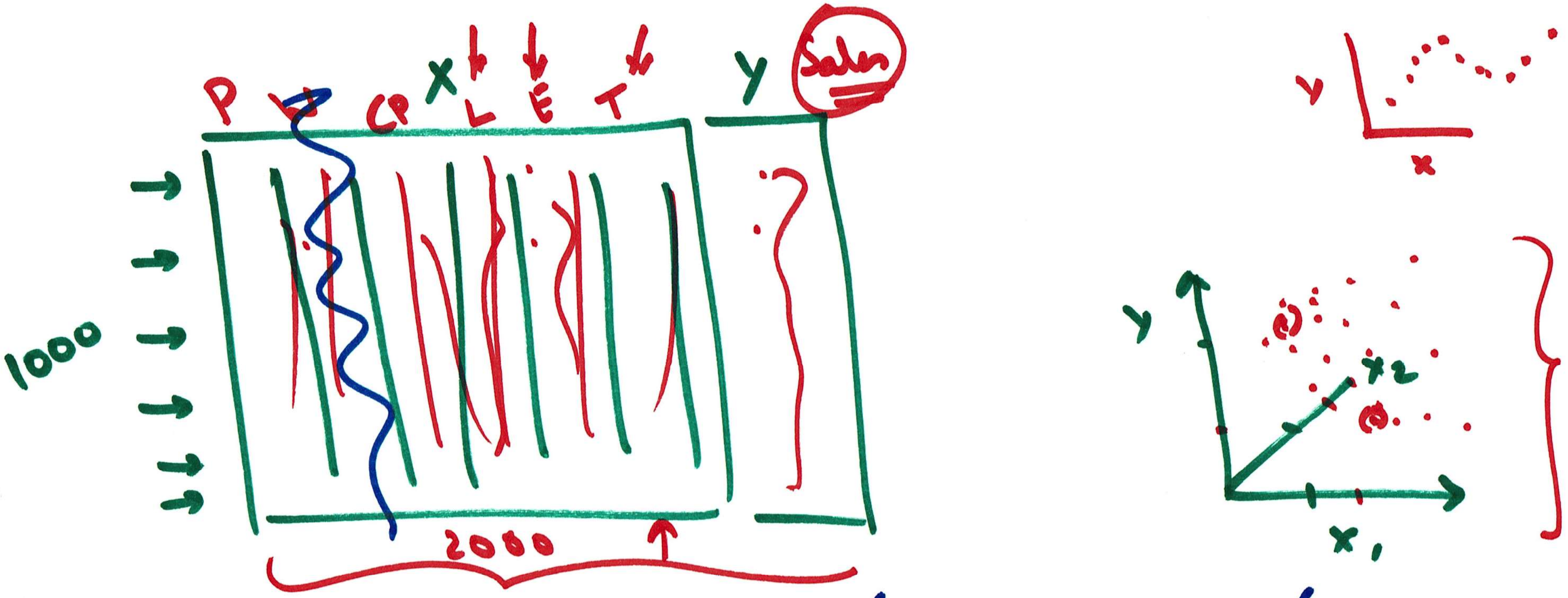
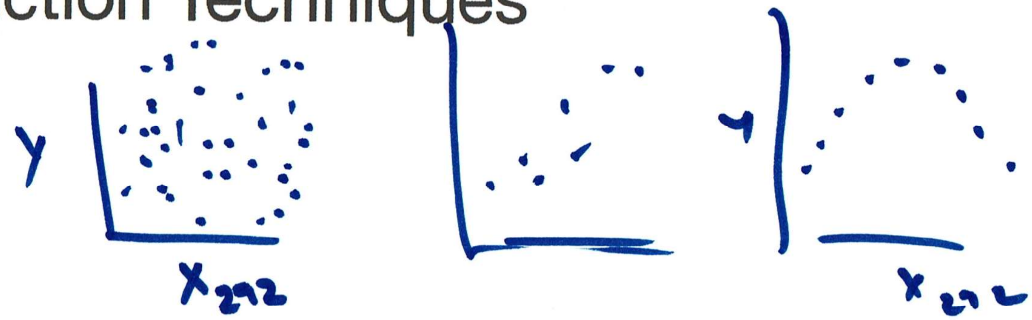
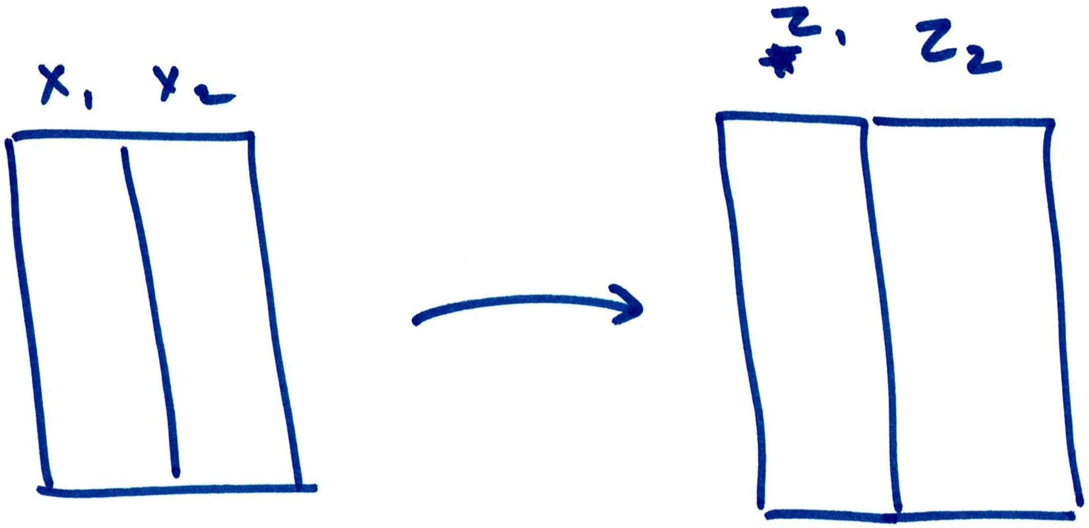


