\rangle\langle v, a_i\rangle}{\|v\|^2}$$

$$+ \frac{|\langle a_i, v\rangle|^2}{\|v\|^4}\langle v, v\rangle$$

$$= \frac{|\langle a_i, v\rangle|^2}{\|v\|^2} + \|a_i\|^2 - 2\frac{|\langle a_i, v\rangle|^2}{\|v\|^2} + \frac{|\langle a_i, v\rangle|^2}{\|v\|^2}$$

$$= \|a_i\|^2$$

$$= |a_{i1}|^2 + \cdots + |a_{id}|^2$$

Therefore

$$q_i^2 = (\text{distance of point } a_i \text{ to the line of } v)^2$$

$$= \|a_i\|^2 - p_i^2$$

$$= \|a_i\|^2 - \left(\frac{|\langle a_i, v\rangle|}{\|v\|}\right)^2$$

We want to minimize

$\sum_{i=1}^{n} q_i^2$ = the sum of the squares of the distances to the line

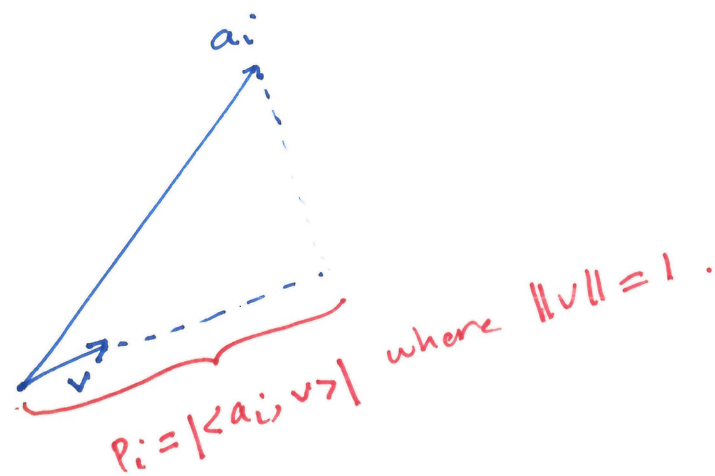
which is equivalent to maximizing

$\sum_{i=1}^{n} p_i^2$ = the sum of the squares of the lengths of the projections onto the lines.

We can choose v to be a vector of norm 1, since the line determined by v remains the same (does not depend on norm of v).

In that case,

$$p_i = \frac{|\langle a_i, v\rangle|}{\|v\|} = |\langle a_i, v\rangle|$$



$$\begin{bmatrix} \langle a_1, v\rangle \\ \langle a_2, v\rangle \\ \vdots \\ \langle a_n, v\rangle \end{bmatrix}_{n\times 1} = \begin{bmatrix} \boxed{\quad a_1 \quad} \\ \boxed{\quad a_2 \quad} \\ \vdots \\ \boxed{\quad a_n \quad} \end{bmatrix}_{n\times d} \begin{bmatrix} v_1 \\ \vdots \\ v_d \end{bmatrix}_{d\times 1}$$

$$\left[\langle a_i, v\rangle \right] = A_{n\times d} \quad v_{d\times 1}$$

Remember that we want to maximize

$$\sum_{i=1}^{n} p_i^2 .$$

But,

$$\sum_{i=1}^{n} p_i^2 = \sum_{i=1}^{n} |\langle a_i, v \rangle|^2 = \|Av\|_2^2$$

So, we should look for a vector $v \in \mathbb{R}^d$ such that $\|v\| = 1$ and $\|Av\|$ is maximized.

Let's call this vector v_1.

$$\boxed{v_1 = \underset{\|v\|=1}{\operatorname{argmax}} \|Av\|}$$

By definition of $\|A\|_2$,

$$\|A\|_2 = \max_{\|v\|=1} \|Av\| = \|Av_1\|$$

Now, perform SVD factorization on A:

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

$$\|A\|_2 = \|U \Sigma V^T\|_2 = \|U\|_2 \|\Sigma\|_2 \|V^T\|_2$$

$$= \|\Sigma\|_2$$

$$= \sigma_1$$

= the first singular value of A.

Therefore, $\|Av_1\| = \sigma_1$

$$\sigma_1^2 = \|Av_1\|^2 = \sum_{i=1}^{n} p_i^2 =$$ sum of squares of projections of $\{a_i\}$ onto the line of v_1.

$\boxed{\text{Claim}}$: v_1 is the first singular vector of A.

Proof. The defining property of v_1 is that it maximizes $\|Av\|$, for all v with $\|v\|=1$.

