Principal Components Analysis:

Suppose I observe n observations of p features:

$$\{(x_{i1}, x_{i2}, \dots, x_{ip})\}_{i=1}^{n}$$

$\downarrow$ design matrix

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & & & \\ x_{n1} & \cdots & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} — & x_1^T & — \\ & \vdots & \\ — & x_i^T & — \\ & \vdots & \\ — & x_n^T & — \end{bmatrix}$$

Let $\bar{x}_j$ denote the j^{th} column mean:

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

$$\bar{x} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)$$

How to perform PCA:

① Center each column vector to create the centered matrix:

$$X^c = \{x_{ij}^c\}_{ij} \text{ where } x_{ij}^c = x_{ij} - \bar{x}_j$$

Visually this corresponds to moving the origin (the mean point of the data cloud).

(Optional: Scale the variance of each feature down to 1.)

② Calculate the sample covariance matrix:

$$\underset{p\times p}{S} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$$

$$= \frac{1}{n-1} X^{c^T} X^c$$

Notice: S is a symmetric, positive semi-def. matrix.

① $S^T = S$ ← symmetry

② $a \in \mathbb{R}$

$a \geq 0$

$ax^2 \geq 0$

$x a x \geq 0$

$x^T S x \geq 0 \quad \forall\ x \in \mathbb{R}^p$

$1{\times}p \quad p{\times}p \quad p{\times}1$

Equivalently:
The eigenvalues of S are all ≥ 0.

③ Calculate the eigendecomposition of S:

$$S = \Phi \Lambda \Phi^T$$

Φ is orthonormal

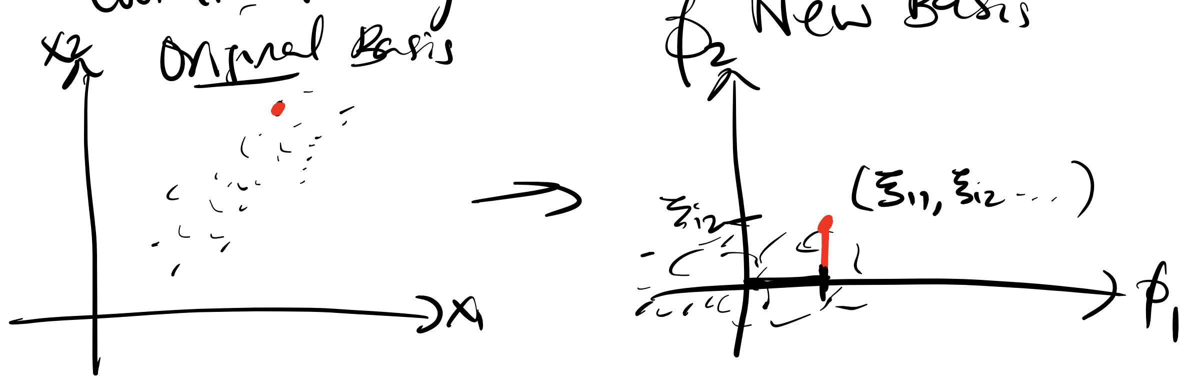
$\Phi^T \Phi = I$

Λ is a diagonal matrix

$\Lambda = \mathrm{diag}(\underset{\geq}{\lambda_1}, \underset{\geq}{\lambda_2}, \underset{\geq}{\lambda_3}, \dots, \lambda_p)$

$$\Phi = \begin{bmatrix} | & | & & | \\ \phi_1 & \phi_2 & \cdots & \phi_p \\ | & | & & | \end{bmatrix}$$

The columns of Φ give us the unit vectors in the direction of our new coordinate system.



What are the coordinates of our data points in the new basis?

Facts:

① The coordinates of the i^{th} datapoint in the new basis are given by the inner product of the i^{th} row of X^c & the directions $\phi_1, \phi_2, \ldots, \phi_p$.

$$\xi_{i1} = \langle x_i^c, \phi_1 \rangle = \langle x_i - \bar{x}, \phi_1 \rangle$$

$$\vdots$$

$$\xi_{ip} = \langle x_i^c, \phi_p \rangle = \langle x_i - \bar{x}, \phi_p \rangle$$

These are called the principal component "scores".

We can re-express the original data in terms of the PC scores & the new basis:

$$x_i = \bar{x} + \xi_{i1}\phi_1 + \xi_{i2}\phi_2 + \cdots + \xi_{ip}\phi_p$$

2. The sample variance of the jth PC score is equal to the jth eigenvalue of S:

$$\underline{\frac{1}{n-1}\sum_{i=1}^{n} \xi_{ij}^2 = \lambda_j} \quad \forall j = 1, \ldots, p$$

Exercise:

A. Show that $\bar{\xi}_j = \frac{1}{n}\sum_{i=1}^{n} \xi_{ij} = 0$

B. Show that $\frac{1}{n-1}\sum_{i=1}^{n} \xi_{ij}^2 = \lambda_j \quad \forall j.$

A. $\xi_{ij} = \langle x_i - \bar{x}, \phi_j \rangle$

$$= (x_i - x)^T \phi_j$$

$$= x_i^{cT} \phi_j$$

$$\frac{1}{n} \sum_{i=1}^n x_i^{cT} \phi_j = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^T \phi_j$$

$$= \frac{1}{n} \left(\sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x} \right)^T \phi_j$$

$$= \frac{1}{n} \left(\sum_{i=1}^n x_i - n\bar{x} \right)^T \phi_j$$

$$= \frac{1}{n} (0)^T \phi_j = 0$$

B. $\frac{1}{n-1} \sum_{i=1}^n \xi_{ij}^2 = \frac{1}{n-1} \sum_{i=1}^n \left(x_i^{cT} \phi_j \right)^2$

$$= \frac{1}{n-1} \sum_{i=1}^n \left(\phi_j^T x_i^c \cdot x_i^{cT} \phi_j \right)$$

$$= \phi_j^T \left(\frac{1}{n-1} \sum_{i=1}^n x_i^c x_i^{cT} \right) \phi_j$$

$$= \phi_j^T \left(\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T \right) \phi_j$$

$$= \phi_j^T S \phi_j$$

① $$= \phi_j^T (\Phi \Lambda \Phi^T) \phi_j$$

$$= \begin{pmatrix}0\\ \vdots\\ 1\\ \vdots\\ 0\end{pmatrix}^T \begin{pmatrix}\lambda_1 & & & 0\\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_p\end{pmatrix}\begin{pmatrix}0\\ \vdots\\ 1\\ \vdots\\ 0\end{pmatrix}$$

