Business data Science
Lecture 19. Nov 1st

Continue ON PCA. Principal Component Analysis
a method for doing dimensionality Reduction.
(its an unsupervised ML technique).
We will see why it can be a very bad idea
to use it naively for supervised Learning).

(SVD, eigenvalue decompositions and PCA are
closely connected tools.)

Dimensionality Reduction: features are 10 dimensional
and you want to create new features that
are 1 dimensional OR 2D.

---

1. PCA is a projection of data points on a linear subspace such that the variance of the projected data is maximized.

2. This (turns out) to be equivalent to finding the subspace that minimizes the error.

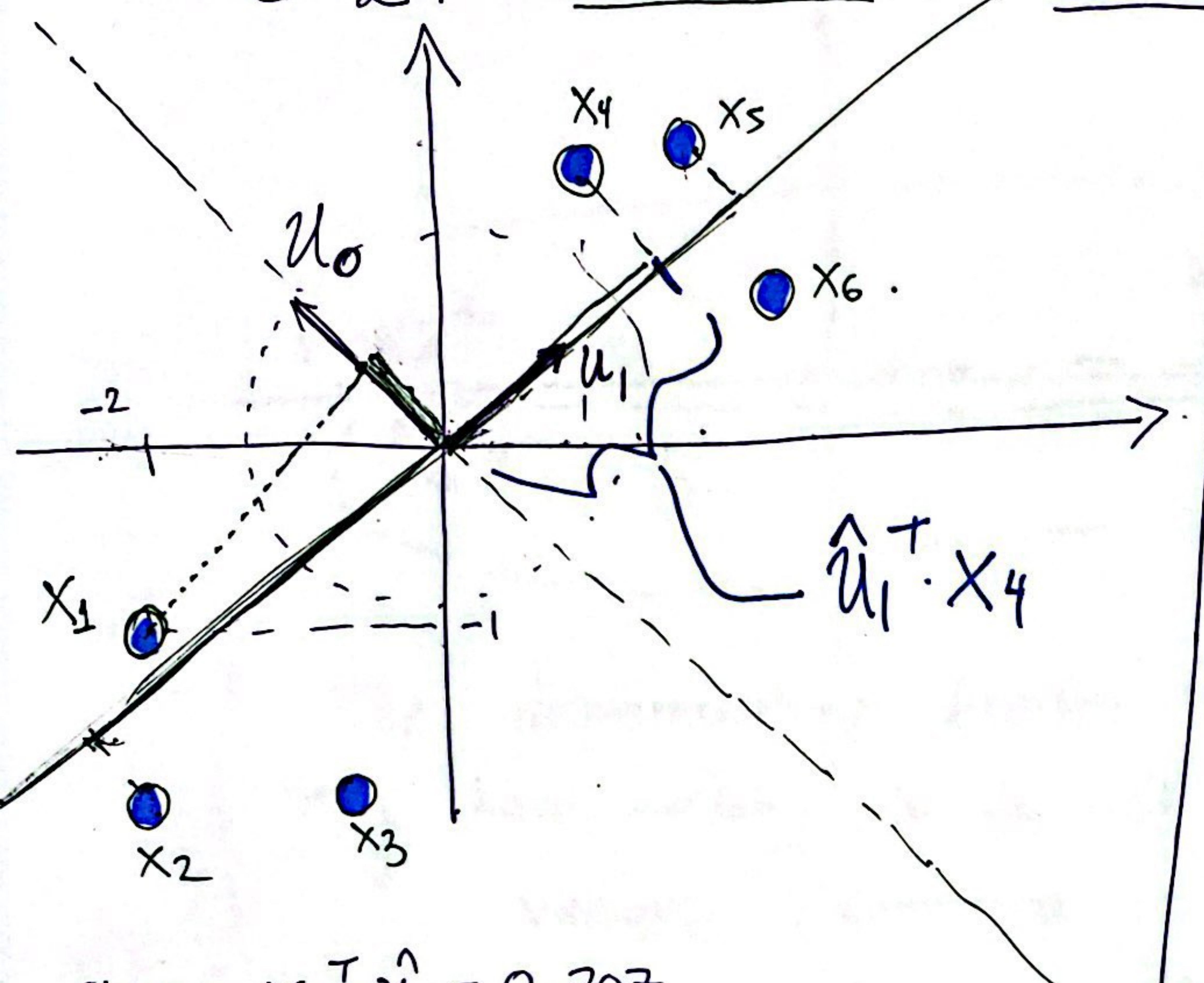
3. Also this turns out to be equivalent to Finding new features that are uncorrelated.
   - PCA whitens the data. (White noise symmetric in all directions).