But,

$$\|Av\| = \|\sigma_1 u_1 v_1^T v + \sigma_2 u_2 v_2^T v + \dots + \sigma_r u_r v_r^T\|$$

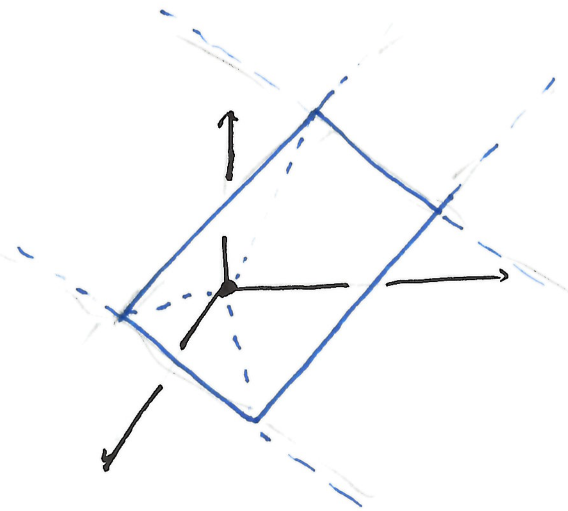
$$\leq |\sigma_1| |\langle v_1, v\rangle| \|u_1\| + \dots + |\sigma_r| |\langle v_r, v\rangle| \|u_r\|$$

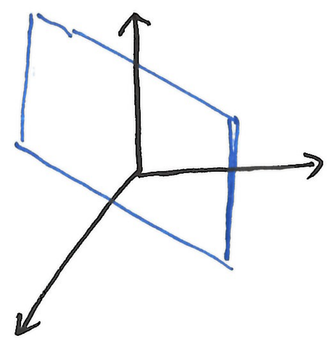
$\|Av\|$ is maximized when $v = v_1$ for which

$\|Av\| = \|Av_1\| = |\sigma_1| \|v_1\|^2 = \sigma_1$ since $\sigma_1 > 0$ and $\|v_1\| = 1 = \|u_1\|$

and $v_1 \perp v_i$ for $i \neq 1$.

One Step Further: Finding The Best Fit 2-dim Subspace

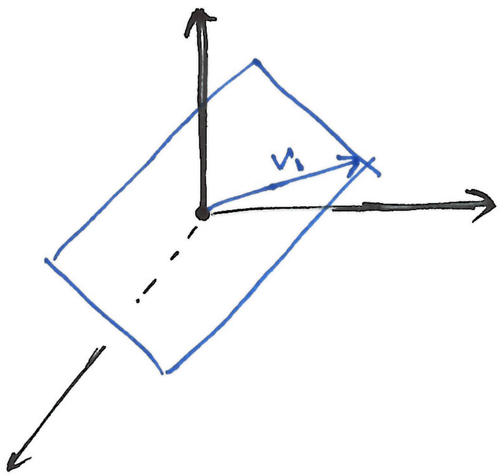




To find the best fit 2-dim subspace for the set of d-dim points $\{a_i\}_{i=1}^{n}$ we use the greedy approach.

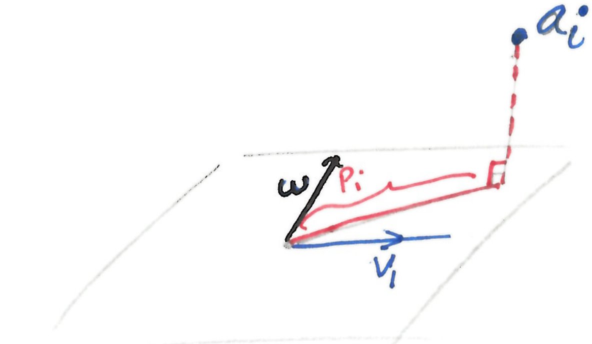
First, we find the best fit line (1-dim) generated by the vector v_1 found from the previous step, i.e. the first singular vector of $A_{n \times d}$.

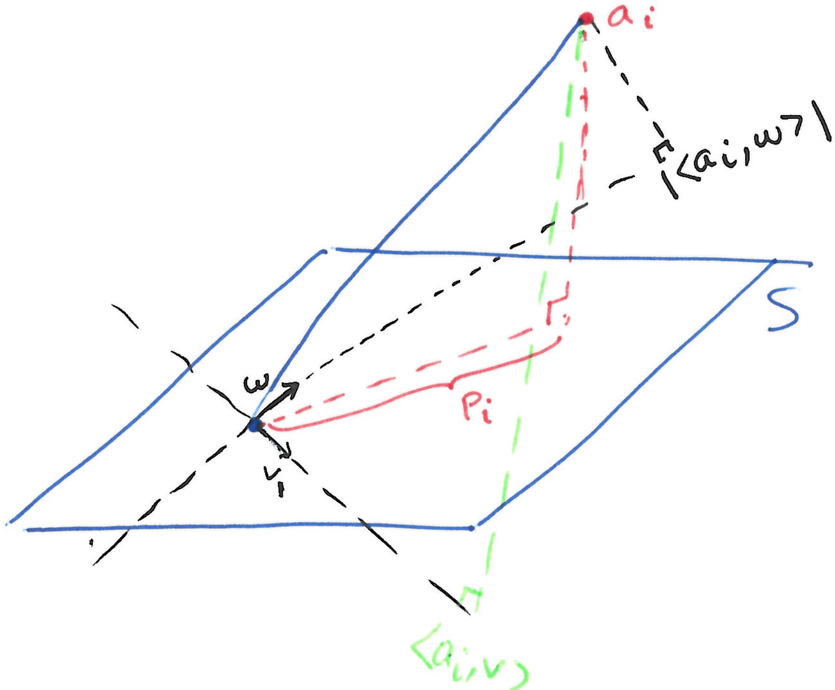
Next, we look for the best fit 2-dim subspace (a plane) containing v_1 (and of course the line spanned by v_1).