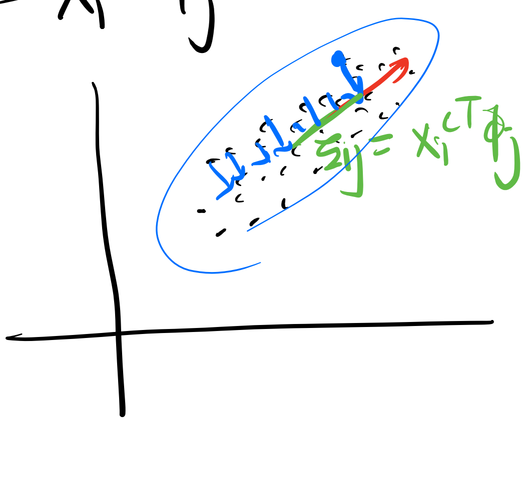
$$= \lambda_j$$

② $$= \phi_j^T S \phi_j$$

$$= \phi_j^T (\lambda_j \phi_j)$$

$$= \lambda_j \underbrace{\phi_j^T \phi_j}_{} = \lambda_j$$

$$\xi_{ij} = x_i^{cT} \phi_j$$



③ If we do SVD on $X^c = UDV^T$ then the columns of V are exactly the same as the eigenvectors of S.

i.e. $\phi_j = v_j \quad \forall j = 1, \dots, p$

$$X^c = UDV^T$$

$$S = \frac{1}{n-1} X^{cT} X^c$$

$$= \frac{1}{n-1} (UDV^T)^T (UDV^T)$$

$$= \frac{1}{n-1} (VDU^TUDV^T)$$

$$= \frac{1}{n-1} VD^2V^T$$

$$= V\left(\frac{1}{n-1}D^2\right)V^T$$

exact same structure as $\Phi \Lambda \Phi^T$

④ Because the spectral decomposition of S arranges the eigenvalues in a decreasing order, the basis vector $\underline{\phi_1}$ which corresponds to λ_1 (the biggest eigenvalue) & is the direction in which the data varies the most, followed by ϕ_2, then ϕ_3, & so on.

⑤ The "total variation" in the data is the trace of S:

$$\sum_{j=1}^{p}\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2\right) = tr(S)$$

sum of diag elements

$$tr(S) = tr(\Phi \Lambda \Phi^T) = tr(\Lambda \Phi^T \Phi)$$
$$= tr(\Lambda)$$
$$= \lambda_1 + \lambda_2 + \cdots + \lambda_p$$

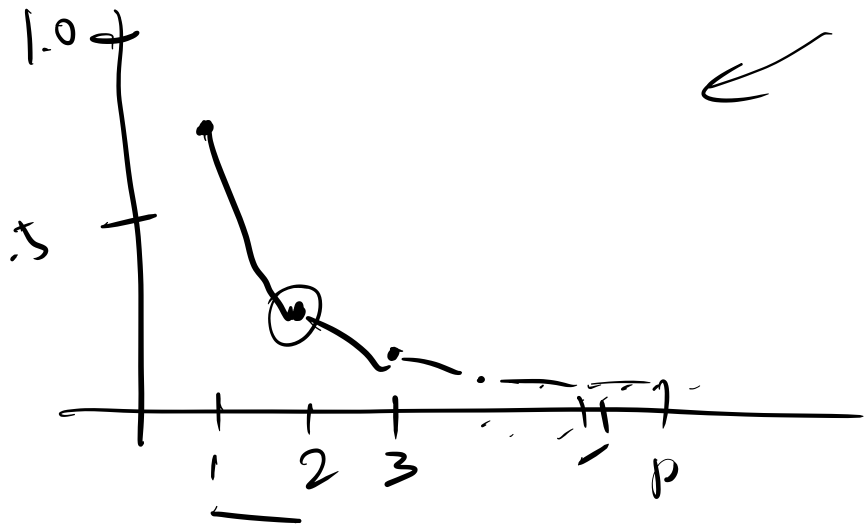
The ratio $\dfrac{\lambda_k}{tr(S)} = \dfrac{\lambda_k}{\lambda_1+\lambda_2+\cdots+\lambda_p}$

the total variation in the data that is

explained by the k^{th} principal component (direction).

$\mathrm{tr}(ABC) = \mathrm{tr}(BCA) = \mathrm{tr}(CAB)$
"cyclic"

Often we plot $\lambda_k / \mathrm{tr}(S)$ against k to see how the eigenvals decay.



We can use the Scree plot to choose a good k to truncate the basis expansion of x_i without losing much info.

$$x_i = \bar{x} + \xi_{i1}\phi_1 + \xi_{i2}\phi_2 + \cdots + \xi_{ip}\phi_p$$

$$= \bar{x} + \xi_{ij}\phi_j + \xi_{i2}\phi_2 + \cdots + \phi_{ik}\phi_k$$

$$= \bar{x} + \sum_{j=1}^{k} \xi_{ij}\phi_j$$