[This special subspace is called the principal subspace
and you can do it any dimension you want.]

2. Given a dataset of points $x_1, x_2 \ldots x_n \in \mathbb{R}^d$.

Lets see what the first principal component is.

$d = 2$. Centered data: $x_i^c = x_i - \bar{x}$



Example

$X_1 = \begin{bmatrix} -2 \\ -1 \end{bmatrix}$. $u_0 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$.

· Project X_1 on u_0.

· First we normalize u_0.

$$\hat{u}_0 = u_0 \cdot \frac{1}{\|u_0\|} = u_0 \cdot \frac{1}{\sqrt{u_0^T \cdot u_0}}$$

$$\hat{u}_0 = \frac{1}{\sqrt{2}} \cdot \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Projection of x_1 on u_0. is $x_1^T \cdot \hat{u}_0$

$$= \frac{1}{\sqrt{2}} \cdot (2 - 1) = \frac{1}{\sqrt{2}} = 0.707.$$

$$\hat{X}_1 = (\text{Projected} \cdot \text{Length.}) \cdot \text{Projeted direction}$$

$$= \frac{1}{\sqrt{2}} \cdot \hat{u}_0$$

$$\hat{X}_1 = (X_1^T \cdot \hat{u}_0) \cdot \hat{u}_0.$$

$$= \frac{1}{\sqrt{u_0^T \cdot u_0}} \cdot (X_1^T \cdot u_0) \frac{1}{\sqrt{u_0^T \cdot u_0}} \cdot u_0$$

$$\hat{X}_1 = \frac{1}{u_0^T \cdot u_0} (X_1^T \cdot u_0) \cdot u_0.$$

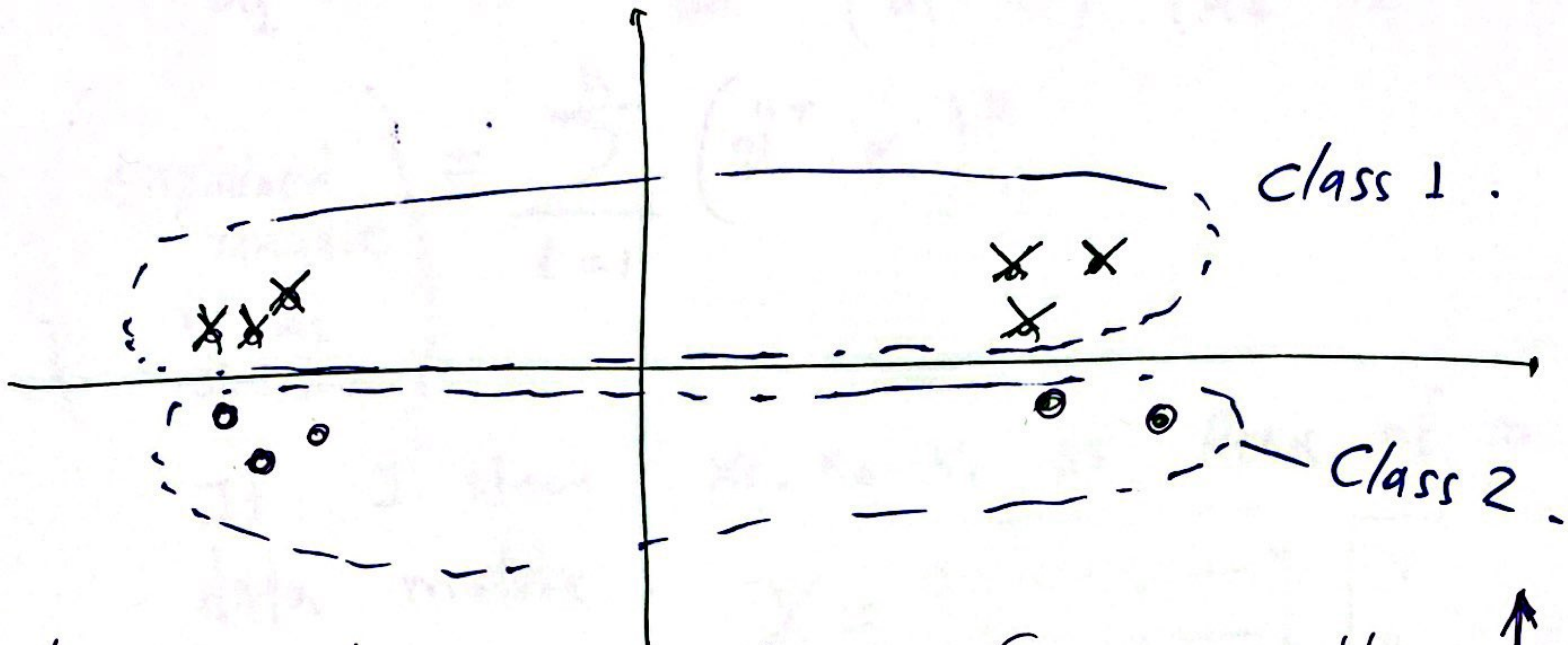
since $x_1^T \cdot \hat{u}_0 = 0.707$ the explained variance in the direction $\hat{u}_0$ is 0.707.

