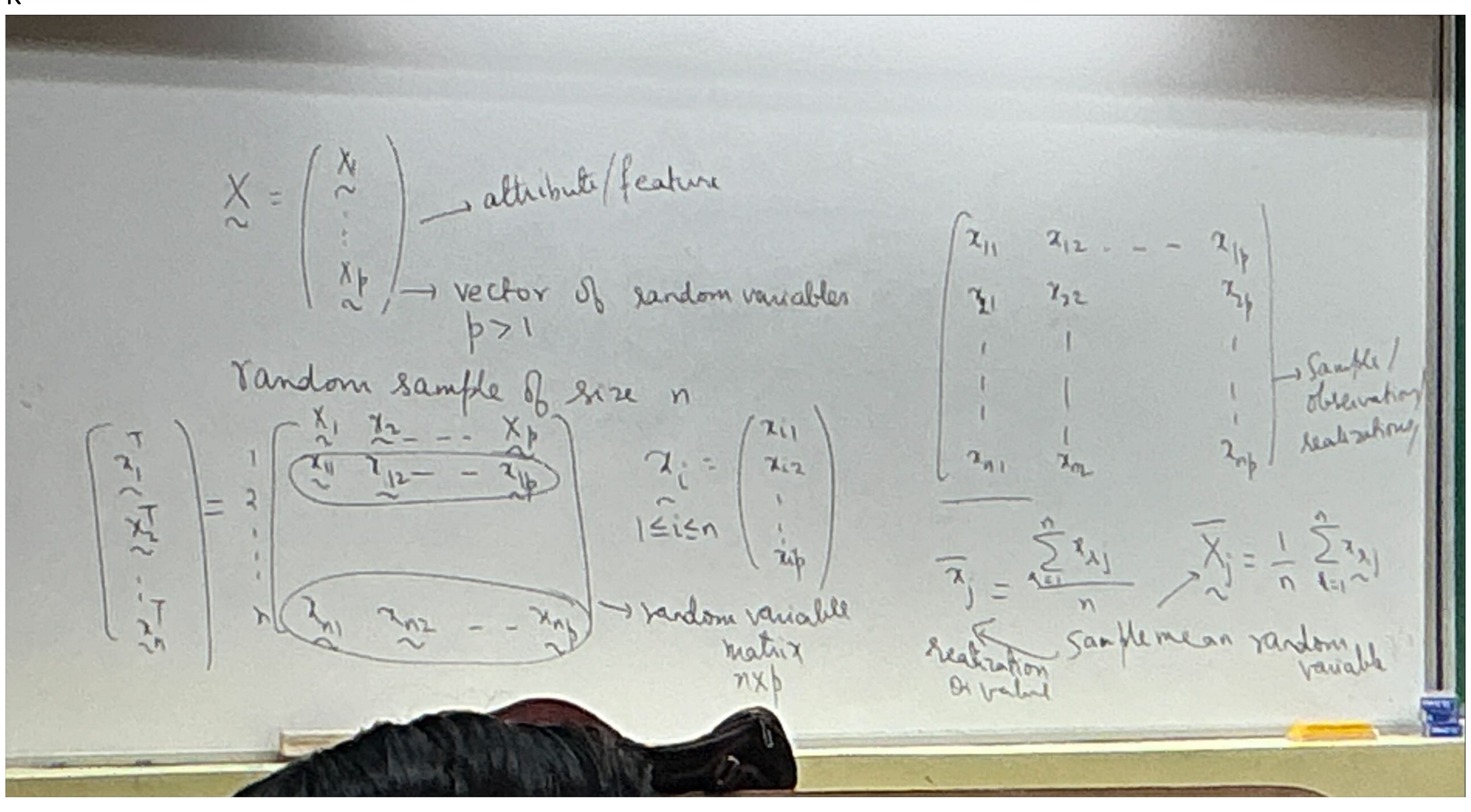
Lec 2

Tuesday, 29 July 2025 10:26 AM

K



$\underset{\sim}{X} = \begin{bmatrix} \underset{\sim}{X_1} \\ \vdots \\ \underset{\sim}{X_p} \end{bmatrix}_{p \times 1}$ random vector.

for one variable var $\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$.

$\underset{\sim}{X}\ \underset{\sim}{Y}$ $\quad cov(\underset{\sim}{X}, \underset{\sim}{Y}) = E[(X - \bar{X})(Y - \bar{Y})]$.

2. r.v.s.

$\underset{\sim}{Z} = \begin{pmatrix} \underset{\sim}{X} \\ \underset{\sim}{Y} \end{pmatrix}$

$\sim$ denotes random var or vector.

post experimentation $\quad s_{ij} = \frac{1}{n} \sum_{r=1}^{n} (x_{ri} - \bar{x}_i)(x_{rj} - \bar{x}_j)$.

