DA_2020 L25

PCA IN A NUTSHELL

(See Higgs & Attwood chap. 2.V)

* Structured data ($N \times P$) matrix

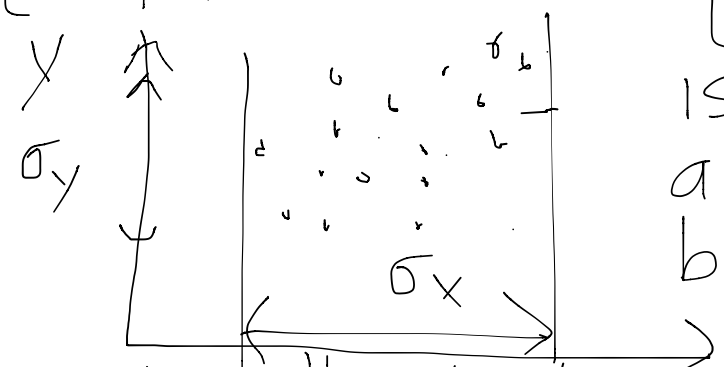
$$\begin{pmatrix} & & \\ & X_{ij} \leftarrow & \\ & & \end{pmatrix}$$

$i = 1, \ldots, N$ rows $\approx$ objects

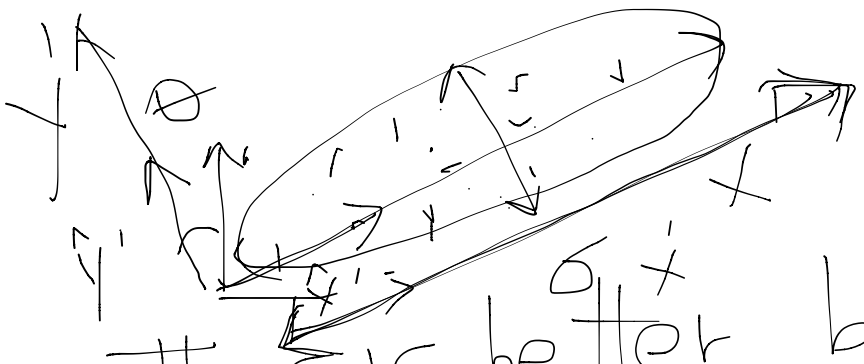
$j = 1, \ldots, P$ measured quantities

We are collecting P dimensional data (rows)

the idea of PCA



Q.
is x y an optimal base to describe the objects x



This is better because we have more variability along x' than along x

Let us find the optimal base throug linear algebra (see Hint box 2.2)

1 DATA X_{ij} $i = 1, \ldots, N$ $j = 1, \ldots, P$

2 AVERAGE Parameters over objects

$$\rightarrow \mu_j = \frac{1}{N} \sum_{i=1}^{N} X_{ij} \qquad j = 1, \ldots, P$$

3 MEASURE DISPERSIONS

$$\rightarrow \sigma_j = \left(\frac{1}{N} \sum_{i=1}^{N} (X_{ij} - \mu_j)^2 \right)^{1/2}$$

4 Z-transform (standardize data)

$$Z_{ij} \equiv \frac{X_{ij} - \mu_j}{\sigma_j}$$

5 change the base

Consider P vectors

$\underline{V}_j = (V_{j1}, V_{j2}, \dots V_{jP})$ ← row

as a new base over which to project the original Z_{ij} (coordinates) of the N data vectors

* to be a base a set of vectors should be orthonormal

$$\sum_{k=1}^{P} V_{jk}^2 = 1 \quad (\forall j = 1, P) \qquad \sum_{k=1}^{P} V_{jk} V_{sk} = \delta_{js} \quad (\forall j, s = 1, \dots P)$$

new

$$y_{ij} = \sum_{k=1}^{P} V_{jk} Z_{ik}$$ ← old

$i = 1, \dots, N$

$j = 1, \dots, P$

Z → y Transform

old → new

The new base is not arbitrary we went That the data z_{ij} projected over the first vector of the new base HAVE MAX VARIANCE (max sum of residuals)

$$\max_{\text{on all bases}} \left(\frac{1}{N} \sum_{i=1}^{N} y_{1i}^2 \right)$$

y_{1i}: along direction 1

We start with X_{ij} ←

* Define the Covariance matrix of the Properties

$$C_{ik} = \frac{1}{N\sigma_j\sigma_k} \sum_{i,j=1}^{N} (X_{ij} - \mu_j)(X_{ik} - \mu_k)$$

