PRINCIPLE COMPONENT ANALYSIS (PCA)

Motivation example:

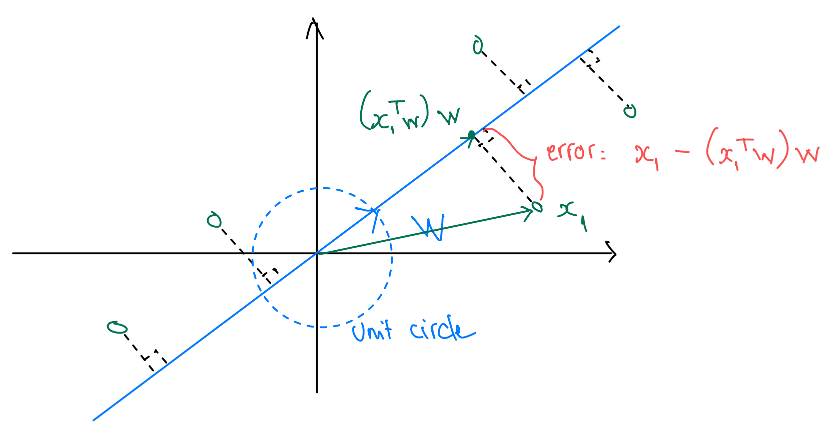
- PCA is common technique used in unsupervised learning (representative learning) for dimensions reduction.

What is unsupervised learning?

Unlike supervised learning, the dataset used in unsupervised learning doesn't have labels.

- PCA used to identify the principal vectors or hyperplanes that capture the most variance in the dataset. These vectors/hyperplanes can be used to reconstruct the dataset with the least amount of errors.

For example:



Goal: Find vector w of unit length that minimize the errors:

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

Solution: Vector w is the eigenvector correspond to the maximum eigenvalue of C

Question: Why vector w that minimize errors is also the vector that capture the most variance in the dataset?

Given: $$f(x) = \min_{\|w\|=1} \sum_{i=1}^{m} \| x_i - (x_i^T w) w \|^2$$

Some w minimize the sum, also minimize the average, so:

$$f'(x) = \min_{\|w\|=1} \frac{1}{m} \sum_{i=1}^{m} \| x_i - (x_i^T w) w \|^2$$

$$= \min_{\|w\|=1} \frac{1}{m} \sum_{i=1}^{m} \left(x_i^T x_i - (x_i^T w)^2 \right)$$

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

< minize w.r.t. w, remove unrelated term $x_i^T x_i$ >

This is the same as:

$$g(x) = \max_{\|w\|=1} \frac{1}{m} \sum_{i=1}^{m} (x_i^T w)^2$$

$$= \max_{\|w\|=1} \frac{1}{m} \sum_{i=1}^{m} (x_i^T w)^T (x_i^T w)$$

$$= \max_{\|w\|=1} w^T \underbrace{\left(\frac{1}{m} \sum_{i=1}^{m} x_i x_i^T \right)}_{\text{covariance matrix } C} w$$

So we conclude that:

$$\min_{\|w\|=1} \sum_{i=1}^{m} \| x_i - (x_i^T w) w \| \quad \text{equivalent} \quad \max_{\|w\|=1} w^T C w$$

w that minimize errors

where $C = \frac{1}{m} \sum_{i=1}^{m} x_i x_i^T$

w that maximize the bilinear form of covariance matrix C

This leads to the solution explanation:

Recall eigenvector / eigenvalue is defined as:

$$C w = \lambda w$$

$$\Leftrightarrow \quad w^T C w = \lambda$$

$$\Leftrightarrow \quad \max_{\|w\|=1} w^T C w = \max_{\|w\|=1} \lambda$$

Hence, w is the eigenvector corresponding to the maximize eigenvalue of C.

How does covariance matrix formed?

We know: $C = \frac{1}{m} \sum^{m} x_i x_i^T$

$$= \frac{1}{m} \left(x_1 \,|\, \dots \,|\, x_m \right) \begin{pmatrix} x_1^T \\ \vdots \\ x_m^T \end{pmatrix}$$

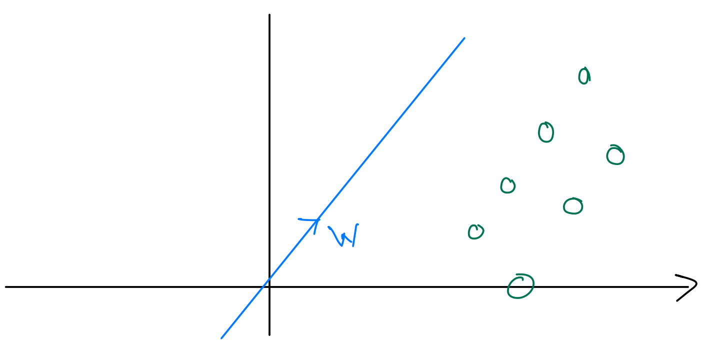
$$= \frac{1}{m} X^T X$$

What if our training set has 2 features of length, one is measured in feets, the other is measured in centimeters?

