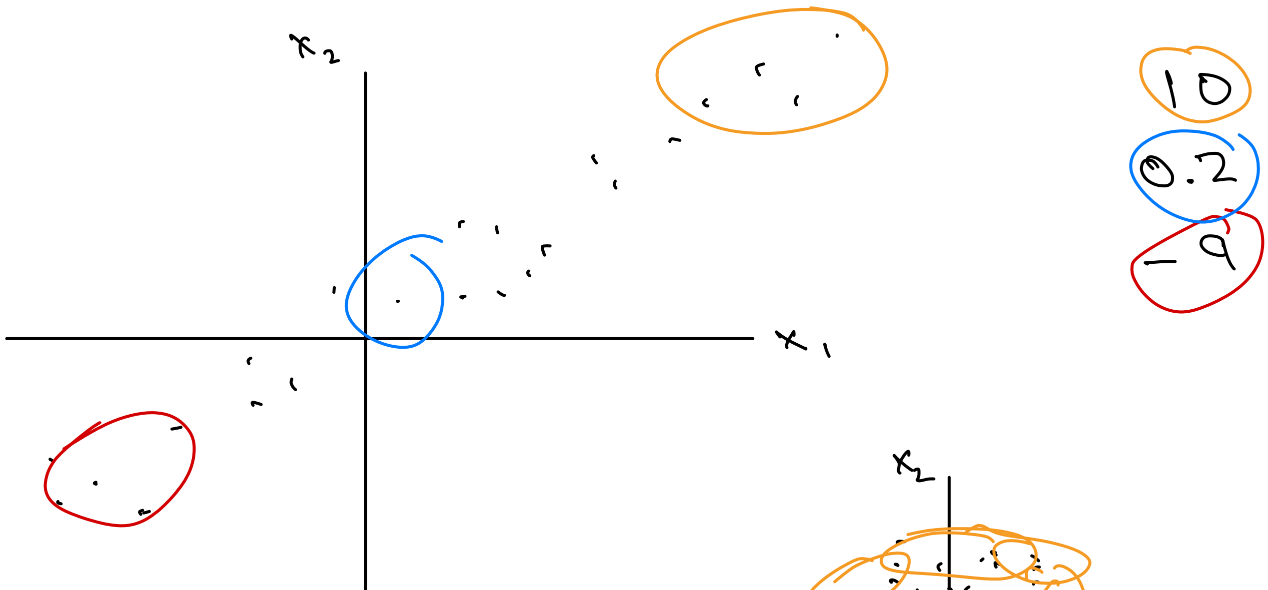
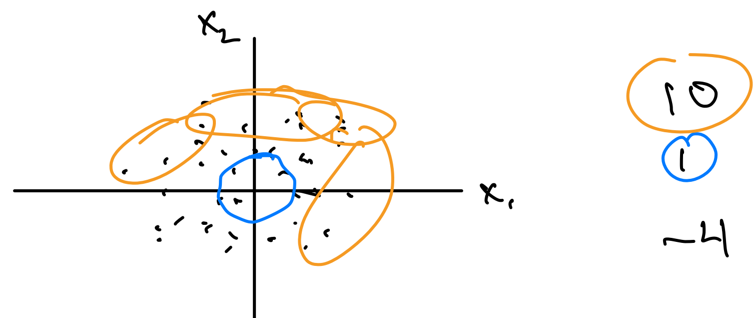
11 Principal Components Analysis

PCA is an unsupervised learning technique - there is no response.



