PCA

* Introduction
* Computing the PCA
* Interpretations of PCA

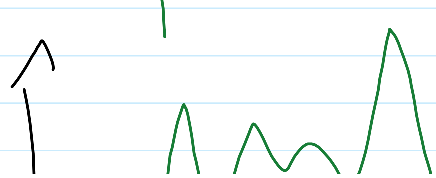
I. Introduction

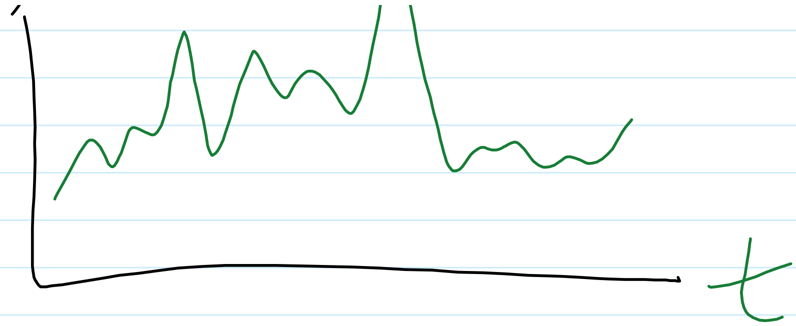
Problem: measured a bunch of data → how to analyze?

EX: Quantitative Finance

	stock1	stock2
t_0	\$	\$
t_1	\$	\$
⋮		

stock price





Which stocks best represent the movement of the market?

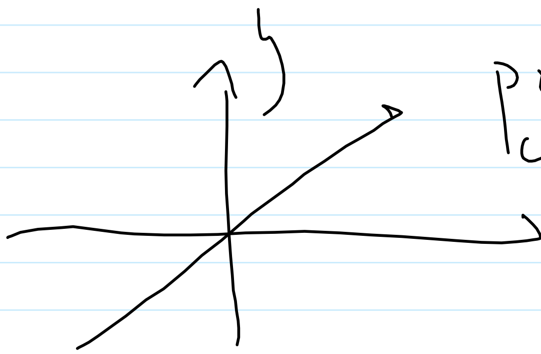
- Not necessarily one stock; could be a combination of stocks, or perhaps a weighted combination
- Moreover, some stocks might be entangled w/ each other

Ex: TSMC makes chips for Apple, but if Apple doing bad, then TSMC :(

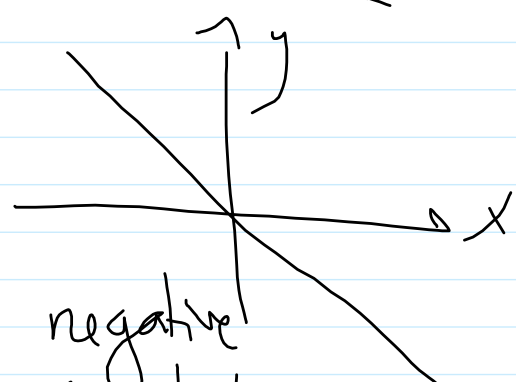
$\Longrightarrow$ PCA: lets us turn a bunch of measurements of various attributes (potentially correlated) into uncorrelated attributes!

⟶ principal components

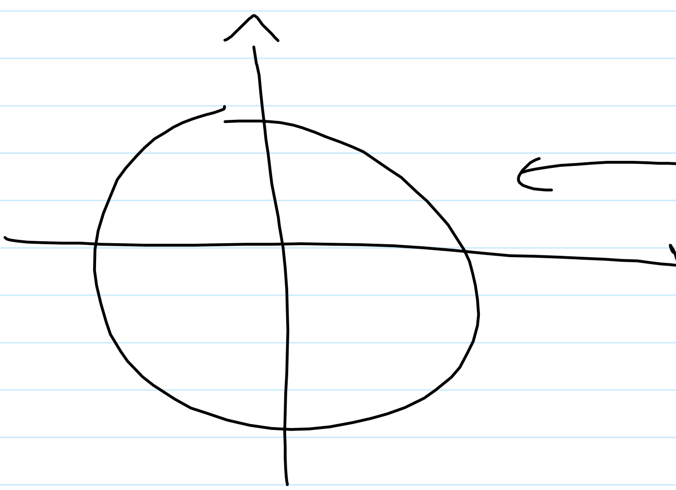
Correlation — tells us how variables move w/ each other <u>linearly</u>



positive correlation



negative correlation



⟵ zero correlation (even though technically these variables are related)

