Some statistics (skipping proofs)

Suppose

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i$$

where

$i = 1, 2, \ldots, n$

$x_{i1}, \ldots, x_{i,p-1}$ are known constants

ε_i are independent $N(0, \sigma^2)$

$\beta_0, \ldots, \beta_{p-1}$ parameters

Defn $MSR = \dfrac{SSR}{p-1}$ regression mean square

$MSE = \dfrac{SSE}{n-p}$ Error mean square

$F^* = \dfrac{MSR}{MSE}$

Theorem If $\beta_1 = \beta_2 = \beta_3 = \cdots = \beta_n = 0$ then F^* follows an F distribution with $p-1$ and $n-p$ degrees of freedom.

To choose between the following two hypotheses

C_1 : $\beta_1 = \beta_2 = \cdots = \beta_n = 0$

C_2 : $\beta_i \neq 0$ for at least one i

We use:

If $F^* \leq F(1-\alpha, p-1, n-p)$ then conclude C_1

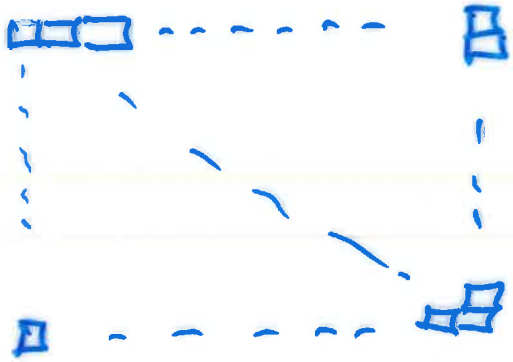
If $F^* > F(1-\alpha, p-1, n-p)$ then conclude C_2

to control Type I errors at level α.

Principal Component Analysis

Consider a collection of data points $w_1, w_2, \ldots, w_n \in \mathbb{R}^p$, where p may be large.

Example $w_1, \ldots, w_n \in \mathbb{R}^{256^2}$ are vectors representing n grey-scale images of faces. A digital image is a 256×256 array of pixels



Each pixel's grey-ness is determined by an integer in $\mathbb{R}$, and the image is thus represented by a 256×256 real matrix. Concatenating rows yields a vector $w \in \mathbb{R}^{65536}$.

Define the mean of $w_1, \dots, w_n \in \mathbb{R}^p$ as

$$\bar{w} = \frac{1}{n}(w_1 + w_2 + \cdots + w_n).$$

Set

$$\boxed{v_i = w_i - \bar{w}}$$

Then $v_1, \dots, v_n \in \mathbb{R}^p$ are data points with mean

$$\bar{v} = \frac{1}{n}(v_1 + \cdots + v_n) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^p$$

We'll use notation

$$v_1 = \begin{pmatrix} x_{1,1} \\ x_{1,2} \\ \vdots \\ x_{1,p} \end{pmatrix}, \quad v_2 = \begin{pmatrix} x_{2,1} \\ x_{2,2} \\ \vdots \\ x_{2,p} \end{pmatrix}, \quad \dots, \quad v_n = \begin{pmatrix} x_{n,1} \\ \vdots \\ x_{n,p} \end{pmatrix}$$

$$\bar{x}_i := \frac{1}{n}(x_{1,i} + x_{2,i} + \cdots + x_{n,i}) = 0$$

average of i^{th} row

Define the covariance matrix

$$C = \begin{pmatrix} c_{11} & \cdots & c_{1p} \\ \vdots & & \vdots \\ c_{p1} & \cdots & c_{pp} \end{pmatrix}$$

by

$$c_{ij} = \frac{1}{n} \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

$$= \frac{1}{n} \sum_{k=1}^{n} x_{ki}\, x_{kj}$$

covariance of i^{th} and j^{th} rows

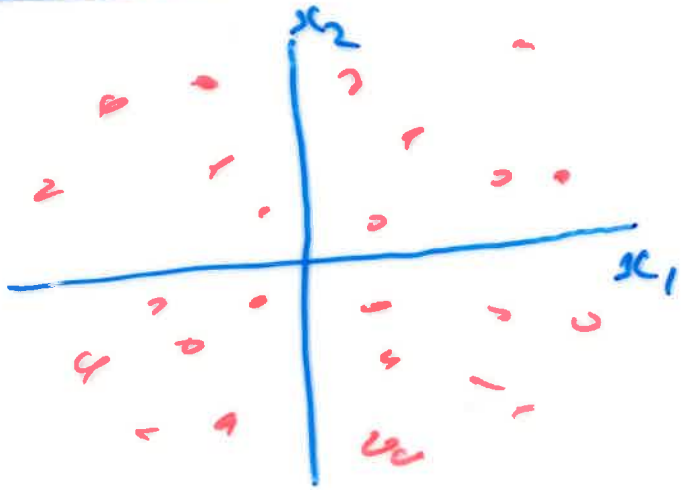
Defn x_{*i} and x_{*j} are uncorrelated if $c_{ij} = 0 = c_{ji}$