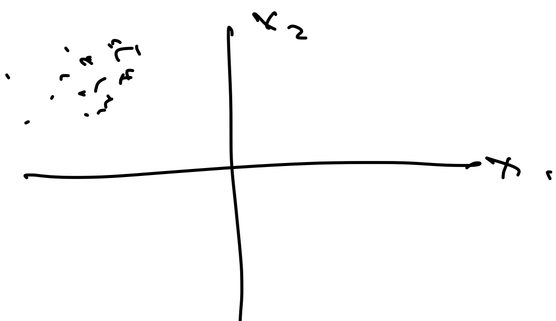


Assumptions

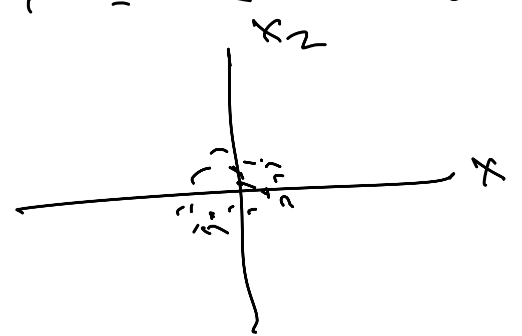
Have $\underline{n}$ observations of $\underline{p}$ variables, $\underline{x}_1, \dots, \underline{x}_n$

$$\underline{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} = \left[\begin{pmatrix} \text{course overall} \\ \text{instr overall} \\ \text{Hrs/week spent on HW} \end{pmatrix} \right]$$

Moreover, assume $\underline{x}_i \leftarrow \underline{x}_i - \bar{\underline{x}}$ are centered



$\rightsquigarrow$



Recall

If M is a symmetric real $p \times p$ matrix its <u>eigendecomposition</u> (or <u>spectral decomposition</u>) is

$$M = A^T D A$$

where A^T is $p \times p$ matrix whose columns are eigenvectors of M & D is diag matrix of eigenvalues.

$$A^T = [\underline{a}_1 \ \underline{a}_2 \cdots \underline{a}_p] \Rightarrow A = \begin{bmatrix} \underline{a}_1^T \\ \underline{a}_2^T \\ \vdots \\ \underline{a}_p^T \end{bmatrix}$$

eigenvectors

→ rows are e.vecs.

Use convention that $\underline{a}_1, \dots, \underline{a}_p$ are normalized so A is orthogonal matrix $A^T A = A A^T = I$. A diagonalizes M:

$$A M A^T = D.$$

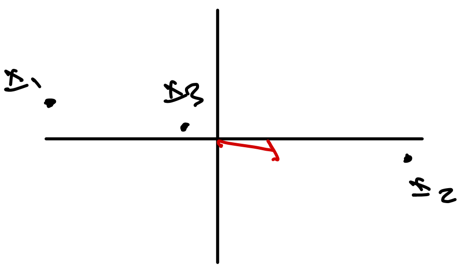
Ex $\quad n=3,\ p=2$

standard basis vectors

$$\underline{x}_1 = \begin{pmatrix} -10 \\ 1 \end{pmatrix} = -10\,\underline{e}_1 + 1\cdot\underline{e}_2 \approx -10.5\,\underline{a}_1$$

$$\underline{x}_2 = \begin{pmatrix} 12 \\ -1/2 \end{pmatrix} = 12\,\underline{e}_1 - \tfrac{1}{2}\,\underline{e}_2 \approx 12.2\,\underline{a}_1$$

$$\underline{x}_3 = \begin{pmatrix} -2 \\ 1/2 \end{pmatrix} = -2\,\underline{e}_1 + \tfrac{1}{2}\,\underline{e}_2 \approx -1.8\,\underline{a}_1$$



Basic idea behind PCA:

$$\underline{x} = b_1\underline{e}_1 + b_2\underline{e}_2 + \cdots + b_p\underline{e}_p$$

re-express as:

$$\underline{x} = \underbrace{a_1}_{\text{biggest}}\underline{a}_1 + \underbrace{a_2}_{\text{next biggest}}\underline{a}_2 + \cdots + \underbrace{a_p}_{\text{smallest}}\underline{a}_p$$