⟹ Use PCA for <u>feature extraction</u>

* From SVD to PCA

- Mathematically, SVD and PCA are very similar
- Dis 4C, Q4 — we will show that if you have done one, you have essentially done the other
- Conceptually, there are some differences

SVD: $A = U \Sigma V^*$

"more general"
- decomposing a matrix
  $\Rightarrow$ helps us understand the action of A when acting on some vector
- abstract meaning: tells us about Null(A), Col(A), etc

– abstract SVD: HW3 Q6
(derivative operator)

PCA: concerned w/ the matrix itself, which contains data points

$\Longrightarrow$ skips straight to the data analysis/feature extraction

II. Computing the PCA

$$\text{data matrix } X = \begin{array}{cc} \text{features} & \\ \begin{bmatrix} \vdash \vec{x}_1^{\,T} \dashv \\ \vdash \vec{x}_2^{\,T} \dashv \\ \vdots \\ \vdash x_m^T \dashv \end{bmatrix} & \text{measurements} \end{array}$$

Each row is a measurement (m measurements)

Each column is an attribute (n attributes)

De-meaned/ $\tilde{X}$ $\begin{bmatrix} \vdash (\vec{x}_1 - \bar{x})^T \dashv \end{bmatrix}$

De-meaned/ Centered Data matrix $\tilde{X} = \begin{bmatrix} \vdash (\bar{x}_1 - \bar{x})^T \dashv \\ \vdash (\bar{x}_2 - \bar{x})^T \dashv \\ \vdots \\ \vdash (\bar{x}_m - \bar{x})^T \dashv \end{bmatrix}$

$$\bar{x}^T = [\underline{\mu_1} \quad \underline{\underline{\mu_2}} \quad \cdots \quad \underline{\underline{\underline{\mu_n}}}]$$

$$\mu_1 = \frac{1}{m} \sum_{j=1}^{m} x_{j1}$$

(Take avg of all values in a column, i.e. for a particular attribute)

Also written as: $\bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i$

$$\boxed{\tilde{X} = X - \frac{1}{m} \vec{\mathbb{1}} \vec{\mathbb{1}}^T X}$$

vector of all 1's

Covariance Matrix

$$\boxed{S = \frac{1}{m} \tilde{X}^T \tilde{X} \quad \text{or} \quad \frac{1}{m-1} \tilde{X}^T \tilde{X}}$$

$$\boxed{S = \frac{1}{m} X^T X \quad \text{or} \quad \frac{1}{m-1} X^T X}$$

population variance

sample variance

* We will accept either m or $m-1$ on an exam

* In many applications, $m \gg 1$ so $m \approx m-1$

Extra

Note: Why is the covariance matrix defined this way?

Extra

Let's consider two attributes "x" and "y" for now:

$$X = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_m & y_m \end{bmatrix}$$

To clarify what is happening, I will

Write X in terms of column vectors:

$$X = \begin{bmatrix} | & | \\ \vec{x} & \vec{y} \\ | & | \end{bmatrix}$$

$$\tilde{X} = \begin{bmatrix} | & | \\ \vec{x} - \bar{\vec{x}} & \vec{y} - \bar{\vec{y}} \\ | & | \end{bmatrix} \qquad \bar{\vec{y}} = \bar{y}\vec{\mathbb{1}}$$

$$\frac{1}{m}\tilde{X}^T\tilde{X} = \frac{1}{m}\begin{bmatrix} \text{---}(\vec{x} - \bar{\vec{x}})^T\text{---} \\ \text{---}(\vec{y} - \bar{\vec{y}})^T\text{---} \end{bmatrix}\begin{bmatrix} | & | \\ \vec{x} - \bar{\vec{x}} & \vec{y} - \bar{\vec{y}} \\ | & | \end{bmatrix}$$

$$= \frac{1}{m}\begin{bmatrix} (\vec{x} - \bar{\vec{x}})^T(\vec{x} - \bar{\vec{x}}) & (\vec{x} - \bar{\vec{x}})^T(\vec{y} - \bar{\vec{y}}) \\ (\vec{y} - \bar{\vec{y}})^T(\vec{x} - \bar{\vec{x}}) & (\vec{y} - \bar{\vec{y}})^T(\vec{y} - \bar{\vec{y}}) \end{bmatrix}$$