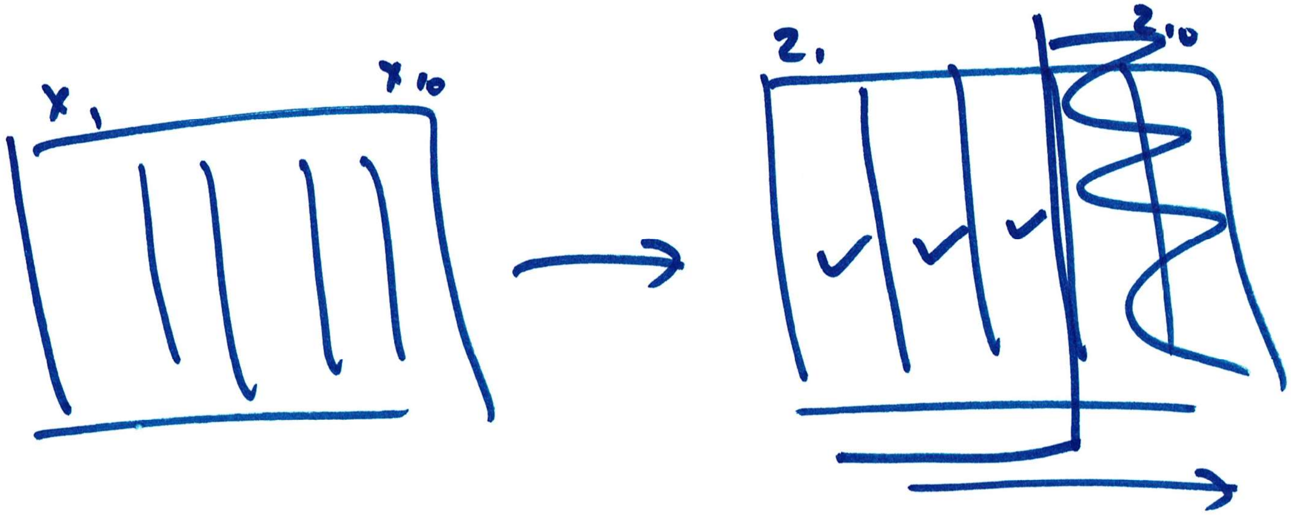
$$y = a + b_1 \square + b_2 \square \dots \dots + b_{2000} \square$$