11.1 Probabilistic Approach

$\boxed{\text{Ex}}$ Suppose $\underline{x}$ is a random vector (mean zero) with cov. matrix Σ, want matrix A that decorrelates $\underline{x}$:

$$Var(A\underline{x}) = [\text{want to be}] = D = \text{diagonal matrix.}$$

$\underline{z} = A\underline{x}$ are uncorrelated.

$$Var(A\underline{x}) = A\Sigma A^T = D$$

[Σ is real, $p \times p$, symmetric] $\Leftrightarrow$ Using A from the eigen decomp of Σ will be the solution. If rows of A are eigvecs of Σ, then

$$\Sigma = A^T D A \quad + \quad A\Sigma A^T = D \quad + D \text{ holds eigenvalues of } \Sigma.$$

$\boxed{\text{Note}}$ Same argument for finite sample version.

$\underline{x}_1, \dots, \underline{x}_n$ over p features, need a convention for storing in a matrix:

$$\underset{n\times p}{X} = \begin{bmatrix} \underline{x}_1^T \\ \underline{x}_2^T \\ \vdots \\ \underline{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

(usual design matrix w/out column of 1s)

$$X^T = [\underline{x}_1\ \underline{x}_2 \cdots \underline{x}_n] \quad p\times n.$$

$\boxed{\text{Def}}$ The <u>sample covariance matrix</u> of X is

$$\hat{\Sigma} = \frac{1}{n-1} X^T X \quad (p\times p \text{ matrix})$$

= covariance matrix of "old features"

Want: A set of "new features" $\underline{z}_i = A\underline{x}_i$ that are uncorrelated + sorted in order of decreasing variability.

<u>Warning</u>: Not saying $\underline{z}_i$ + $\underline{z}_j$ to be uncorrelated

<u>Are saying</u>: elements in $\underline{z}_i = \begin{pmatrix} z_{i1} \\ z_{i2} \\ \vdots \\ z_{ip} \end{pmatrix}$ to be uncorrelated. (Act like p uncorr. random variables).

Store $\underline{z}_1, \ldots, \underline{z}_n$ in matrix

$$Z^T = [\underline{z}_1\ \underline{z}_2 \cdots \underline{z}_n]$$

$$Z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & & z_{2p} \\ \vdots & \vdots & & \vdots \\ z_{n1} & z_{n2} & & z_{np} \end{bmatrix}$$ = $n \times p$ design matrix of p "new features"

$$Z^T = [\underline{z}_1\ \underline{z}_2 \cdots \underline{z}_n] = [A\underline{x}_1\ \ A\underline{x}_2 \ \cdots\ A\underline{x}_n] = A X^T$$

want : cov matrix of $\underline{z}$ to be diagonal

$$\text{Diag} = \hat{\Sigma}_z = \frac{1}{n-1} Z^T Z = \frac{1}{n-1}(A X^T)(X A^T)$$

$$= \frac{1}{n-1} A X^T X A^T = A\left(\frac{1}{n-1} X^T X\right) A^T = A \hat{\Sigma} A^T$$

$\Rightarrow$ Set rows of A to be eigenvectors $\hat{\Sigma}$

If you do this, then $\hat{\Sigma}_z$ = Diag with eigenvalues of $\hat{\Sigma}$

$$\hat{\Sigma}_z = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$

- λ_1 is sample variance of "new features"

$$z_{11}, z_{21}, \dots, z_{n1}$$

- λ_j is sample variance of jth "new feature"

$$z_{1j}, z_{2j}, \ldots, z_{nj}$$

PCA

$$\underline{x}_i = x_{i1}\underline{e}_1 + x_{i2}\underline{e}_2 + \cdots + x_{ip}\underline{e}_p$$

$$= z_{i1}\underline{a}_1 + z_{i2}\underline{a}_2 + \cdots + z_{ip}\underline{a}_p$$

eigenvectors

principal components ("new features")

= scores = component scores

= "loadings"

$(z_{i1}\underline{a}_1)$

= loading