UNSUPERVISED LEARNING

└ REPRESENTATION LEARNING.

Goal: Given a set of "data points", "understand" something "useful" about them.

Data points → vectors in $\mathbb{R}^d$ $\quad \begin{bmatrix} \text{height} \\ \text{weight} \\ \text{age} \end{bmatrix} \in \mathbb{R}^3$

Running theme: "COMPREHENSION IS COMPRESSION" (George Chaitin)

↳ understanding

↳ Learning

Problem:

Input: $\{x_1, x_2, \dots, x_n\}$ $\quad x_i \in \mathbb{R}^d \leftarrow$ # of features

Output: Some "compressed" representation of the dataset

Example:

$$\left\{ \begin{array}{c} x_1 \\ \begin{bmatrix} -7 \\ -14 \end{bmatrix} \end{array}, \begin{array}{c} x_2 \\ \begin{bmatrix} 2.5 \\ 5 \end{bmatrix} \end{array}, \begin{array}{c} x_3 \\ \begin{bmatrix} 0.5 \\ 1 \end{bmatrix} \end{array}, \begin{array}{c} x_4 \\ \begin{bmatrix} 0 \\ 0 \end{bmatrix} \end{array} \right\}$$

Question: How many real numbers are needed to store this dataset $\boxed{8}$

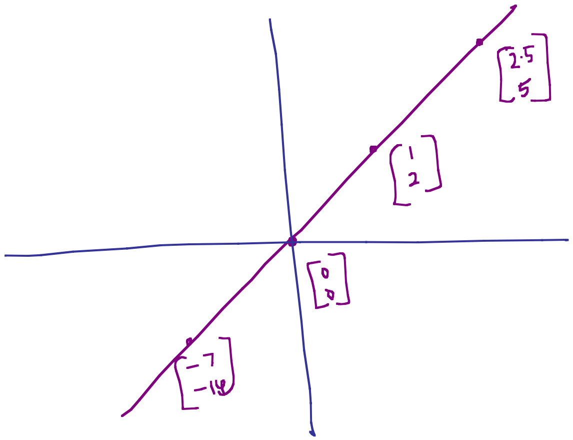
Representative

$\begin{bmatrix} 1 \\ 2 \end{bmatrix} \in \mathbb{R}^2$

Co-efficients

$\{-7, 2.5, 0.5, 0\}$ $\boxed{6}$

NOTE: using representative & co-effecients can "RECONSTRUCT" the dataset exactly.



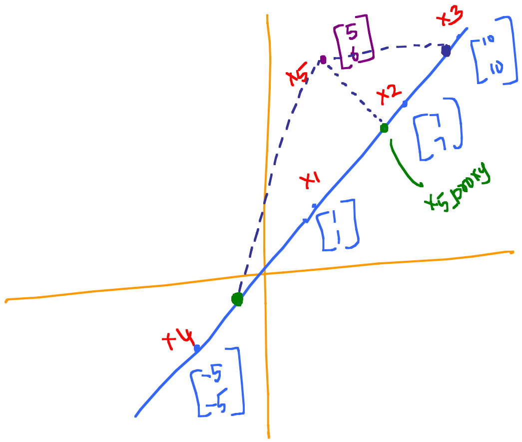
Repr	Co-eff
$\begin{bmatrix}1\\2\end{bmatrix}$	$\{-7, 2.5, 0.5, 0\}$
$\begin{bmatrix}1/\sqrt{5}\\2/\sqrt{5}\end{bmatrix}$	$\{-7\sqrt{5}, 2.5\sqrt{5}, 0.5\sqrt{5}, 0\}$

NOTE:

Any vector along the purple line can be chosen as a representative [except $\begin{bmatrix}0\\0\end{bmatrix}$]

Original: # real numbers $= dn$

real numbers in compressed representation $= d + n$



Rep

$\left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$

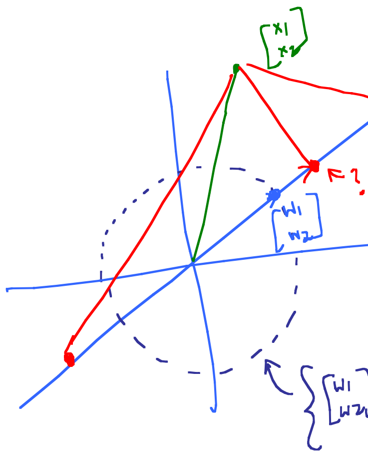
Co-eff

$\{ (-5,-5), (1,1), (7,7), (5,6), (10,10) \}$

$x_3 = 10 \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 10 \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 10 \\ 10 \end{bmatrix}$

Q: Who can "pretend" to be a proxy for x_5 along the blue line?