Variance is a measure of "spread", i.e., deviation from the mean

m

$$\mathrm{Var}\,X = \sigma_X^2 = \frac{1}{m}\sum_{i=1} (x_i - \bar{x})^2$$

deviation of x_i from the average $\bar{x}$

This equals

$$(\vec{x} - \vec{\bar{x}})^T(\vec{x} - \vec{\bar{x}})$$

$$S = \frac{1}{m}\tilde{X}^T\tilde{X} = \begin{bmatrix} \mathrm{Var}\,x & \mathrm{Cov}(x,y) \\ \mathrm{Cov}(y,x) & \mathrm{Var}\,y \end{bmatrix}$$

$\rightarrow$ diagonal entries of S are variances

$\rightarrow$ off diag of S are covariances

Extra   Extra

Procedure for PCA:

1. Calculate covariance matrix

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

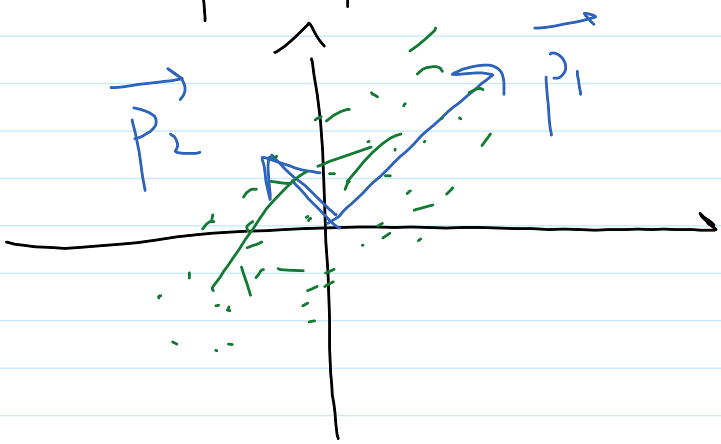
(remember to center your data!)

$X = X - X$

② Diagonalize $S = P \Lambda P^T$

(Guaranteed by spectral theorem)

- eigenvectors $\vec{p}_1$ " $\vec{p}_n$ are the principle components
- eigenvals $\lambda_1 \ldots \lambda_n$ are the variances along those principle components



III. Interpretation of PCA

a) Variance maximization

"principle components capture the ..."

directions of maximum spread'

$$\tilde{X} = \begin{bmatrix} - (\vec{x}_1 - \bar{x})^T - \\ \vdots \\ - (x_m - \bar{x})^T - \end{bmatrix}$$

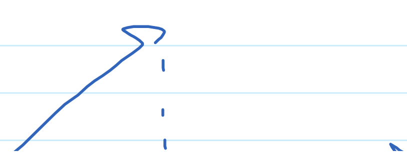
Then $\tilde{X}\vec{w}$, where $\vec{w}$ is a unit vector, is given by

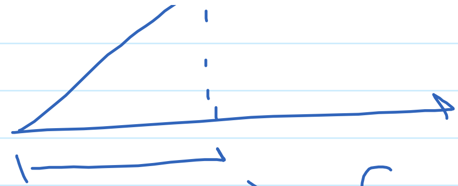
$$\tilde{X}\vec{w} = \begin{bmatrix} (\vec{x}_1 - \bar{x})^T \vec{w} \\ (\vec{x}_2 - \bar{x})^T \vec{w} \\ \vdots \\ (\vec{x}_m - \bar{x}_m)^T \vec{w} \end{bmatrix}$$

$$= \begin{bmatrix} \langle \vec{x}_1 - \bar{x}, \vec{w} \rangle \\ \vdots \\ \langle \vec{x}_m - \bar{x}, \vec{w} \rangle \end{bmatrix}$$