$i,k = 1, \ldots, p$

$$= \frac{1}{N} \sum_{i=1}^{N} Z_{ij} \cdot Z_{ik} \leftarrow$$

The sample covariance matrix is the matrix of correlation coefficients

The matrix C_{ik} is real, symmetric

A theorem follows of linear algebra that I can always transform this type of matrices into diagonal form

$P \times P$

$$U \cdot \check{C} \cdot U^{-1} = \tilde{C} \quad | \cdot U^{-1}$$

$$\tilde{C} = \begin{pmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \ddots \\ & & & \lambda_p \end{pmatrix}$$

In general $\lambda_1, \lambda_2 \ldots \lambda_p$ (for generic U and U^{-1}) ARE NOT IN ORDER

IN PCA we require that $\lambda_1 > \lambda_2 >> \lambda_p$ (through permutation of rows and columns of U U^{-1})

The λ_s are called EIGENVALUES of C_{jk} and the new base vectors are called EIGENVECTORS In PCA they are ORDERED from max to min

$$\sum_{j=1}^{p} V_{nj} C_{jk} = \lambda_n V_{nk}$$

$n, j, k = 1, \ldots, p$

note if we had defined vectors as (VECTORS) $\begin{pmatrix} V_{n1} \\ V_{n2} \\ \vdots \\ V_{np} \end{pmatrix}$ we would have here $\sum_{j=1} C_{jk} V_{nj}$

NOTE Essential

The variance over the first (ordered) direction in the PCA space is

$$\frac{1}{N}\sum_{i=1}^{N} y_{i1}^2 = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{p}\sum_{k=1}^{p} V_{1j} z_{ij} V_{1k} z_{ik}$$

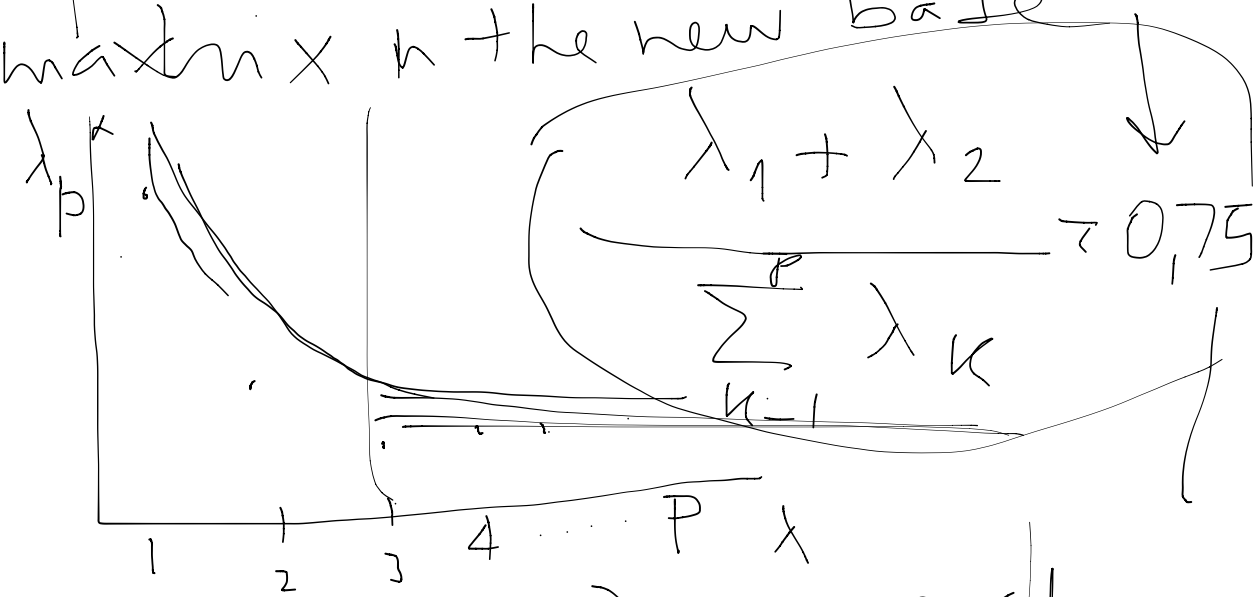
$$= \sum_{j=1}^{p}\sum_{k=1}^{p} V_{1j} C_{jk} V_{1k} = \sum_{k=1}^{p} \lambda_1 V_{1k}^2$$