You should normalize that shit, the formula is:

$$d_j := d_j - \sigma_j \quad \forall j \in [d], \quad \sigma_j = \sqrt{\frac{1}{m} \sum^{m} x_{ij}^2}$$

So far, we know how to find "best" line w that go through $(0,0)$, What if our data is not centered?

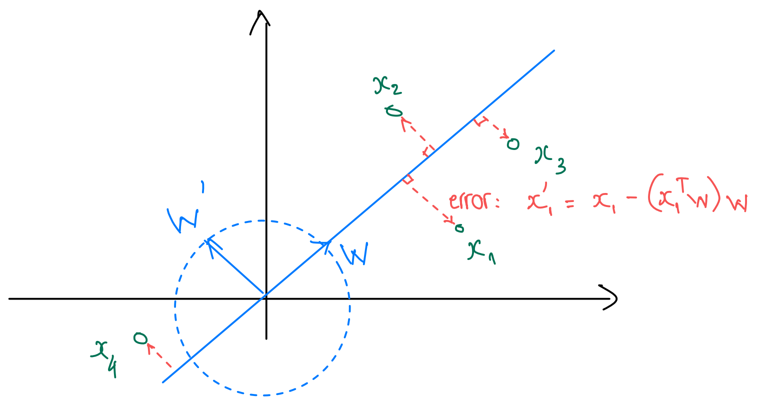


Solution: Before finding w, "center" all the data points:

- Find mean: $\mu = \frac{1}{m} \sum_{i=1}^{m} x_i$
- Subtract mean from data:

$$x_i = x_i - \mu \quad \forall i$$

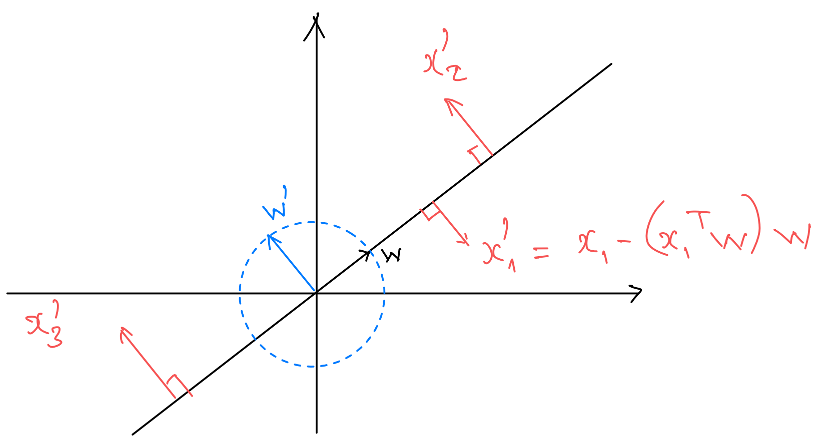
Now that we found w that can reconstruct the dataset with some errors, What if there is some information inside those errors that we also want to compress ?



<u>Solution:</u> Find another w that can reconstruct those errors :

- Update: $S' = \{x_1', \dots x_m'\}$ where $x_i' = x_i - (x_i^T w) w$
- Find $w' = \underset{\|w\|=1}{\arg\max}\ w^T C' w$

Great, now we can find 2 vectors w and w. What is the relationship between them ?



<u>Solution:</u>

Since w' is the line "best" describe the errors, and all errors are perpendicular to w.

$\Rightarrow \quad w' \perp w$

$\Rightarrow \quad w'^T w = 0$

for any
$$\begin{cases} w' = \underset{\|w\|=1}{\arg\max}\ w^T C' w \\ w = \underset{\|w\|=1}{\arg\max}\ w^T C w \\ C = \frac{1}{m} \sum_{i=1}^{m} x_i x_i^T, \quad C' = \frac{1}{m} \sum^{m} x_i' x_i'^T \\ x_i' = x_i - (x_i^T w) w \end{cases}$$

w' is the second relavant eigenvector

Now that we know how to iteratively find w that "compressed" the dataset, when should we stop iterating ?

- The most intuitive answer is when the residuals / errors equals to 0. Theorectically, this is always the case after d rounds:

$$x_i - \left[(x_i^T w_1) w_1 + \dots + (x_i^T w_d) w_d \right] = \vec{0} \in \mathbb{R}^d, \ \forall i \in [m]$$

$$\Rightarrow \quad x_i = (x_i^T w_1)\underline{w_1} + \dots + (x_i^T w_d)\underline{w_d} \quad , \forall i \in [m]$$

Since $w_1 \perp \dots \perp w_d$, we are essentially changing the basis of x_i

Why change the basis reduced dimension ?

- Informally, the standard basis vectors are not always good at expressing the dataset. Some basis vector may contain more noise than information, some may be redundant.
- Formally, if the data exists in a low-dimensional space, then residues become 0 much earlier than d rounds.