Ans: Projection of x_5 onto the blue line.



$\left\{ \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} : w_1^2 + w_2^2 = 1 \right\}$

$$c \cdot \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \left(\frac{x_1 w_1 + x_2 w_2}{w_1^2 + w_2^2} \right) \cdot \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

$$= \left(\frac{x^T w}{\|w\|^2} \right) \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

$\min_{c}$ length^2(error vector)

↪ $\begin{bmatrix} x_1 - c w_1 \\ x_2 - c w_2 \end{bmatrix}$

$$\min_{c} \ (x_1 - c w_1)^2 + (x_2 - c w_2)^2 .$$

$$c^* = \left(\frac{x_1 w_1 + x_2 w_2}{w_1^2 + w_2^2} \right)$$

(scalar)

Inner product/ dot product of x and w.

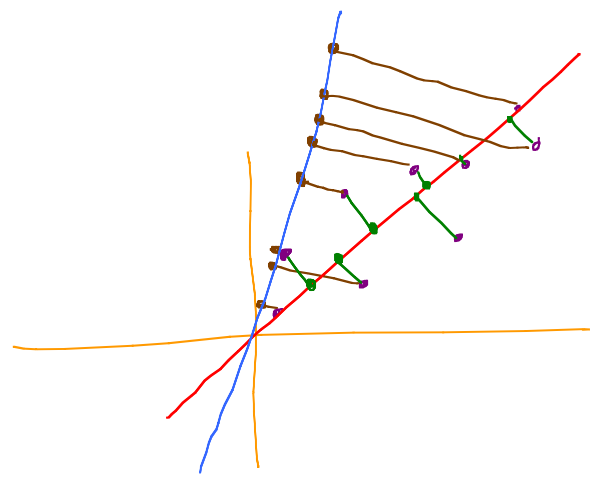
length$^2 \left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \right)$ ←

NOTE: Can always pick $\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ s.t length$\left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \right) = 1$

$$\Rightarrow \quad c^* = (x^T w) \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

Goal: Develop a way to find a "compressed" representation of data when data points not-necessarily fall on line

↳ RULE, NOT EXCEPTION



Goal: Find the line that has the least "reconstruction" error.

$\|z\|^2 = z^T z$

Dataset: $\{x_1, x_2, \ldots, x_n\}$ $\quad x_i \in \mathbb{R}^d$

$$\text{ERROR}(\text{line}, \text{dataset}) = \sum_{i=1}^{n} \text{error}(\text{line}, x_i)$$

$$= \sum_{i=1}^{n} \text{length}^2\left(x_i - (x_i^T w) w\right)$$

$$= \sum_{i=1}^{n} \| x_i - (x_i^T w) \cdot w \|^2$$

Represented using w s.t $\|w\|^2 = 1$

$$f(w) = \frac{1}{n} \sum_{i=1}^{n} \| \underbrace{\overset{\in \mathbb{R}^d}{x_i} - \overset{\in \mathbb{R}}{(x_i^T w)} \cdot \overset{\in \mathbb{R}^d}{w}} \|^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left(x_i - (x_i^T w) \cdot w\right)^T \left(x_i - (x_i^T w) \cdot w\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[x_i^T x_i - \cancel{(x_i^T w)^2} - (x_i^T w)^2 + \cancel{(x_i^T w)^2 \cdot 1} \right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\underbrace{x_i^T x_i} - (x_i^T w)^2 \right)$$

$$\min_{w:\ \|w\|^2=1} g(w) = \frac{1}{n}\sum_{i=1}^{n} -(x_i^T w)^2$$

$$\max_{w:\ \|w\|^2=1} \frac{1}{n}\sum_{i=1}^{n} (x_i^T w)^2 = \frac{1}{n}\sum_{i=1}^{n} (\underset{1\times d}{w^T}\ \underset{d\times 1}{x_i})(\underset{1\times d}{x_i^T}\ \underset{d\times 1}{w})$$

$$= \frac{1}{n}\sum_{i=1}^{n} \underline{w^T}(x_i x_i^T)\underline{w}$$

$$= w^T \Big(\underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T}_{C} \Big) w$$