$$= \lambda_1 (1)$$

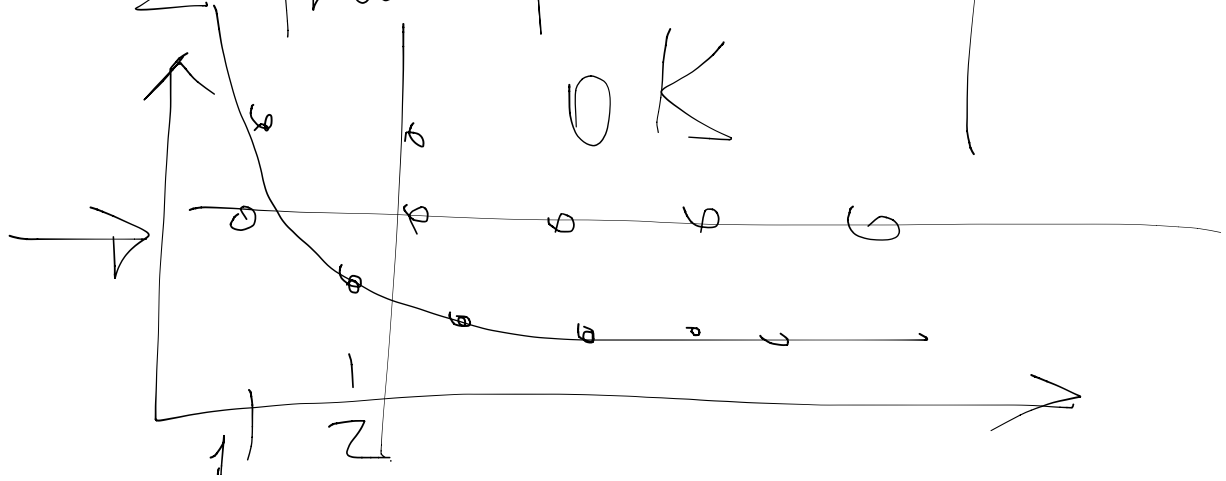
The sum of the eigenvectors is then the total variance seen from the new base

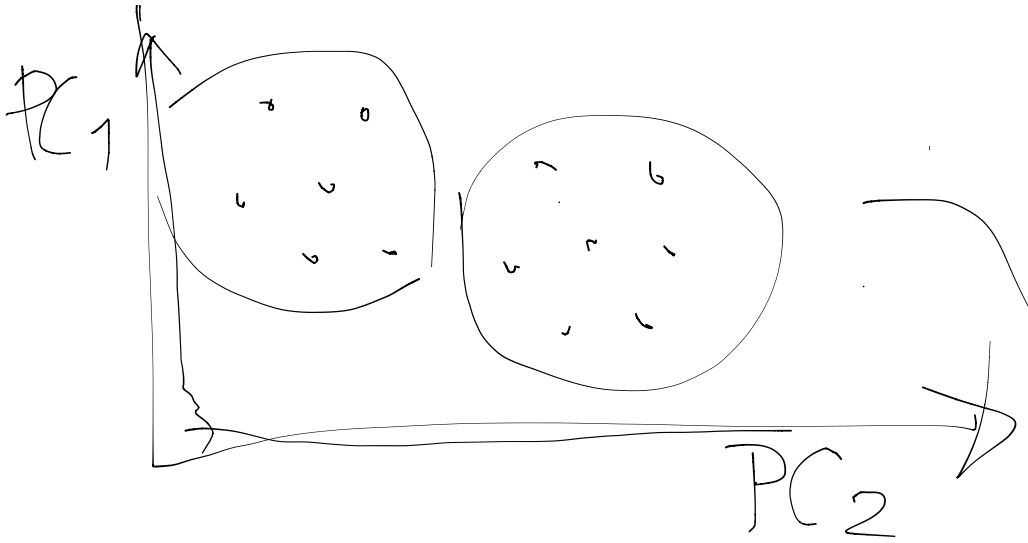
$$= \lambda_1 + \lambda_2 + \dots + \lambda_p$$

