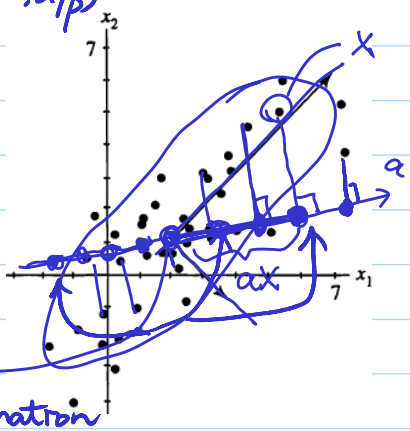
Principal Component Analysis (PCA)

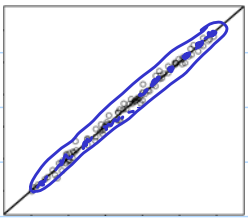
- PCA belongs to the class of *projection methods*, which choose one or few linear combinations of the original p variables to *maximize* some measure of "*interestingness*."

$$Y_1 = \mathbf{a}_1'\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$$

$$\mathbf{X}_{(n\times p)} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \longrightarrow \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1k} \\ y_{21} & y_{22} & \cdots & y_{2k} \\ \cdots & \cdots & & \cdots \\ y_{n1} & y_{n2} & \cdots & y_{nk} \end{bmatrix}, \quad k \ll p$$



- In PCA, the goal is to reduce the *dimensionality* (**Q**: why?) of a data set comprised of a large number of interrelated variable, while retaining as much as possible the *variation* (**Q**: why?) present in the data



Q: 1-dim? or 2-dim?