C $d\times d$ matrix

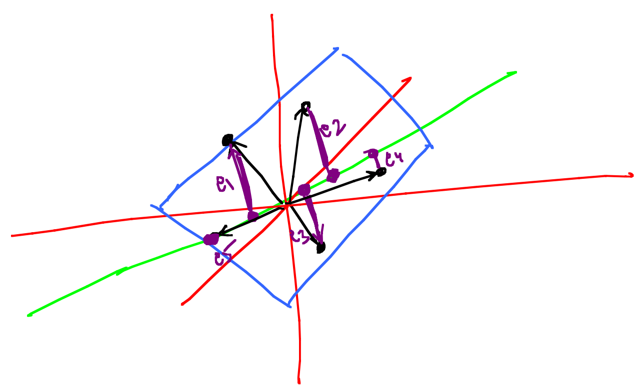
Equivalently

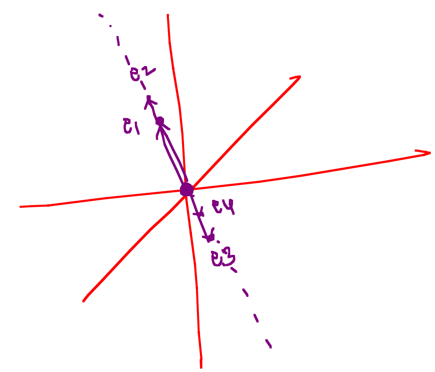
$$\max_{w:\ \|w\|^2 = 1} \quad w^T C w$$

$$\boxed{C = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T}$$

↳ Co-variance matrix

Soln: w is the eigenvector corresponding to the maximum eigen value of C





$x \in \mathbb{R}^d$

$\downarrow$ Find w

$(x^T w) \cdot w$

$\downarrow$ Residue / error

$\underbrace{x - (x^T w) \cdot w}$

Might not be error but has "information"

POSSIBLE ALGORITHM

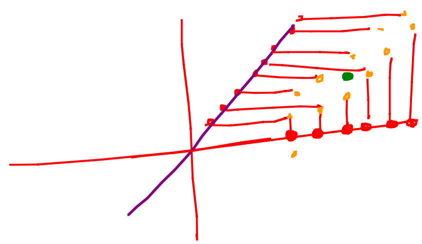
Input: $\{x_1, \ldots, x_n\}$ $\quad x_i \in \mathbb{R}^d$

$\rightarrow$ $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\qquad x_i = x_i - \mu \quad \forall i$

$\rightarrow$ Find "best" line $w_1 \in \mathbb{R}^d$

$\rightarrow$ Replace $x_i \leftarrow x_i - (x_i^T w) w$

$\rightarrow$ Repeat to obtain w_2



Issue: Data may not be centered.

Questions

→ How to solve $\max_{\substack{w \\ \|w\|^2=1}} w^T C w$?

→ How many times to repeat the procedure?

→ Where exactly is "compression" is happening?

→ What "representations" are we learning?

Note Title 05-08-2022

$$D = \{x_1, x_2, \dots, x_n\} \quad x_i \in \mathbb{R}^d .$$

$$W_1 = \underset{w:\ \|w\|^2 = 1}{\operatorname{argmax}} \; w^T C w \qquad C = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$$

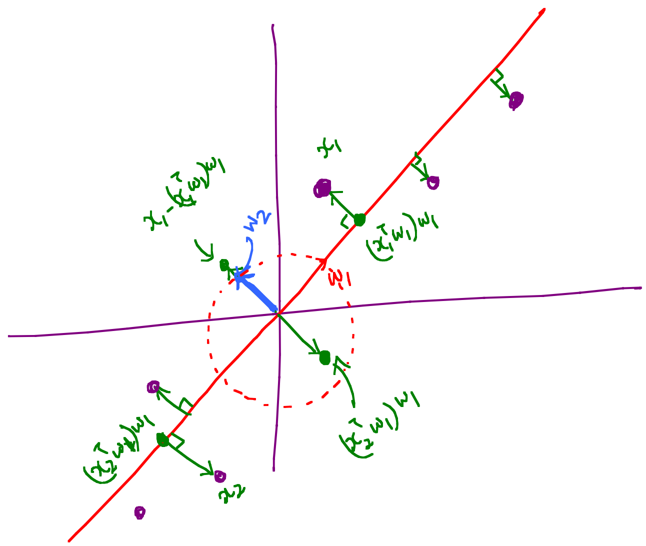
$$\boxed{\begin{array}{c} x_1' \\ x_2' \\ \vdots \\ x_n' \end{array}} \begin{array}{l} \rightarrow \; x_1 - (x_1^T w_1) w_1 \\ \rightarrow \; x_2 - (x_2^T w_1) w_1 \\ \quad \vdots \\ \rightarrow \; x_n - (x_n^T w_1) w_1 \end{array}$$

