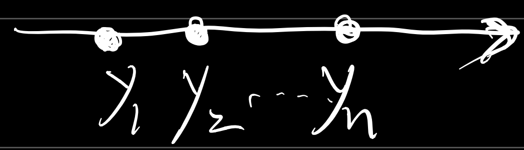
Recall: (1) For 1D sample points $y_1, \cdots, y_n \in \mathbb{R}$
their <u>sample mean</u> is

$$\mu \overset{\text{def}}{=} \frac{1}{n} \sum_{j=1}^{n} y_j ,$$

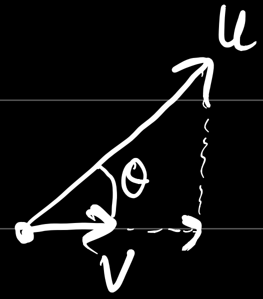


their <u>sample variance</u> is

$$\frac{1}{n} \sum_{j=1}^{n} (y_j - \mu)^2$$

(2) Given a unit vector $v \in S^{n-1}$,
$(S^{n-1} = \{ x \in \mathbb{R}^n : \|x\| = 1 \})$
the projection of $u \in \mathbb{R}^n$ onto v is

$$\mathrm{Proj}_v u = (\|u\| \cos\theta)\, v$$

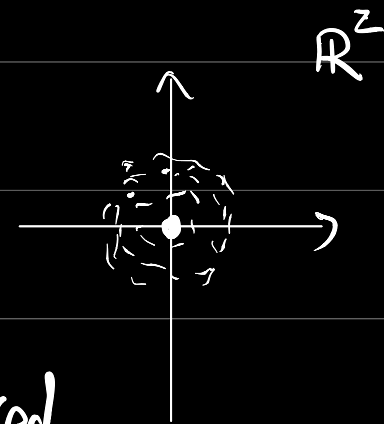


$$\overset{\|v\|=1}{=} (\underbrace{\|u\|\, \|v\| \cos\theta}_{u \cdot v})\, v$$

$$= (u^T v)\, v$$

3.1.3. Theory:

Let $x^{(1)}, \cdots, x^{(n)} \in \mathbb{R}^p$ be centered sample points, i.e.



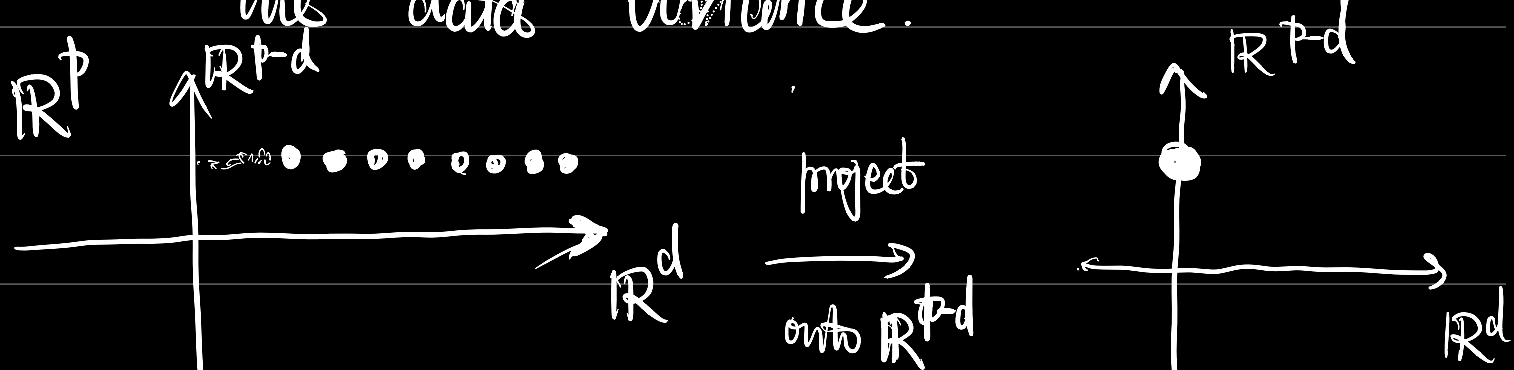
$$\mu \overset{\text{def}}{=} \frac{1}{n} \sum_{j=1}^{n} x^{(j)} = 0$$

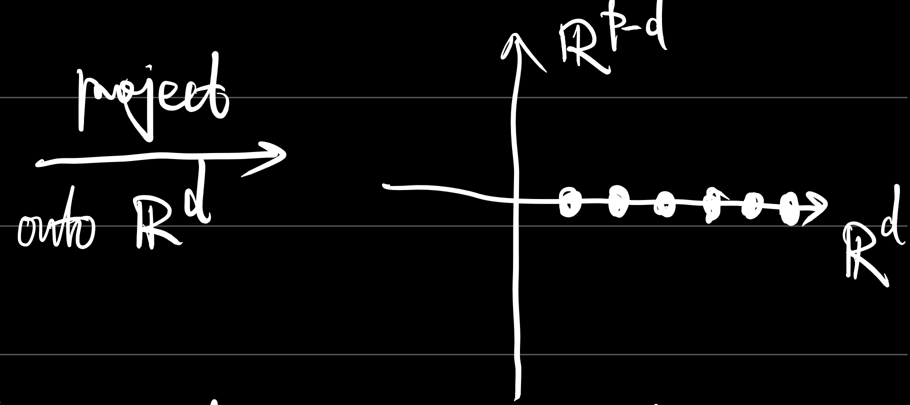
(If $x^{(1)}, \cdots, x^{(n)}$ are not centered, replace $x^{(j)}$ by $x^{(j)} - \mu$, then