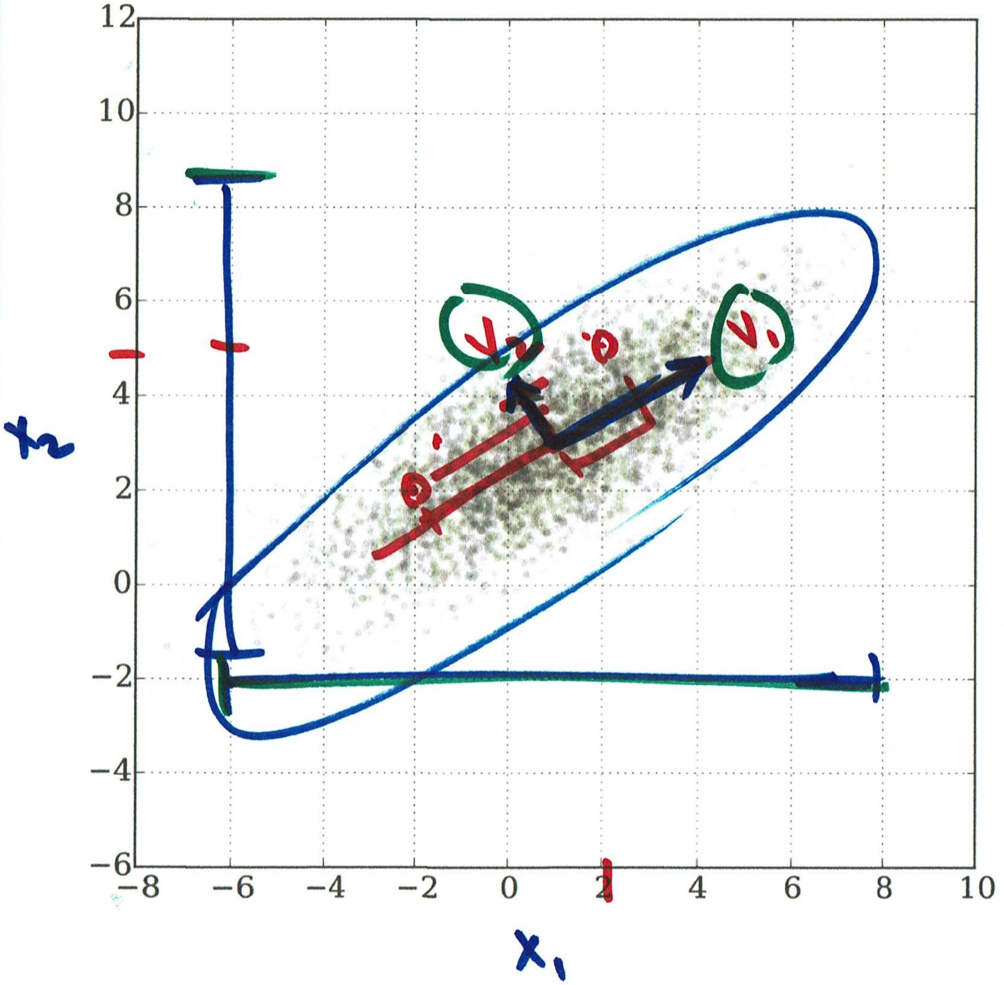
Dim. Reduction Techniques



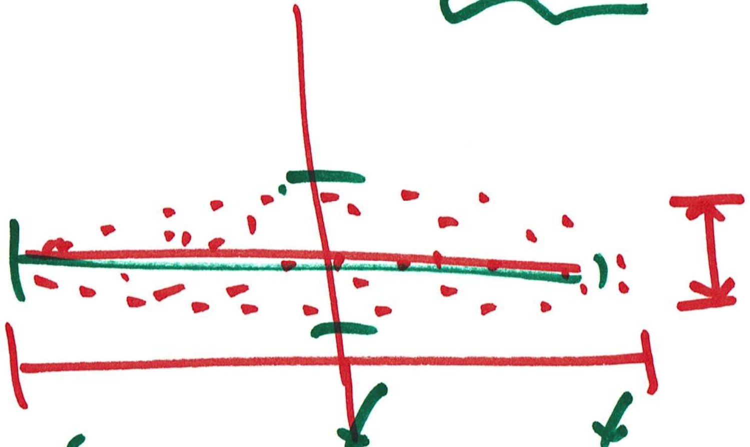
- Feature elimination
  - Simply identify and remove variables (columns) that are not important
  - The disadvantage is that we would gain no insight from those dropped variables and loose any information they contain
- Feature extraction
  - Create a few new variables from the old variables
  - **PCA** Principal Component Analysis: is the most popular feature extraction technique (linear)
  - t-SNE (non-linear)