$$W_2 = \underset{w:\ \|w\|_2^2 = 1}{\operatorname{argmax}} \; w^T C' w$$

$$C' = \frac{1}{n}\sum_{i=1}^{n} x_i' x_i'^T$$

Note Title 05-08-2022

Question: What can we say about w_1 & w_2?



Observation:

→ All residues are orthogonal to w_1

→ Any line which minimizes sum of errors w.r.t residues must also be orthogonal to w_1 [ARGUE WHY]

$$\Rightarrow \boxed{w_2^T w_1 = 0}$$

Note Title 05-08-2022

By continuing this procedure, we get $\underbrace{\{w_1, w_2, \ldots, w_d\}}_{\text{ORTHONORMAL VECTORS}}$ s.t $\|w_k\|_2^2 = 1 \ \forall k$ and $w_i^T w_j = 0 \ \forall i \neq j$

Residue after round 1

$$\underbrace{\{x_1 - (x_1^T w_1) w_1, \ \ldots, \ x_n - (x_n^T w_1) w_1\}}$$

all vectors in $\in \mathbb{R}^d$

- $w_2 \rightarrow$ Best line that fits residues.
- $w_1^T w_2 = 0$

Note Title 05-08-2022

Residues after round 2

$$\left\{ \underbrace{x_1 - (x_1^T w_1) w_1 - \left(\left(x_1 - (x_1^T w_1) w_1 \right)^T w_2 \right) w_2}, \quad \ldots \right\}$$

$$= \left\{ x_1 - (x_1^T w_1) w_1 - \left(x_1^T w_2 - (x_1^T w_1) \cdot \underbrace{w_1^T w_2}_{=0} \right) w_2, \quad \ldots \right\}$$

$$= \left\{ \underbrace{x_1 - (x_1^T w_1) w_1 - (x_1^T w_2) w_2}, \quad \ldots \right\}$$

Residues after d-rounds.

$$\forall i \qquad x_i - \left((x_i^T w_1) w_1 + (x_i^T w_2) w_2 + \cdots + (x_i^T w_d) w_d \right) = \vec{0} \in \mathbb{R}^d.$$

Note Title 05-08-2022

$$\forall i \quad x_i = (x_i^T w_1) w_1 + (x_i^T w_2) w_2 + \cdots + (x_i^T w_d) w_d.$$

What have we gained?

- If data lives in a "low" dimensional linear sub-space, then residues become 0 much earlier than d rounds.

Example: Say Dataset is such that after 3-rounds, residues become 0.

$$\text{Dataset} = \{x_1, \ldots, x_n\} \quad x_i \in \mathbb{R}^{100}.$$

Note Title 05-08-2022

$\forall i \quad x_i = (x_i^T w_1) w_1 + (x_i^T w_2) w_2 + (x_i^T w_3) w_3.$

$\{w_1, w_2, w_3\} \in \mathbb{R}^{100}$

Rep

$\{w_1, w_2, w_3\}$

↳ common for dataset

Co-efficients

$x_i \rightarrow [x_i^T w_1 \quad x_i^T w_2 \quad x_i^T w_3] \in \mathbb{R}^3$

↳ Datapoint specific

original: $100 \times n$

$100 \times 100 = 10000$

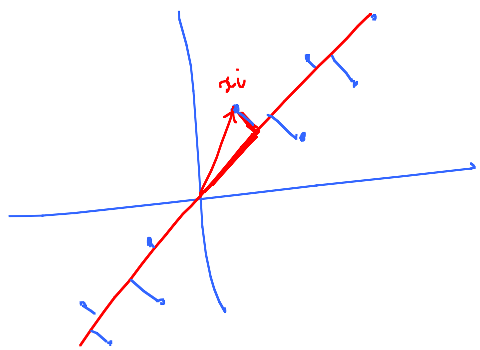
$\boxed{d \times n}$

now: $3 \times 100 + 3n$

$3 \times 100 + 3 \times 100 = 600$

$\boxed{d \times k + k \times n}$

Question: What if data "approximately" lies in a low-dimensional space?