Remember that $\text{proj}_{\vec{w}} \vec{v} = \dfrac{\langle \vec{v}, \vec{w} \rangle}{\cancel{\langle \vec{w}, \vec{w} \rangle}} \vec{w}$

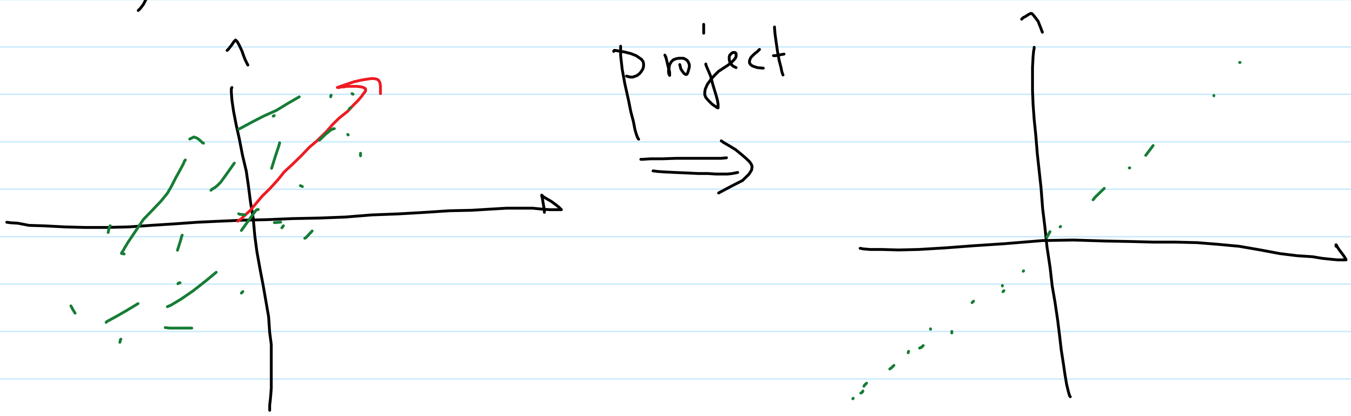
Scalar projection





$\langle \vec{v}, \vec{w} \rangle$ for $\|\vec{w}\| = 1$

$\tilde{X}\vec{w}$ projects the data contained in $\tilde{X}$ onto $\vec{w}$!!!



$\tilde{X}\vec{w}$ is an $m \times 1$ vector

$\Longrightarrow$ what is the variance?

$$\sigma_x^2 = \frac{1}{m}\sum_{i=1}^{m} (x_i - \bar{x})^2$$

Variance of projection looks like the norm squared!

$$\longrightarrow \boxed{\frac{1}{m}\|\tilde{X}\vec{w}\|^2 = \frac{1}{m}\vec{w}^T\tilde{X}^T\tilde{X}\vec{w}}$$

Want to find $\vec{w}$ s.t. this is

Want to find $\vec{w}$ s.t. this is maximized!

$\Rightarrow$ From SVD, already know the dir of max amplification

Dis 4B, last problem:

max gain: σ_1 for input $\vec{v}_1$

where $A^T A \vec{v}_1 = \sigma_1^2 \vec{v}_1$

$\Rightarrow$ $\vec{p}_1$ is the unit vector that maximizes $\vec{w}^T \tilde{X}^T \tilde{X} \vec{w} = \|\tilde{X}\vec{w}\|^2$

$\Rightarrow$ turns out that $\vec{p}_1$ is the eigenvector of $\tilde{X}^T\tilde{X}$ with the largest eigenvalue $\lambda_1 = \sigma_1^2$

Formally, the optimization problem:

variance along

$$\vec{p}_k = \max_{\vec{w}} \left\{ \vec{w}^T \tilde{X}^T \tilde{X} \vec{w} \right\}$$ variance along $\vec{w}$ direction

such that $\|\vec{w}\| = 1$ and

$$\vec{w}^T \vec{p}_1 = w^T \vec{p}_2 = \ldots = \vec{w}^T \vec{p}_{k-1} = 0$$

— In other words, (1) $\vec{p}_1$ maximizes the variance

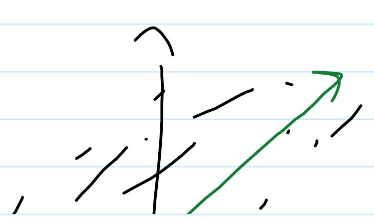
(2) $\vec{p}_2$ maximizes the variance in $\perp$ directions to $\vec{p}_1$