Explained variance by $\hat{u}_1$ is $(\hat{u}_1^T \cdot x_1)^2 + (\hat{u}_1^T \cdot x_2)^2 + .. (\hat{u}_1^T \cdot x_n)^2$.

$$= \sum_i (\hat{u}_1^T \cdot x_i)^2.$$

if I stack x_i as Rows of a Matrix $X = n \begin{bmatrix} -x_1- \\ -x_2- \\ \vdots \end{bmatrix}$

When is PCA a terrible direction for classification?
1st component



The discriminating direction for your problem may have nothing to do with the direction of maximum variance ($\longleftrightarrow$.).

(When this happens, PCA is a disaster for supervised problems.)

~~Given data points $X_1, X_2 .. X_n \in R^d$.~~
~~I am unit for a unit length vector u_1~~
~~that maximizes~~

Given $x_1, x_2 .. x_n \in R^d$ (centered data!)

explained variance in a unit direction $\hat{u}_1$ is $(\hat{u}_1^T \cdot x_1)^2 + (\hat{u}_1^T \cdot x_2)^2 + .. (\hat{u}_1^T \cdot x_n)^2$.

$$\begin{pmatrix}\text{explained} \\ \text{variance} \\ \text{in } \hat{u}_1 \\ \text{direction.}\end{pmatrix} = \sum_{i=1}^{n} (\hat{u}_1^T \cdot x_i)^2.$$

If I stack $x_1, x_2 .. x_n$ as Rows of a data matrix

$$X = \begin{array}{c|c} & \xleftrightarrow{d} \\ \updownarrow n & \begin{bmatrix} - x_1 - \\ - x_2 - \\ \vdots \\ - x_n - \end{bmatrix} \end{array}$$

v.
$\|v\|^2 = v^T v$.
(Trick 1.).

$(Xu)^T = u^T X^T$
(trick 2).
$(ABC)^T = C^T B^T A^T$.

$$X \cdot \hat{u}_1 = \begin{array}{c|c} \updownarrow n & \begin{bmatrix} - x_1 - \\ - x_2 - \\ \vdots \\ - x_n - \end{bmatrix} \end{array} \cdot \begin{bmatrix} | \\ \hat{u}_1 \\ | \end{bmatrix} = \begin{bmatrix} x_1^T \cdot \hat{u}_1 \\ x_2^T \cdot \hat{u}_1 \\ \vdots \\ x_n^T \cdot \hat{u}_1 \end{bmatrix}$$

$$\text{Explained variance} = \|X \cdot \hat{u}_1\|_2^2 = \sum (x_i^T \cdot \hat{u}_1)^2.$$

1st principal component is the vector $\hat{u}_1$

that $$\max_{\|u_1\|=1} \sum (x_i^T \cdot u_1)^2 = \max_{\|u_1\|=1} \|X \cdot u_1\|_2^2.$$

Using trick 1: $$\max_{\|u_1\|=1} (Xu)^T (Xu).$$

trick 2: $$\max_{\|u\|=1} u^T \underbrace{X^T X}_{\text{This is the covariance matrix for centered data.}} \cdot u.$$

$$\max_{\|u\|=1} u^T C \cdot u.$$