PYTHAGOROUS

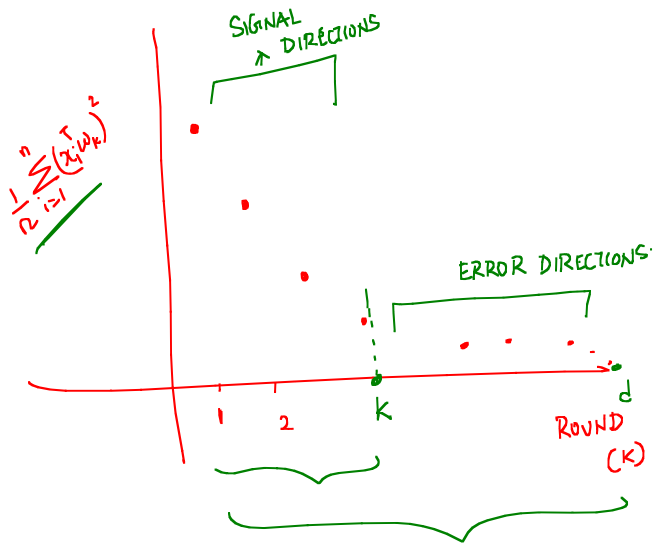
For any $w \in \mathbb{R}^d$, s.t $\|w\|_2^2 = 1$

$$\forall i \quad \|x_i\|^2 = \|x_i - (x_i^T w)\, w\|^2 + \|(x_i^T w)\, w\|^2$$

$$\frac{1}{n}\sum_{i=1}^{n} \|x_i\|^2 = \frac{1}{n}\sum_{i=1}^{n} \|x_i - (x_i^T w)\, w\|^2 + \frac{1}{n}\sum_{i=1}^{n} \|(x_i^T w)\, w\|^2$$

As large as possible.

"Larger the value of $\frac{1}{n}\sum_{i=1}^{n} (x_i^T w)^2$, the better the fit"



ENTER LINEAR ALGEBRA

$$\max_{w:\ \|w\|_2^2=1} w^T C w \qquad C = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$$

$C \rightarrow$ Covariance matrix.

Soln: w_1 is eigenvector corresponding to the "largest" eigenvalue of C [HILBERT's min-max theorem]

In fact $\{w_1, \ldots, w_d\}$, the eigenvectors of C form an orthonormal basis.

$\boxed{w_k \rightarrow \text{Best line one can obtain in round } k}$

What do eigenvalues of C mean?

We know

$$C w_1 = \lambda_1 w_1$$

$$w_1^T C w_1 = w_1^T(\lambda_1 w_1) = \lambda_1$$

$$\lambda_1 = w_1^T C w_1 = w_1^T \left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \right) w_1$$

$$\boxed{\lambda_1 = \frac{1}{n} \sum_{i=1}^{n} (x_i^T w_1)^2}$$

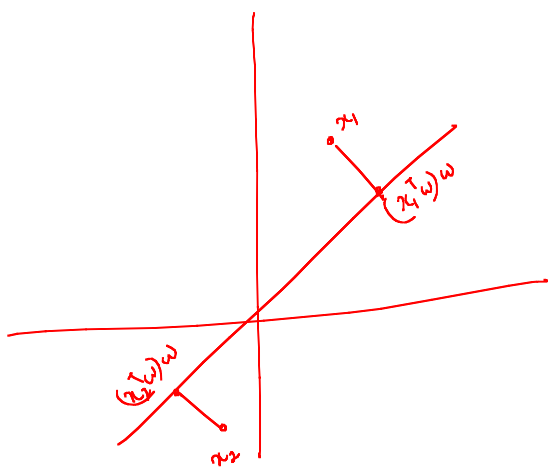
term we used earlier

Rule of thumb for # dimensions

$$\left(\sum_{i=1}^{k} \lambda_i(C) \Big/ \sum_{i=1}^{d} \lambda_i(C) \right) \geq 0.95$$

usually in practice

Note Title 05-08-2022



$$\{(x_1^T w), \; \ldots, (x_n^T w)\}$$

Average?

$$\frac{1}{n}\sum_{i=1}^{n}(x_i^T w) = \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^T}_{0^T w} w$$

[for a centered dataset].