⋮

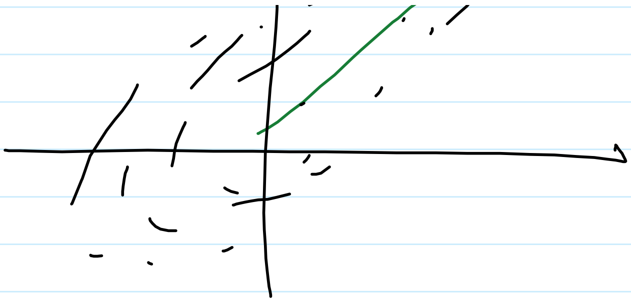
etc

Once again, from SVD:

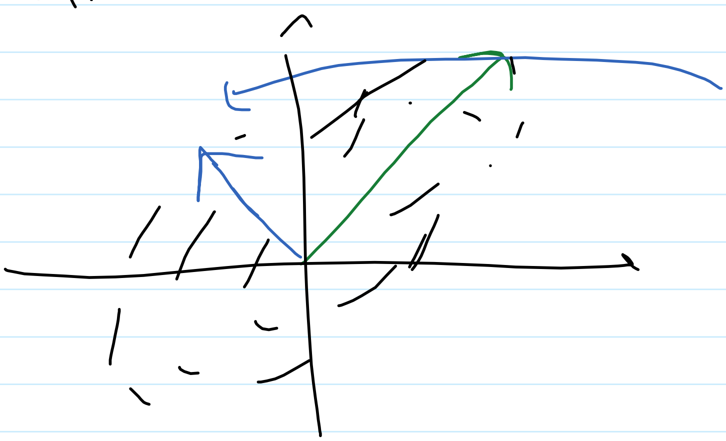
- max amp is in the $\vec{p}_1$ dir
- max amp in $\perp$ directions to $\vec{p}_1$ is given by $\vec{p}_2$, with gain σ_2



"line of best-fit"

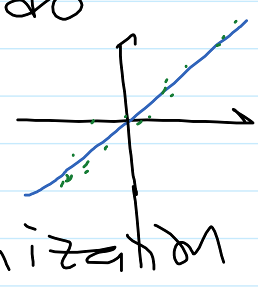


What's the second best fitting line?
i.e. what other directions capture variance in the data



$\Longrightarrow$ "line of best fit" reduces data to one-dimension

$\Longrightarrow$ can use PCA to do dimension reduction



b) Reconstruction Error Minimization

Problem: approximate data point as

$$\boxed{\vec{x}_i \approx \hat{\alpha}_1 \vec{v}_1 + \dots + \hat{\alpha}_l \vec{v}_l}$$

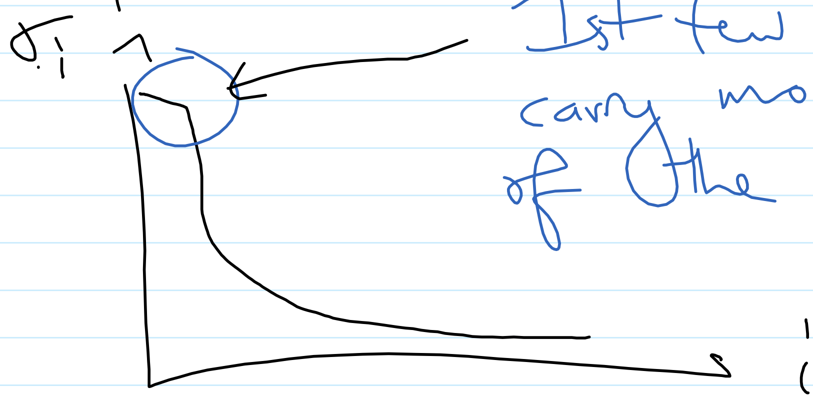
$$\boxed{\vec{X}_i \approx \hat{\alpha}_1 \vec{v}_1 + \dots + \hat{\alpha}_k \vec{v}_k}$$

$\Longrightarrow$ Store $k < n$ numbers to represent each data point

What are the best basis vectors $\vec{v}_i$?