$Cov(X_i, X_j)$.

$s_{ii} = Var(x_i)$

$= Cov(x_i, x_i)$.

$S = [s_{ij}]_{p \times p}$.

Variance - covariance, sample matrix.

After experimentation, we have the realisation of the variables → deterministic.

Population from which sample is drawn

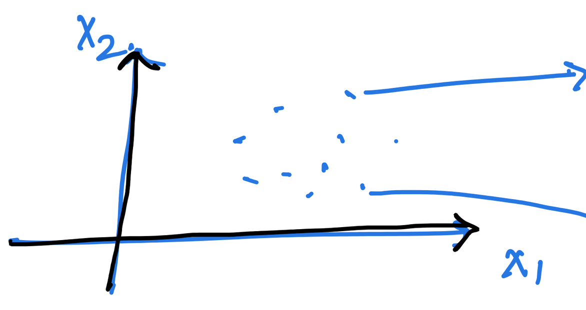
$\underset{p \times 1}{\underset{\sim}{\mu}}$. $\quad \underset{p \times p}{\Sigma}$

Let look at the sample data from p-dim perspective (row wise).

each sample in $\mathbb{R}^p$ & n samples.

plot these samples - we get a scatter plot in p-dim.

$p = 2$.



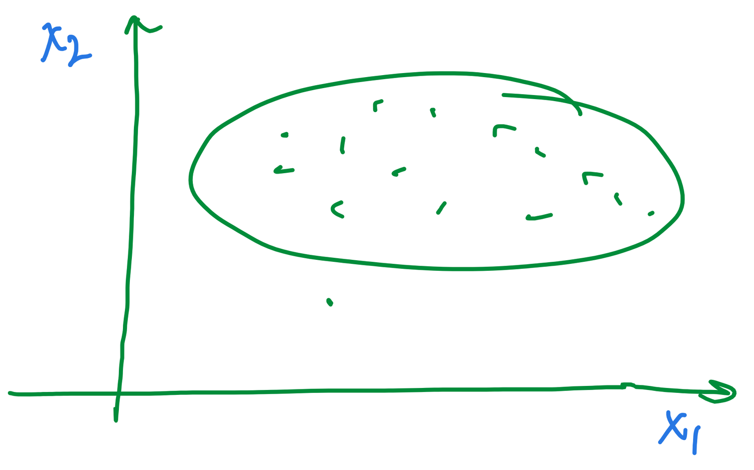
sample pt 10.

$p \times 1$, $p = 2$.

sample pt 1.

Scatter plot of sample.

These pts. are independent

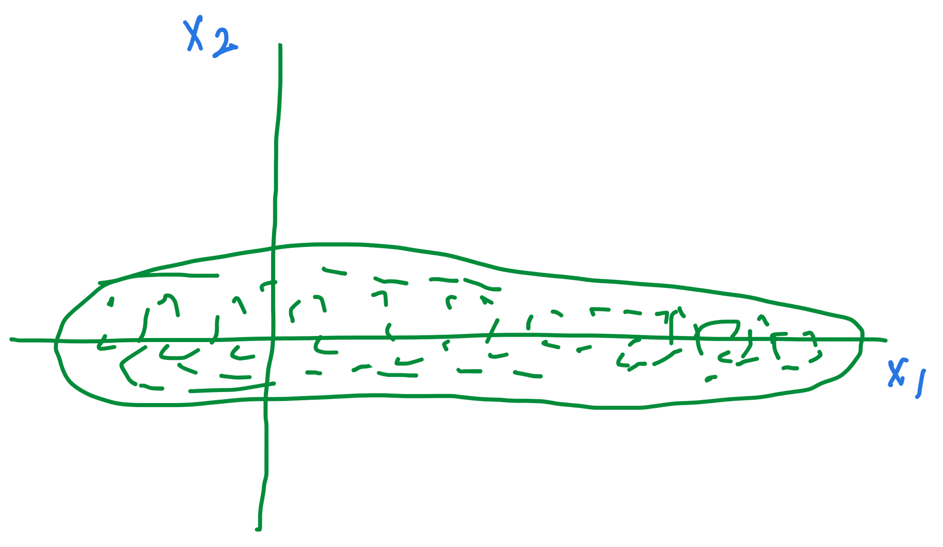


It takes almost an elliptical shape there might be some outliers.

$S_{2 \times 2}$. $\begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix}$.

α, β ratio not very large.

$\alpha \simeq \beta$.



x_1 dominating

Another ex.

variation along 1 feature is much more as compared to along the other feature.

Feedback example.