A posteriori, if the 'spectrum' of the Covariance matrix in the new base



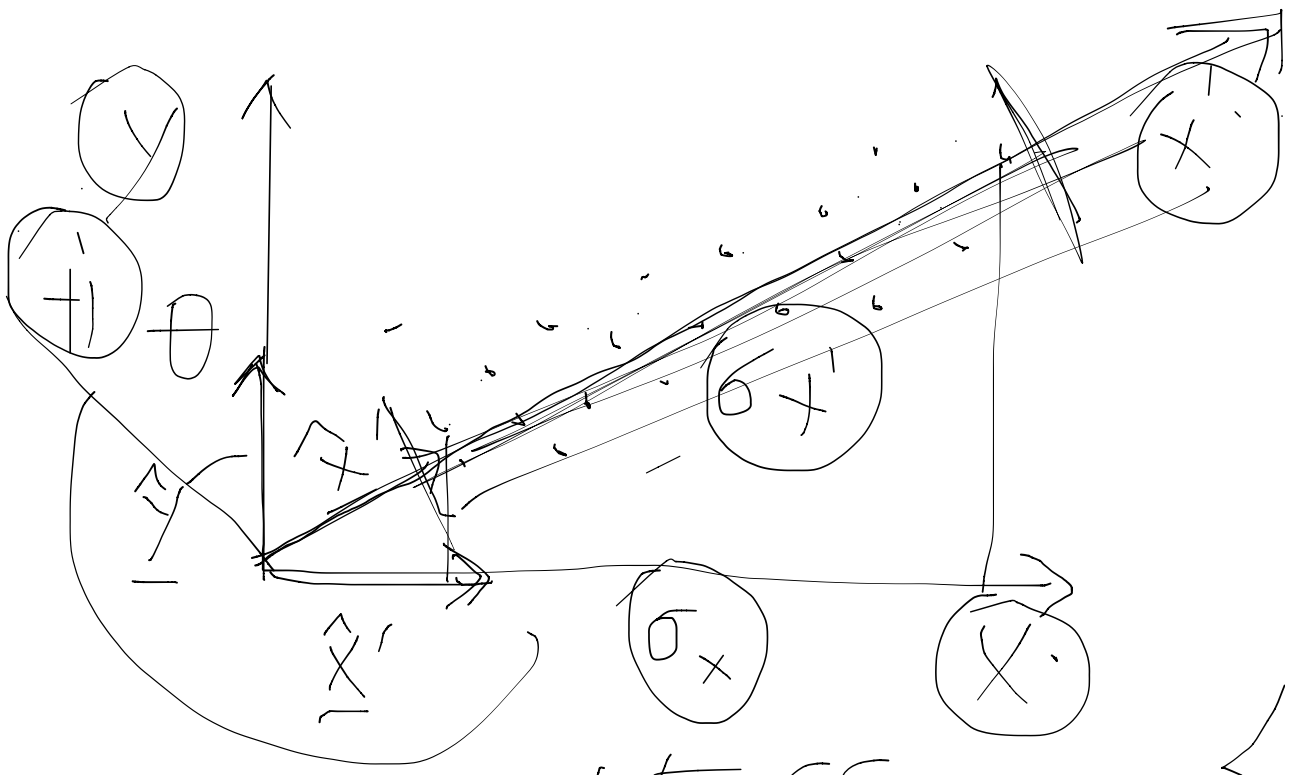
we get a Dimensional reduction we can take only the first 2 Principal directions





Possible natural clustering !

each point in this plane is the scalar product of the original $\underline{x}_i$ vectors $\underline{V}_1$ and V_2



PCA IS LESS DIMENSIONS BUT MORE VARIABIL.

Concentrated varia bility