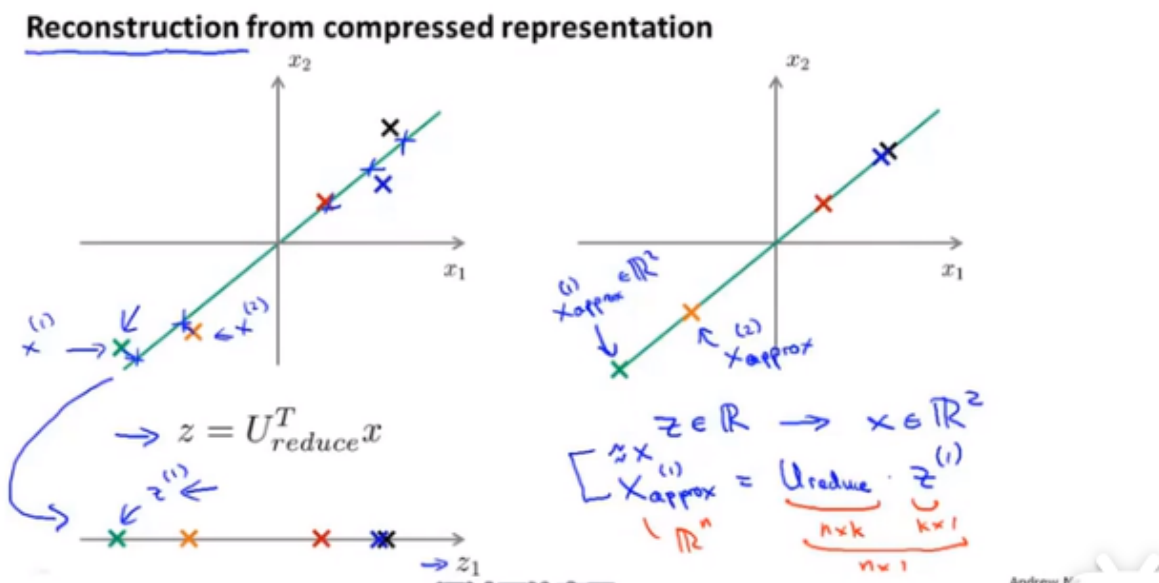


PCA压缩重现



原本n维向量的数据样本，经过公式到了一维向量z，现在利用公式回归得到原本的n维向量x的近似表示，这就叫原始数据的重建。

如何选择PCA中的维度k

选择合适的k要达成的目的：

Choosing k (number of principal components)

Average squared projection error: $\frac{1}{m} \sum_{i=1}^m \|x^{(i)} - x_{approx}^{(i)}\|^2$

Total variation in the data: $\frac{1}{m} \sum_{i=1}^m \|x^{(i)}\|^2$

Typically, choose k to be smallest value so that

$$\rightarrow \frac{\frac{1}{m} \sum_{i=1}^m \|x^{(i)} - x_{approx}^{(i)}\|^2}{\frac{1}{m} \sum_{i=1}^m \|x^{(i)}\|^2} \leq \frac{0.01}{0.05} \quad \frac{(1\%)}{(5\%)} \quad \frac{(10\%)}{(10\%)}$$

\rightarrow "99% of variance is retained"
~~95%~~ to 90%

其中一种选择k的方法是从1开始增加k，直到找到满足公式的选择。但调用svd函数可以直接得到S矩阵的值，可以根据它来直接调整k

Choosing k (number of principal components)

Algorithm:

Try PCA with $k=1$ ~~$k=2$~~ ~~$k=3$~~ ~~$k=4$~~

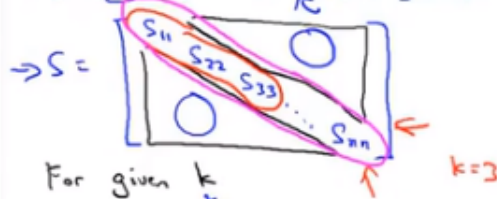
Compute $U_{reduce}, z^{(1)}, z^{(2)}, \dots, z^{(m)}, x_{approx}^{(1)}, \dots, x_{approx}^{(m)}$

Check if

$$\frac{\frac{1}{m} \sum_{i=1}^m \|x^{(i)} - x_{approx}^{(i)}\|^2}{\frac{1}{m} \sum_{i=1}^m \|x^{(i)}\|^2} \leq 0.01?$$

$k=17$

$$\rightarrow [U, S, V] = \text{svd}(\text{Sigma})$$



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

$$\rightarrow \frac{\sum_{i=1}^k S_{ii}}{\sum_{i=1}^n S_{ii}} \geq 0.99$$

Andrew K.

直接选择 k

Choosing k (number of principal components)

$$\rightarrow [U, S, V] = \text{svd}(\text{Sigma})$$

Pick smallest value of k for which

$$\frac{\sum_{i=1}^k S_{ii}}{\sum_{i=1}^m S_{ii}} \geq 0.99$$

$k=100$

(99% of variance retained)