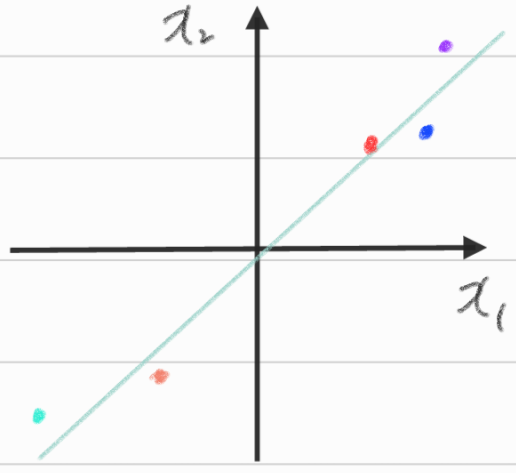
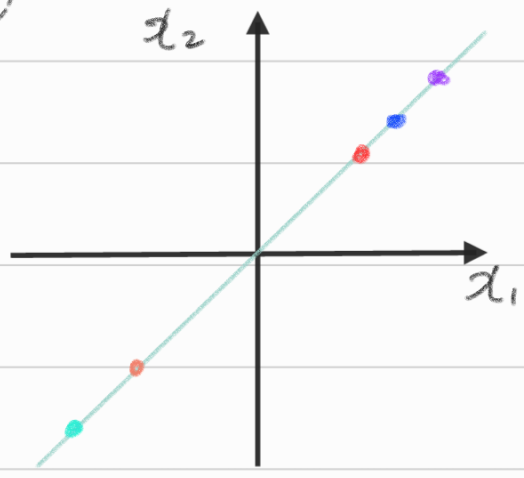
Reconstruction from Compressed Representation

$x^{(i)} \rightarrow z^{(i)} \rightarrow x^{(i)}$

$z = U_{reduce}^{T} \cdot x$ (U_{reduce}^{T}: $k \times n$) $\quad z \in \mathbb{R}^k$

$x_{근사값} = U_{reduce} \cdot z$ ($n \times 1$ = $n \times k$ · $k \times 1$) $\quad x \in \mathbb{R}^n$

① 

③ 

② $z = U_{reduce}^{T} x$

① → ② → ③

Choosing the Number of Principal Components

Average squared projection error - Minimize

Total variation in the data (from origin how far)

Typically, choose k to be smallest value so that

$$\frac{\text{Average squared projection error}}{\text{Total variation in the data}} \leq 0.01 \ (1\%)$$

$$\frac{\frac{1}{m}\sum_{i=1}^{m}\|x^{(i)} - x_{approx}^{(i)}\|^2}{\frac{1}{m}\sum_{i=1}^{m}\|x^{(i)}\|^2}$$

99% of variance is retained

95%

[U, S, V] 에서 $S = \begin{bmatrix} S_{11} & & 0 \\ & S_{22} & \\ 0 & & \ddots \\ & & & S_{nn} \end{bmatrix}$ (Covariance matrix)

For given k,

$$1 - \frac{\sum_{i=1}^{k} S_{ii}}{\sum_{i=1}^{n} S_{ii}} \leq 0.01$$

Advice for Applying PCA

지도학습)

Extract inputs:

Unlabeled dataset $x^{(m)} \in \mathbb{R}^{10000}$

$z^{(m)} \in \mathbb{R}^{1000}$

$(x^{(m)}, y^{(m)}) \rightarrow (z^{(m)}, y^{(m)})$

When running PCA, run PCA only on the training set portion

PCA 적용

- Compression : memory reduce, speed up algorithm
- Visualization : 1~3차원
* Overfitting 막아주기도 하지만 bad use,
  Regularization (규제) 사용권장