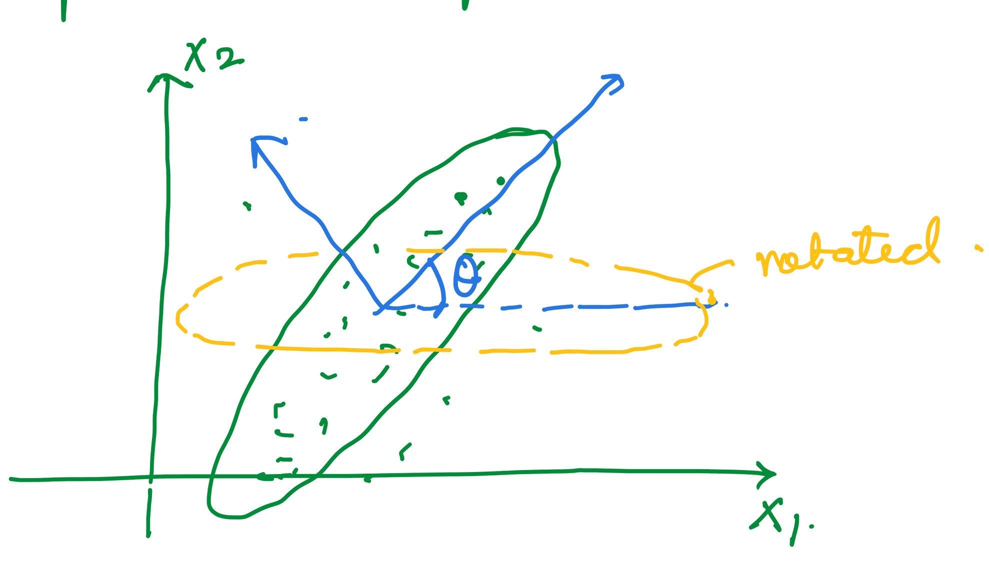
Independent so covariance zero

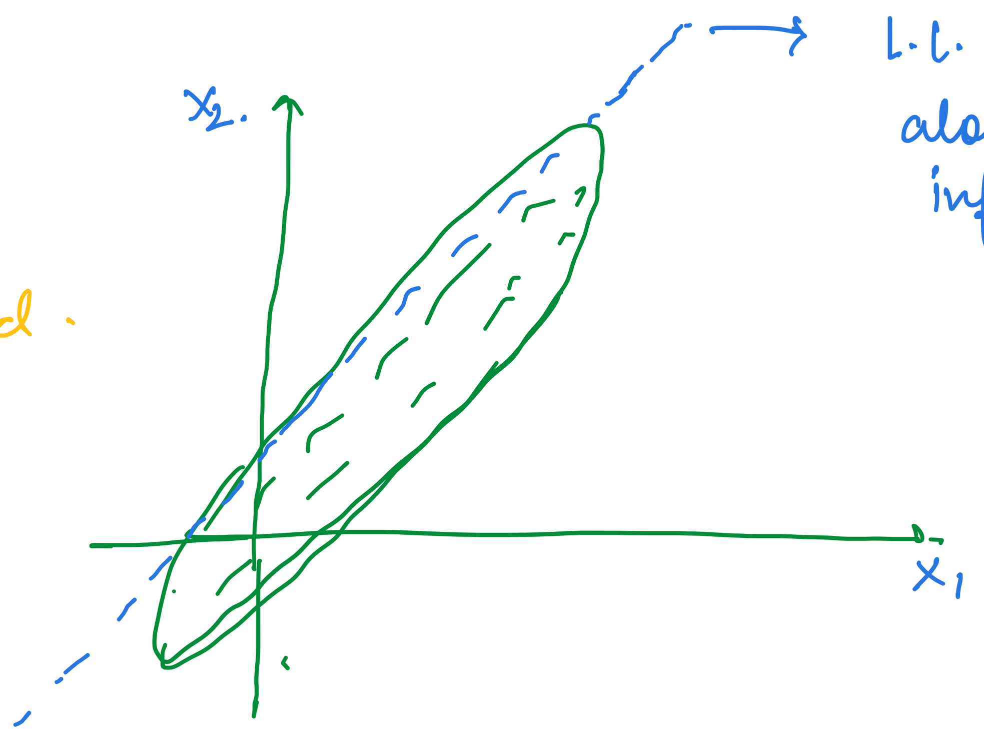
$S_{2 \times 2}$. $\begin{bmatrix} \gamma & 0 \\ 0 & \delta \end{bmatrix}$.

$\gamma >>> \delta$.

Since diagonal matrix, these values are the eigen values α, β.

\# Dependent @ $p = 2$.





L.C. of x_1 & x_2.

along this axis → vital information.

rotate θ angle to make it standard ellipse.

Similarly for the next.

Can also shift entire data to origin. So that centroid of the datapoints = 0.

one feature dominating

some linear combinations of both the features may be a good idea to use. Instead of 2 columns, a new col.