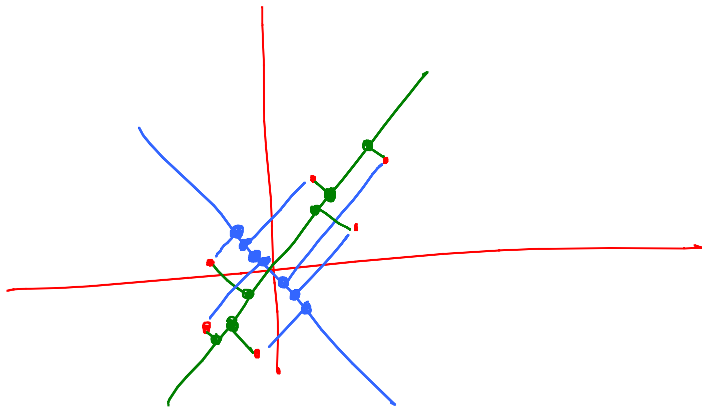
$$= 0$$

Variance

$$\frac{1}{n}\sum_{i=1}^{n}(x_i^T w - \underbrace{\text{mean}}_{=0})^2 = \boxed{\frac{1}{n}\sum_{i=1}^{n}(x_i^T w)^2}$$

ERROR MINIMIZATION on CENTERED DATASET $\iff$ VARIANCE MAXIMIZATION

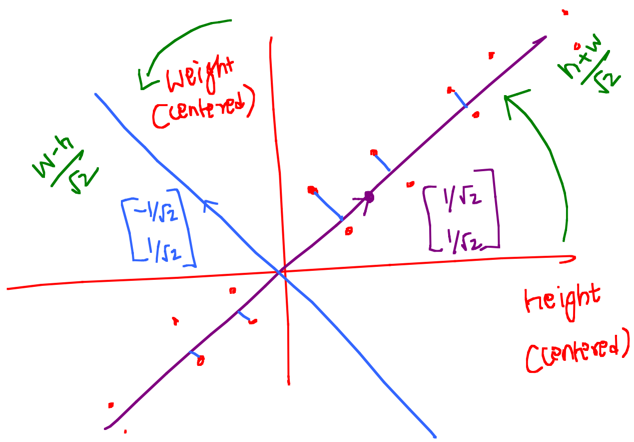
Note Title 05-08-2022

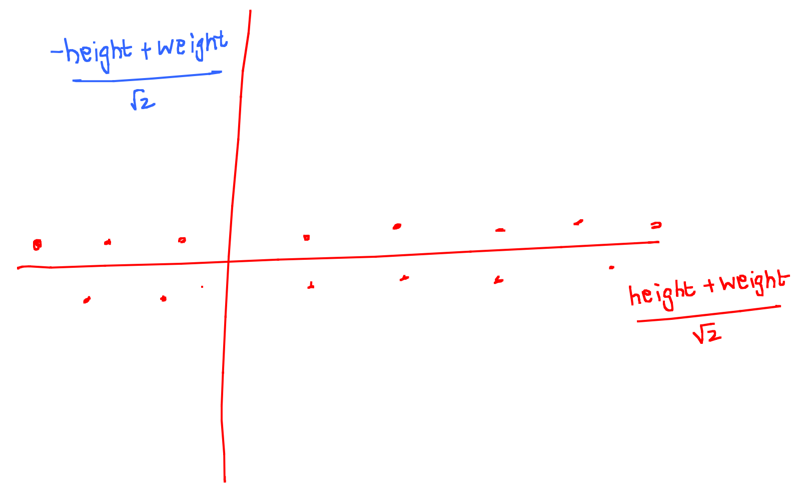


Want directions where Projections don't "crowd-up"

i.e., Variance is not small.

ONE MORE EXAMPLE





height gives some information about weight.

$(h+w)/\sqrt{2}$ does not give any information about $\frac{w-h}{\sqrt{2}}$.

"de-correlated"

Note Title 05-08-2022

PRINCIPAL COMPONENT ANALYSIS - $\{ \underbrace{w_1, \ldots, w_k}_{\text{PRINCIPAL COMPONENT}} \}$

"DIMENSIONALITY REDUCTION"

PCA Finds combination of features that are de-correlated. [loosely speaking independent of each other].

"EIGENFACES"