Illustration ($p = 2$)

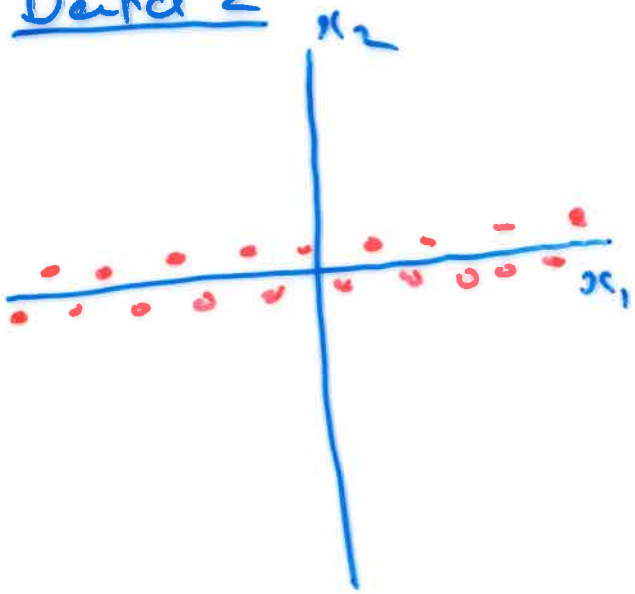
Consider four data sets

$\{v_1, \ldots, v_n\} \subseteq \mathbb{R}^2$

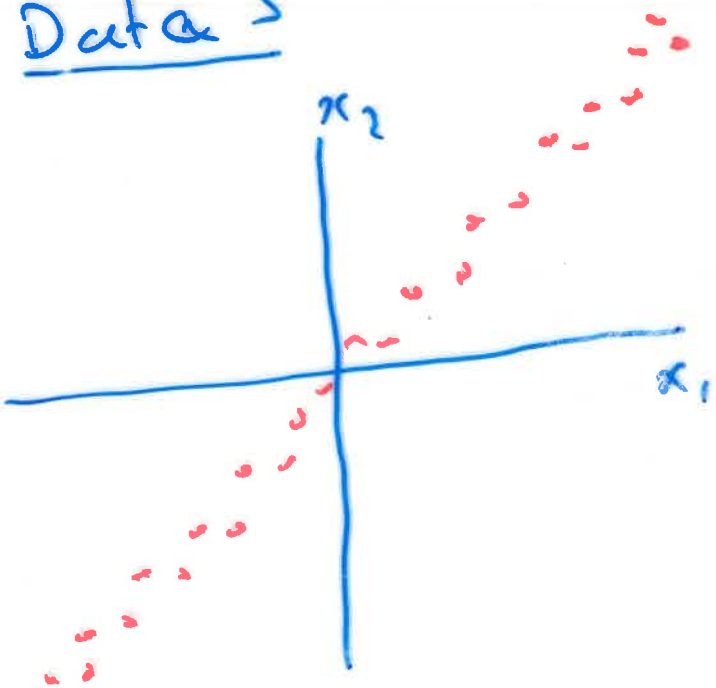
Data 1



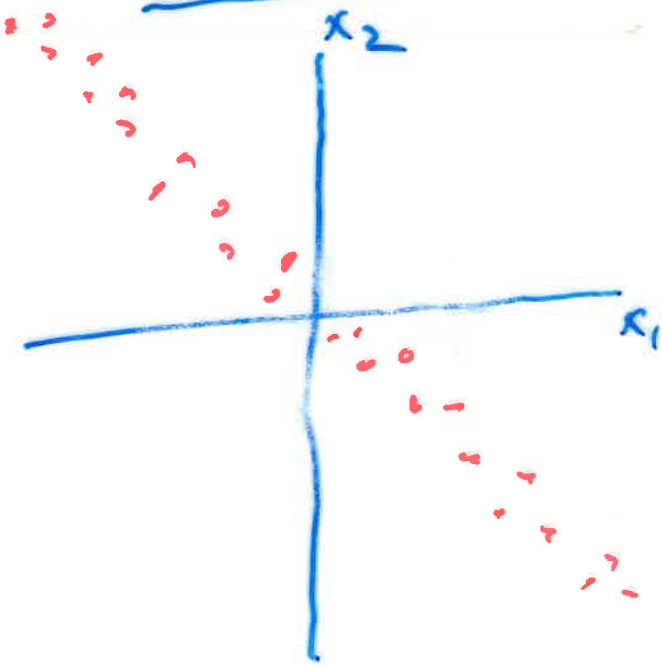
Data 2



Data 3



Data 4



for each of the four data sets let's consider the covariance matrix

$$C = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$$

Data set	c_{11}	c_{22}	$c_{12} = c_{21}$
1	large	large	0
2	large	small	0
3	large	large	large (positive)
4	large	large	large (negative)