$\Longrightarrow$ turns out to be PCA basis

Often project data onto 1st few principle components



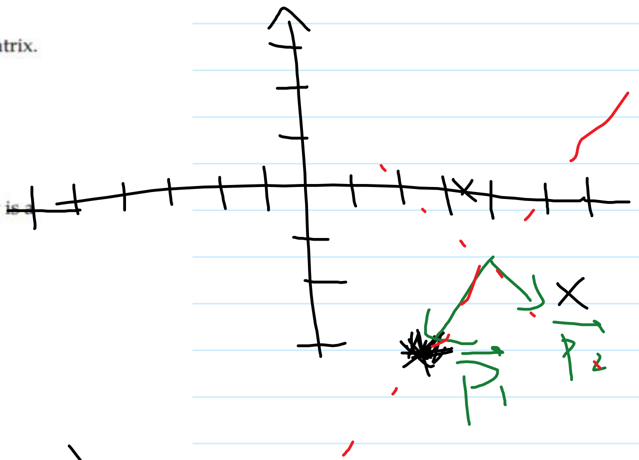
1st few carry most of the weight

3 PCA

Suppose we had the following data points $(x_i, y_i) \in \mathbb{R}^2$ aggregated in the following matrix.

$$A = \begin{bmatrix} 5 & -6 \\ 7 & 0 \\ 11 & -4 \\ 5 & -6 \end{bmatrix}$$

a) Find the covariance matrix S of A if each column is a type of data and each row is a measurement.



$m = 4$

$$S = \frac{1}{m} \tilde{A}^T \tilde{A}$$

$$\mu_1 = \frac{1}{4}(5+7+11+5) = 7$$

$$\mu_2 = \frac{1}{4}(-6+0+(-4)+(-6)) = -4$$

$$\tilde{A} = \begin{bmatrix} -2 & -2 \\ 0 & 4 \\ 4 & 0 \\ -2 & -2 \end{bmatrix}$$

$$\frac{1}{4} \begin{bmatrix} -2 & 0 & 4 & -2 \\ -2 & 4 & 0 & -2 \end{bmatrix} \begin{bmatrix} -2 & -2 \\ 0 & 4 \\ 4 & 0 \\ -2 & -2 \end{bmatrix} = \frac{1}{4} \begin{bmatrix} 24 & 8 \\ 8 & 24 \end{bmatrix}$$

$$\boxed{S = \begin{bmatrix} 6 & 2 \\ 2 & 6 \end{bmatrix}}$$

b) Since S is a symmetric matrix, we can eigendecompose it into the form $S = P\Lambda P^T$, where P contains the orthonormal principal components of S and Λ is a diagonal matrix with the squared weights of the corresponding principal components. Find the eigenvalues of S and order them from largest to smallest, $\lambda_1 > \lambda_2$.

$$S = \begin{bmatrix} 6 & 2 \\ 2 & 6 \end{bmatrix}$$

$$\lambda_1 = 8 \qquad \lambda_2 = 4$$

$$\vec{v}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \vec{v}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

c) Find the orthonormal eigenvectors $\vec{p}_i$ of S (all eigenvectors are mutually orthogonal and have unit length).

$$\vec{p}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \vec{p}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

d) What are the principal components of the system? What are the weights of each principal component?