x_1
x_2
z_1
z_2


x1
x10
z1
z10



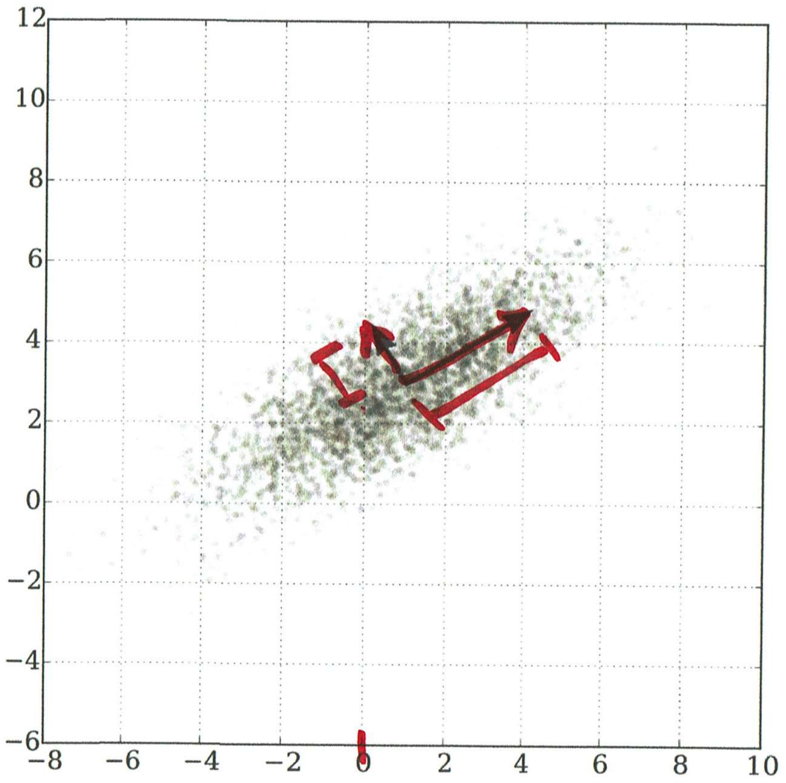
$$\begin{array}{cc} X_1 & X_2 \\ 2 & 5 \\ -2 & 2 \\ 1.2 & 0.8 \end{array} \rightarrow \begin{array}{cc} Z_1 & Z_2 \\ 1 & 0.8 \\ -2 & 0.5 \\ 1.9 & 0.1 \end{array}$$