What if the data "approximately" in a low-dimensional space ?

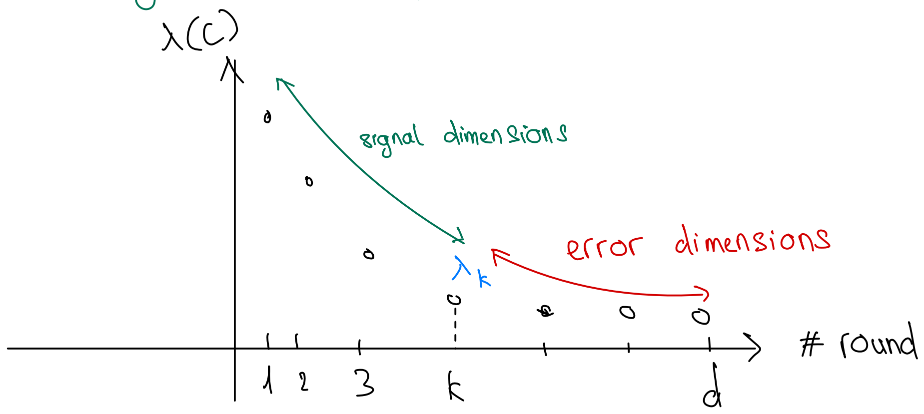
Consider this, using pythagorous, we can say that:

$$\| x_i \|^2 = \| x_i - (x_i^T w) w \|^2 + \| (x_i^T w) w \|^2$$

$$\Leftrightarrow \frac{1}{m}\sum^m \| x_i \|^2 = \underbrace{\frac{1}{m}\sum^m \| x_i - (x_i^T w) w \|^2}_{\text{avg residual / error, as small as possible}} + \underbrace{\frac{1}{m}\sum^m \| (x_i^T w) w \|^2}_{\text{avg projection } x_i \text{ onto } w\text{, as large as possible}}$$

We can prove that eigenvalue: $\lambda(C) = \frac{1}{m}\sum^m \| (x_i^T w) w \|^2$, C is cov matrix

" For each round, the larger $\lambda(C) = \frac{1}{m}\sum^m \| (x_i^T w) w \|^2$, the better the fit "



stop iterating here

Rule of thump for # dimensions

We can say after compression, we retain 95% of information if:

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

signal dimensions

total dimensions

PCA algorithm

Given $S = \{x_1, \dots, x_m\}$ $x_i \in \mathbb{R}^d$, "compressed" S.

Let $X \in \mathbb{R}^{m \times d}$ be the corresponding dataset matrix. $X = \begin{pmatrix} x_1^T \\ \vdots \\ x_m^T \end{pmatrix}$

1. Center X:

- $u = \frac{1}{m} \sum_{i=1}^{m} x_i \quad \forall i \in [m]$ ($u \in \mathbb{R}^d$) mean vector
- $x_i := x_i - u \quad \forall i \in [m]$

or $X := X - \mathbb{1} \cdot u^T$ — vector of ones size $m \times 1$

2. Normalize X:

- $\sigma_j = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (x_{ij} - \mu_j)^2} \quad \forall j \in [d]$ standard deviation feature j
- $x_{ij} := x_{ij} / \sigma_j \quad \forall \begin{matrix} i \in [m] \\ j \in [d] \end{matrix}$

Step 1 and 2 can be thought of as Standardization:

$\forall \begin{matrix} i \in [m] \\ j \in [d] \end{matrix}$, $x_{ij} := \dfrac{x_{ij} - \mu_j}{\sigma_j}$, where $\begin{cases} \mu_j = \frac{1}{m} \sum_{i=1}^{m} x_{ij} \\ \sigma_j = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (x_{ij} - \mu_j)^2} \end{cases}$

3. Construct covariance matrix C:

- $C = \frac{1}{m} \sum_{i=1}^{m} x_i x_i^T \quad \forall i \in [m]$ $\quad C \in \mathbb{R}^{d \times d}$

or $C = \frac{1}{m} X^T X$

4. Eigendecompose covariance matrix C:

- $C = \underbrace{Q}_{\text{eigenvectors}} \cdot \underbrace{D}_{\text{eigenvalues}} \cdot Q^T$

$= (\underbrace{Q_L}_{\text{first } k \text{ eigenvectors}} \mid Q_R) \cdot \begin{pmatrix} D_{TL} & \\ & D_{BR} \end{pmatrix} \cdot \begin{pmatrix} Q_L^T \\ Q_R^T \end{pmatrix}$

5. Get the principal components:

The principal components matrix is: $X \cdot Q_L$

If you want to reconstruct the dataset using principle components, then:

$$X := (X Q_L) \cdot Q_L^T$$

Interpret the eigenvector chosen

Given any vector, covariance matrix will rotate it to roughly the same direction of the chosen eigenvector.