Principal Component Analysis.

Dimensionality Reduction

Latent variable Modeling
(Hidden)

$x_i \rightarrow$ one data instance $\quad x_i \in R^D$

$z_i \rightarrow$ hidden / latent unknown variable.

If $z_i \in \{1, \ldots, K\}$
$\hookrightarrow$ clustering
$\hookrightarrow$ Mixture of models

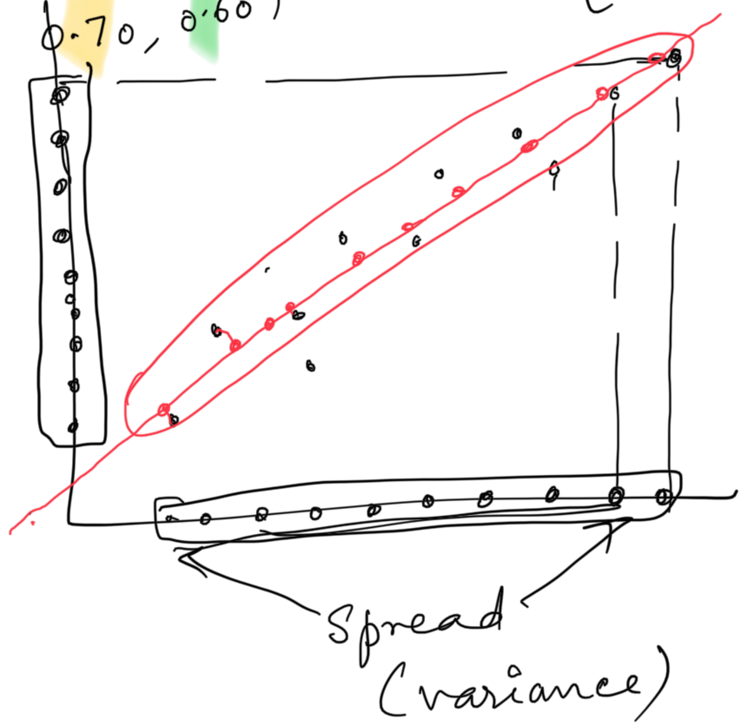
If $z_i \in R^d$
where $d << D$
$\hookrightarrow$ This is a dimensionality reduction problem.

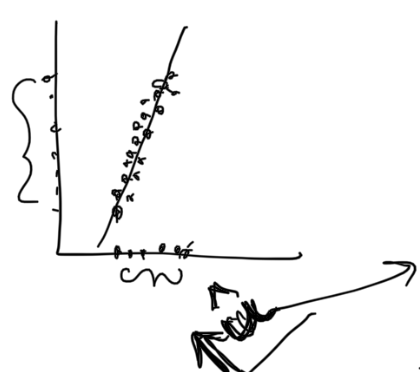
Given X: $\begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \end{bmatrix}$ find Z: $\begin{bmatrix} z_1^T \\ z_2^T \\ \vdots \end{bmatrix}$

$\left[\; x_N^T \;\right] \qquad \left[\; z_N \;\right]$

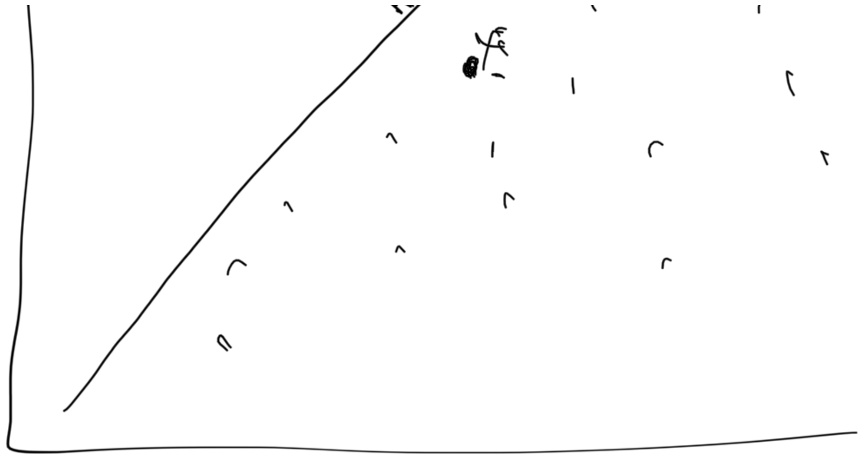
Such that ~~some~~ some property is preserved.

$$X = \begin{bmatrix} 0.20, & 0.10 \\ 0.35, & 0.40 \\ 0.50, & 0.20 \\ 0.65, & 0.10 \\ 0.70, & 0.60 \end{bmatrix} \qquad Z = \begin{bmatrix} ? \end{bmatrix}$$





unit vector



$x_i^T \hat{u}$ — Projection of i^{th} data instance

$$Z = \begin{bmatrix} x_1^T \hat{u} \\ x_2^T \hat{u} \\ \vdots \\ x_N^T \hat{u} \end{bmatrix}$$

Let us assume that the data has 0 mean.

$$\Rightarrow \sum_{i=1}^{N} x_i = 0$$

$$\text{var: } \frac{1}{N} \sum_{i=1}^{N} (x_i^T \hat{u})^2$$

$$= \frac{1}{N} \sum_{i=1} \underline{\hat{u}}^T x_i x_i^T \underline{u}$$

$$= \hat{u}^T \underbrace{\left(\frac{1}{N} \sum_{i=1}^{N} x_i x_i^T \right)} \hat{u}$$

This is the sample covariance matrix of X

$$S = \frac{1}{N} \sum_{i=1}^{N} (x_i - \underline{\bar{x}})(x_i - \underline{\bar{x}})^T$$

$$= \frac{1}{N} \sum_{i=1}^{N} x_i x_i^T \quad [\text{if } \bar{x} \text{ is } 0]$$

$$\text{var} = \underline{\hat{u}^T S \hat{u}}$$

$$\boxed{\begin{aligned} &\arg\max_{\hat{u}} \ \hat{u}^T \mathbf{S} \hat{u} \\ &\text{subject to } \ \underline{\hat{u}^T \hat{u} = 1} \end{aligned}}$$

$\hat{u}$ should be unit length

$$\boxed{\hat{u}^T S \hat{u} + \lambda (\hat{u}^T \hat{u} - 1)}$$