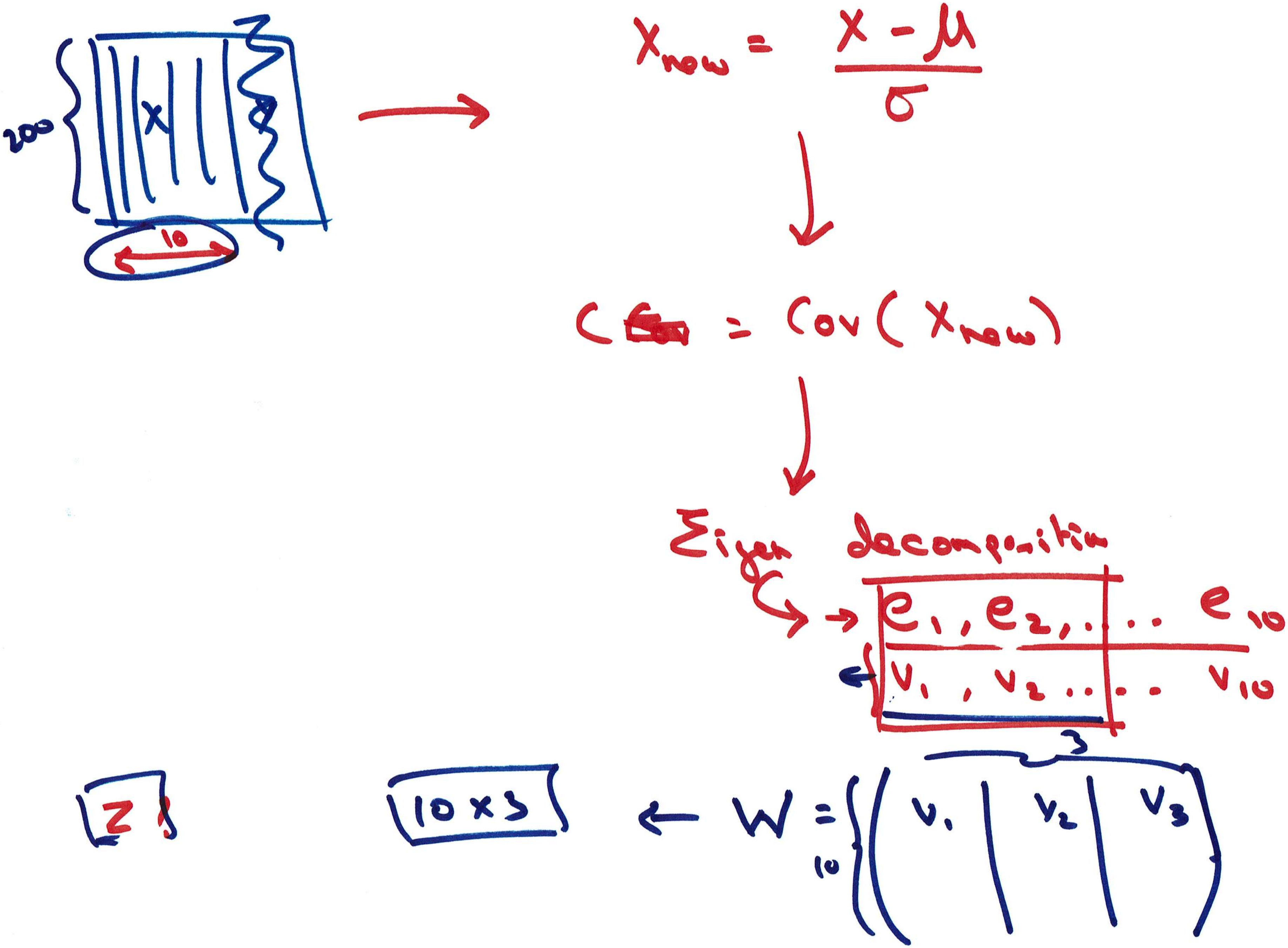


$$\begin{cases} Z_1 = \square\, x_1 + \square\, x_2 \\ Z_2 = \square\, x_1 + \square\, x_2 \end{cases}$$

PCA

- creates new variables using linear combinations of old variables
- is designed to create variables that are independent of one another
- also manages to tell us how important each of these new variables are
- this “importance”, helps us to choose how many variables we will use