Suppose S is a subspace of $\mathbb{R}^d$ containing v_1. Consider the sum of squared lengths of the projections

$$\sum_{i=1}^{n} p_i^2 .$$





Suppose w is any unit vector (i.e. $\|w\| = 1$) in S perpendicular to v_1.

We claim that

$$p_i^2 = |\langle a_i, v_1\rangle|^2 + |\langle a_i, w\rangle|^2$$

Complete $\{v_1, w\}$ to an orthonormal basis $\{v_1, w, v_3', \ldots, v_d'\}$ for $\mathbb{R}^d$.

Write a_i as a linear combination of vectors in this basis.

$$a_i = a_{i1}' v_1 + a_{i2}' w + a_{i3}' v_3' + \dots + a_{id}' v_d'$$

The projection of the vector a_i onto the plane S is given by $a_{i1}' v_1 + a_{i2}' w$.

Therefore,

$$p_i^2 = \| a_{i1}' v_1 + a_{i2}' w \|^2 =$$

$$|a_{i1}'|^2 + |a_{i2}'|^2 =$$

$$|\langle a_i, v_1 \rangle|^2 + |\langle a_i, w \rangle|^2$$

So, we proved our claim.

Using this result, we have

$$\sum_{i=1}^{n} p_i^2 = \sum_{i=1}^{n} |\langle a_i, v_1 \rangle|^2 + \sum_{i=1}^{n} |\langle a_i, w \rangle|^2$$

To maximize $\sum_{i=1}^{n} p_i^2$, we need to choose the unit vector w such that $w \perp v_1$ and $\sum_{i=1}^{n} |\langle a_i, w \rangle|^2$ is maximized.

Again, we note that

$$\sum_{i=1}^{n} |\langle a_i, w \rangle|^2 = \| Aw \|_2^2$$

Therefore, we want to maximize $\|Aw\|$ for all choices of vector w subject to the following conditions:

(i) $\|w\| = 1$

(ii) $w \perp v_1$

We set

$$v_2 = \underset{w \perp v_1,\ \|w\|=1}{\arg\max} \|Aw\|$$

We show that $\|A v_2\|$ is the second singular value of A and v_2 is the second singular vector of A.

Suppose w is any vector in $\mathbb{R}^d$ which is perpendicular to v_1 and has norm 1.

$\boxed{r = \operatorname{rank}(A)}$

$$
\begin{aligned}
Aw &= U \Sigma V^T w \\
&= \sum_{i=1}^{r} \sigma_i \langle v_i, w \rangle u_i \\
&= \sum_{i=2}^{r} \sigma_i \langle v_i, w \rangle u_i
\end{aligned}
$$

since $\langle v_1, w \rangle = 0$.

Because $\sigma_2 \geq \sigma_3 \geq \cdots \geq \sigma_r > 0$, the 2-norm $\|Aw\|$ is maximized precisely when the distribution $\{ |\langle v_i, w \rangle| \}_{i=2}^{r}$ is concentrated at $i=2$, that is when $w = v_2$. In that case,

$$
\max_{\substack{w \perp v_1 \\ \|w\|=1}} \|Aw\| = \|Av_2\| = \left\| \sum_{i=2}^{r} \sigma_i \langle v_i, w \rangle u_i \right\|
$$

$$
= |\sigma_2 \langle v_2, v_2 \rangle| \, \|u_2\| = |\sigma_2| = \sigma_2
$$

Continuing this process, we have

$$v_1 = \underset{\|v\|=1}{\operatorname{argmax}} \|Av\| \quad , \quad \|Av_1\| = \sigma_1$$

$$v_2 = \underset{\substack{\|v\|=1 \\ v \perp v_1}}{\operatorname{argmax}} \|Av\| \quad , \quad \|Av_2\| = \sigma_2$$

$$v_3 = \underset{\|v\|=1}{\operatorname{argmax}} \|Av\| \quad , \quad \|Av_3\| = \sigma_3$$

$$\vdots$$

$$v_r = \underset{\substack{\|v\|=1 \\ v \perp \{v_1, \dots, v_{r-1}\}}}{\operatorname{argmax}} \|Av\| \quad , \quad \|Av_r\| = \sigma_r$$

$$0 = \underset{\substack{\|v\|=1 \\ v \perp \{v_1, \dots, v_r\}}}{\operatorname{argmax}} \|Av\| \quad , \quad 0 = \sigma_{r+1}$$

where $r = \operatorname{rank}(A)$.