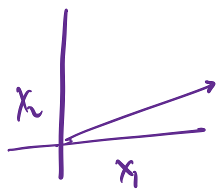
10/27/2021 : Finishing up PCA.

Recap PCA is for dimension reduction by getting rid of redundancy in the data.

Get a new set of variables that are linear combinations of the old; and will give us the same information or (almost as much information) with a much smaller set of variables



2 ways of getting the linear combinations from $x_1, \ldots x_p$

$z_i = a_1 x_i^{(1)} + a_2 x_i^{(2)} + \cdots + a_p x_i^{(p)}$

For every set of coefficients $a_1, \ldots a_p$,
get a new set of points $z_1, \ldots z_n$

Compute sample variance for these $\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2$

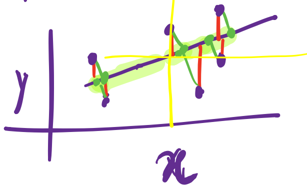
& find $\vec{a} = (a_1, \ldots a_p)$ that maximizes sample variance (find the a_i that spreads the po

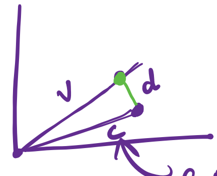
If we center the data,
then ~~the~~ maximizing sample variance is like maximizing dist from origin

---

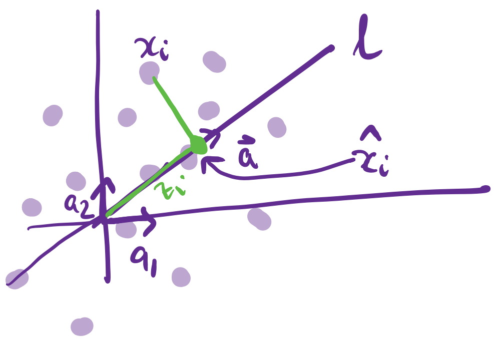
2nd method : Finding the best line that fits the data.

Reg

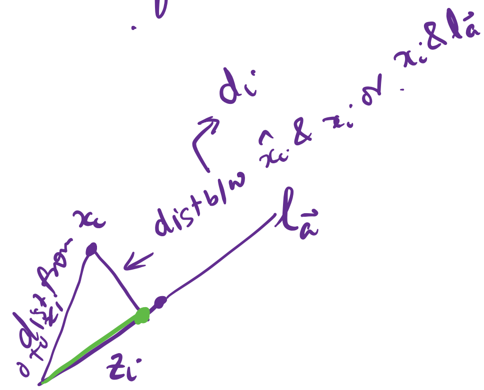




So maximizing v^2 & minimizing d^2 are equivalent, so our 2 methods of computing $\vec{a}$ will arrive at the same coefficients !!



$\hat{x}_i$ is the projection of x_i onto the line l.
$\vec{a}$ is a unit vector in the direction of l.



$p = 2$ (# of dimensions)

$$\hat{x}_i = \vec{a} \cdot \left(a_1 x_i^{(1)} + a_2 x_i^{(2)}\right)$$

magnitude of x_i

$$z_i^2 + d_i^2 = |x_i|^2$$

maximizing z_i^2 is equiv. to minimizing d_i^2

Moving on to Multivariate regression.