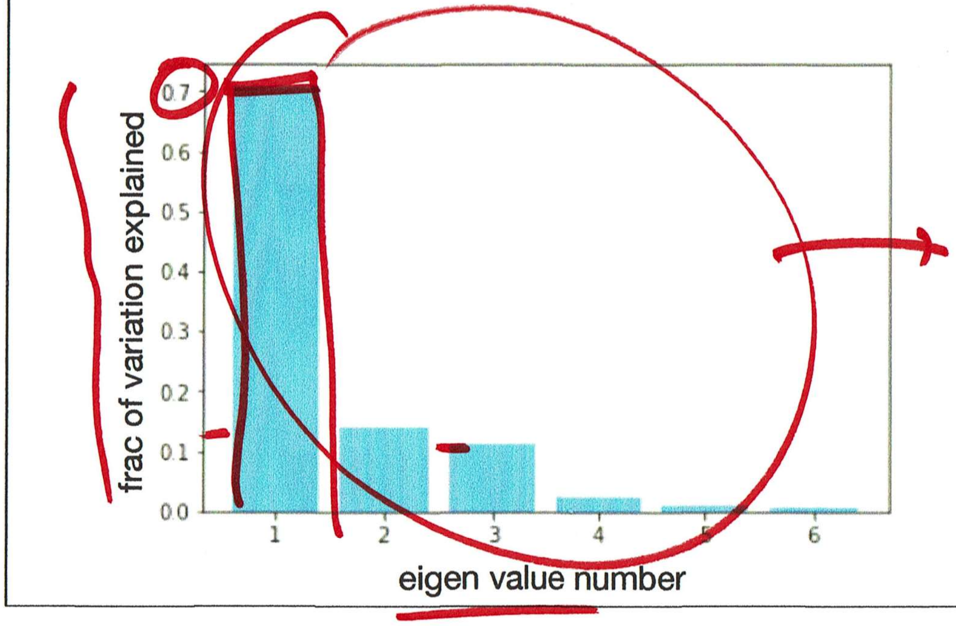


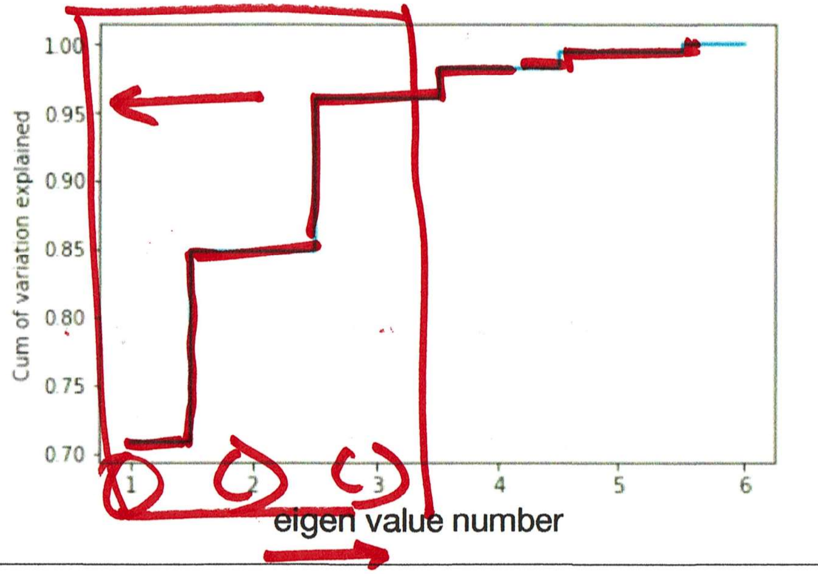

200
X
10
$X_{now} = \frac{X - \mu}{\sigma}$
$C = Cov(X_{now})$
Eigen decompositin
$e_1, e_2, \ldots \ldots e_{10}$
$v_1, v_2 \ldots \ldots v_{10}$
3
$W =$
10
v_1
v_2
v_3
10 x 3
Z

- Scale the data and compute the covariance matrix
- Break the covariance matrix into magnitude and direction. Eigen Vectors and the Eigen Values of the covariance matrix can be thought of as the natural axis/directions and magnitudes along those axis, of the data
  - The eigen values also can be used to calculate the percentage of variation explained by each component
- Sort in the eigen values in desending order and calculate the cumulative percentage of variation explained

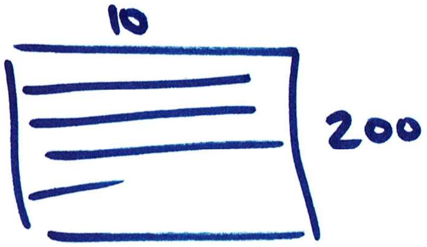
$$\frac{e_1}{\Sigma e_i}, \frac{e_2}{\Sigma e_i}, \frac{e_3}{\Sigma e_i}$$

- Pick the number of principal components you will use
- Transform to new variables

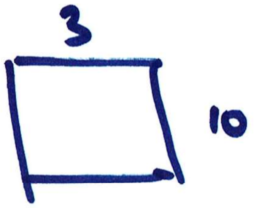




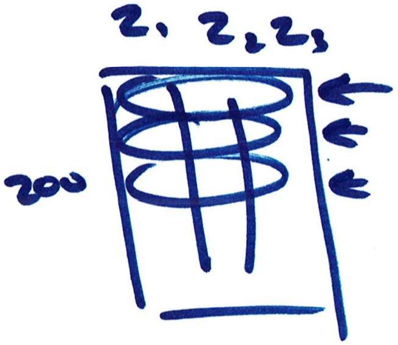
X =



W :



$$Z = \overset{200\times10}{X}\ \overset{10\times3}{W} \Rightarrow \overset{200\times3}{Z}$$



$$\boxed{S = 100 + 20 z_1 + 15 z_2 - 10 z_3}$$

$$z_1 = \square\, x_1 + \square\, x_2 + \square\, x_3 \ldots x_{10}$$

$$z_2 = \ldots$$

$$z_3 = \ldots$$