std dev

$$\text{weights} \longrightarrow \sigma_i = \sqrt{\lambda_i}$$

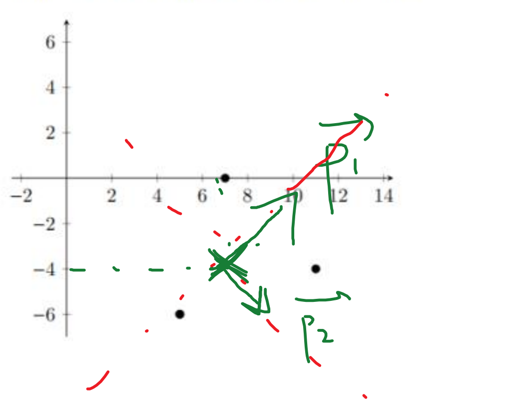
$$\sigma_1 = \sqrt{8} = 2\sqrt{2} \longrightarrow \vec{p}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\sigma_2 = \sqrt{4} = 2 \longrightarrow \vec{p}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ \ldots \end{pmatrix}$$

$$\sigma_2 = \sqrt{4} = 2 \longrightarrow \vec{p_2} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

$$\sigma_1 \vec{p_1} = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \qquad \sigma_2 \vec{p_2} = \begin{pmatrix} \sqrt{2} \\ -\sqrt{2} \end{pmatrix}$$

e) Plot the two principal components scaled by their weights on the following graph. Remember that we subtracted the column means from each column.



$$(7,-4) \longrightarrow (7+2, -4+2)$$
$$= (9, -2)$$
$$(7,-4) \longrightarrow (7+\sqrt{2}, -4-\sqrt{2})$$

4 Using the SVD for PCA

In the previous question, we viewed the principal components as the eigenvectors of the covariance matrix $S = \frac{1}{m}\tilde{A}^T\tilde{A}$. In this question, we see how Principal Component Analysis relates to the Singular Value Decomposition.

a) Given m data points $(x_i, y_i)_{i=1}^m$ in $\mathbb{R}^2$, what is our data matrix A?

each col is an attribute

$$A = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_m & y_m \end{bmatrix}$$

each row is a measurement

b) If the SVD of M is $M = U\Sigma V^T$, what are the eigenvectors and eigenvalues of M^TM?

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

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

$$M^TM = V\Sigma^T \underbrace{\cancel{U^TU}}_{I} \Sigma V^T$$

$$= V\Sigma^T\Sigma V^T$$

diagonalization equation!

diagonalization equation!

$\Longrightarrow$ $\boxed{\begin{array}{l} \lambda_i = \sigma_i^2 \\ \text{eigenvectors are } \vec{v}_i \end{array}}$

c) How can we use the SVD to compute our principal components?

Goal of PCA is to diagonalize $S = \frac{1}{m}\tilde{A}^T\tilde{A}$

(1) We start w/ data matrix A:

a) de-mean / center A

$\Longrightarrow \tilde{A} = A - \bar{A}$

b) scale $\tilde{A}$ such that

$$M = \alpha\tilde{A} \Longrightarrow M^TM = \frac{1}{m}\tilde{A}^T\tilde{A}$$

$$M^TM = \alpha^2\tilde{A}^T\tilde{A} = \frac{1}{m}\tilde{A}^T\tilde{A}$$

$$\Longrightarrow \alpha = \frac{1}{\sqrt{m}}$$

$$\boxed{\text{Define } M = \frac{1}{\sqrt{m}}\tilde{A}}$$

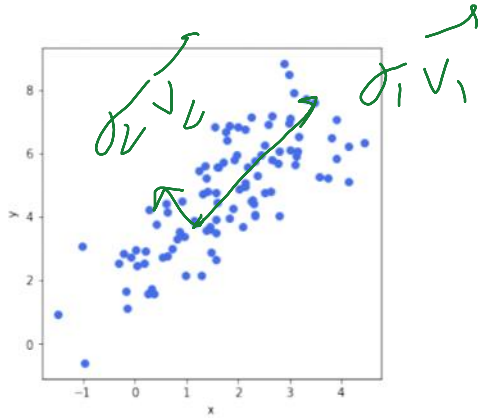
c) Calculated SVD of M!

$M = U\Sigma V^T \longleftarrow$ eigenvecs

$$M = U \Sigma V^T \longleftarrow \text{eigenvecs of } M^T M = \frac{1}{m}\tilde{A}^T \tilde{A} = S$$

- σ_i give weights
- V_i give principal components

d) If the given data looked like the following figure, what would you expect $\sigma_1 \vec{v}_1$ and $\sigma_2 \vec{v}_2$ to be?



e) Sketch the projection of the demeaned data onto the principal components $\vec{v}_1$ and $\vec{v}_2$.