$$\frac{1}{n} \sum_{j=1}^{n} (x^{(j)} - \mu) = \frac{1}{n} \sum_{j=1}^{n} x^{(j)} - \frac{1}{n} \sum_{j=1}^{n} \mu$$

$$= \mu - \frac{1}{n} n \cdot \mu = 0 \;)$$

Q in PCA: find low-dimensional repre of the sample points that maximizes the data variance.

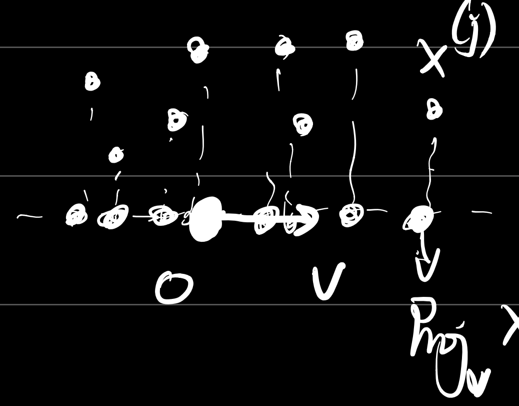




i.e. find directions of projection that yield maximum data variance.

The projected sample points onto a unit vector $v \in S^{n-1}$ is

$$\mathrm{Proj}_v x^{(1)} = (x^{(1)T} v) v, \; \dots, \; \mathrm{Proj}_v x^{(n)} = (x^{(n)T} v) v$$



mean of the projection

$$= \frac{1}{n} \sum_{j=1}^{n} \mathrm{Proj}_v x^{(j)}$$

$$= \frac{1}{n} \left[\sum_{j=1}^{n} \underbrace{\left(x^{(j)T} v \right)}_{\text{magnitude of projections}} \right] v = \left[\Bigg(\underbrace{\frac{1}{n} \sum_{j=1}^{n} x^{(j)}}_{} \Bigg)^T v \right] v = 0$$

$\mu = 0$

variance of the projections

$$\overset{\text{def of variance}}{=} \frac{1}{n}\sum_{j=1}^{n} \| \mathrm{Proj}_v x^{(j)} - 0 \|^2 = \frac{1}{n}\sum_{j=1}^{n} \| \mathrm{Proj}_v x^{(j)} \|^2$$

$$\overset{\text{def of proj}}{=} \frac{1}{n}\sum_{j=1}^{n} \| \underbrace{(x^{(j)T} v)}_{\text{scalar}} v \|^2 \overset{\|v\|=1}{=} \frac{1}{n}\sum_{j=1}^{n} (x^{(j)T} v)^2$$

$$= \frac{1}{n}\sum_{j=1}^{n} \underbrace{(x^{(j)T} v)^T}_{\text{scalar}} (x^{(j)T} v)$$

$$= \frac{1}{n}\sum_{j=1}^{n} v^T x^{(j)} x^{(j)T} v$$

$$= v^T \underbrace{\left(\frac{1}{n}\sum_{j=1}^{n} x^{(j)} x^{(j)T} \right)}_{\overset{\text{def}}{=} C} v$$

$$= v^T C v$$

where the (k,l)-entry of C is

$$C_{kl} = \frac{1}{n}\sum_{j=1}^{n} \left(x^{(j)} x^{(j)T} \right)_{kl} = \frac{1}{n}\sum_{j=1}^{n} x_k^{(j)} x_l^{(j)}$$

$$\overset{\circledast}{=} \frac{1}{n}(XX^T)_{kl}$$

where $X = \begin{bmatrix} | & | & & | \\ x^{(1)}, & x^{(2)}, & \cdots, & x^{(n)} \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{p\times n}$

is the sample matrix.

$\circledast$ is because

$$XX^T = \begin{pmatrix} | & & | \\ x^{(1)} & \cdots & x^{(n)} \\ | & & | \end{pmatrix} \begin{pmatrix} — & x^{(1)} & — \\ & \vdots & \\ — & x^{(n)} & — \end{pmatrix}$$

kth row: $\qquad$ l th column

$$(XX^T)_{kl} = (x_k^{(1)}, \cdots, x_k^{(n)}) \cdot \begin{pmatrix} x_l^{(1)} \\ \vdots \\ x_l^{(n)} \end{pmatrix} = x_k^{(1)} x_l^{(1)} + \cdots + x_k^{(n)} x_l^{(n)}$$

Putting together, variance of the projections

$$= v^T C v = v^T \left(\frac{1}{n} XX^T\right) v = \frac{1}{n} v^T (XX^T) v$$

Here $C^T = \left(\frac{1}{n} XX^T\right)^T = C$ is symmetric.

thus the projected variance is the Rayleigh quotient $v^T C v$, which is maximized if v is an eigenvector associated to the largest eigenvalue of C, or equivalently, of XX^T.