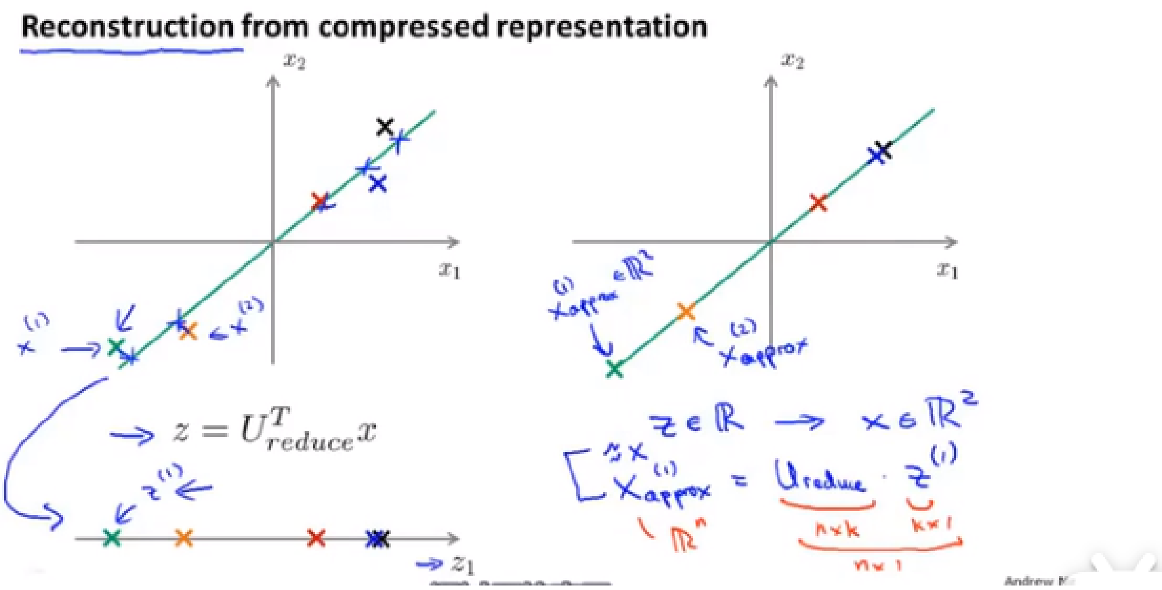
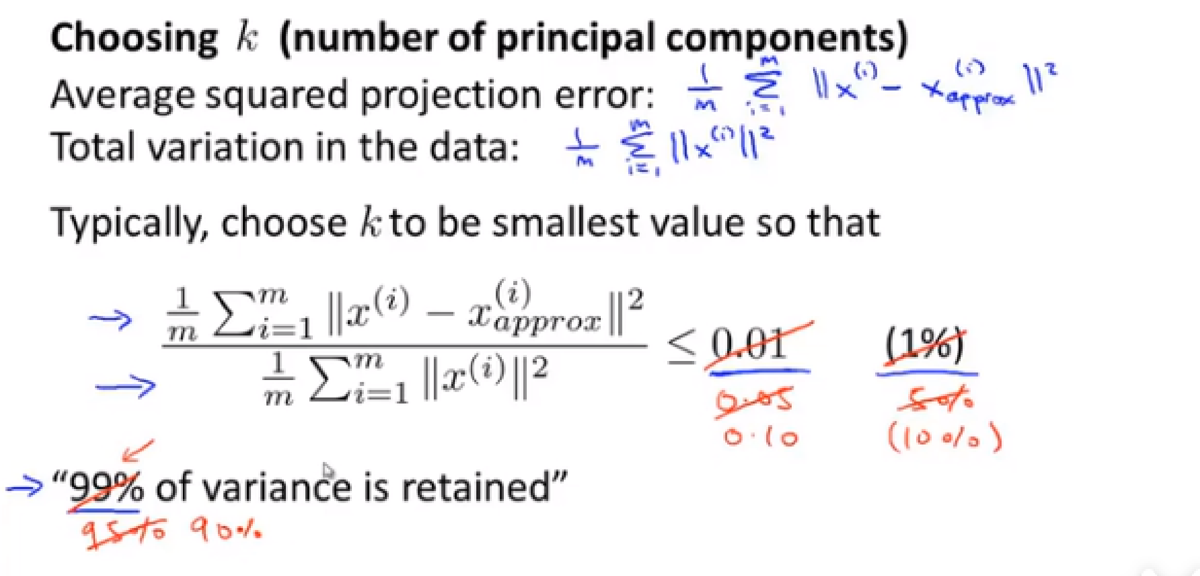
PCA压缩重现



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。

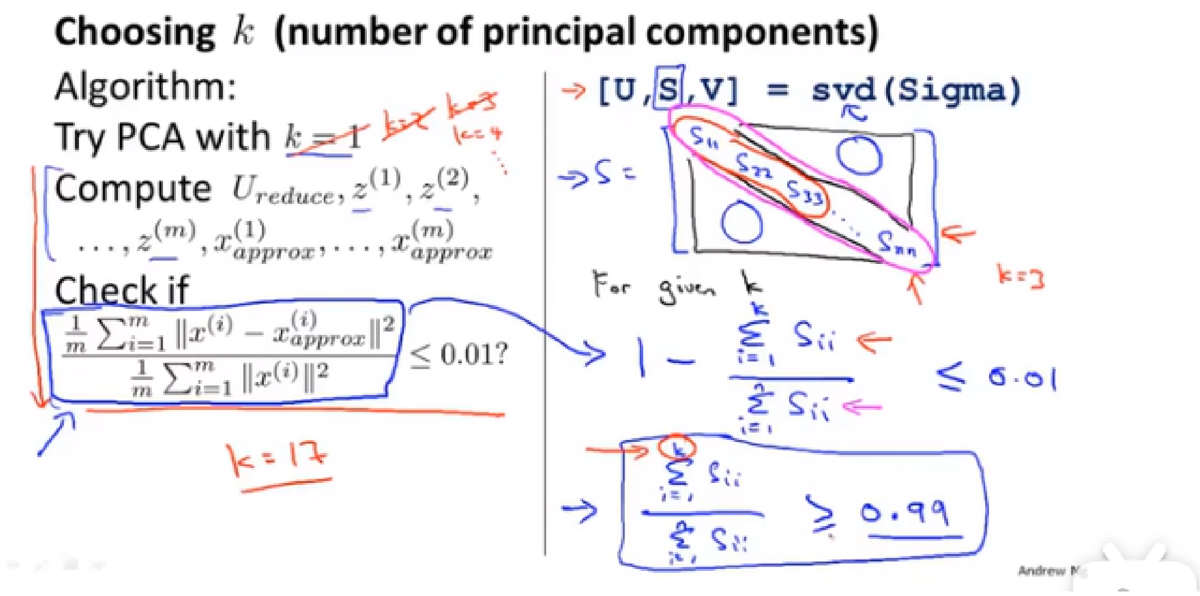
如何选择PCA中的维度k

选择合适的k要达成的目的：

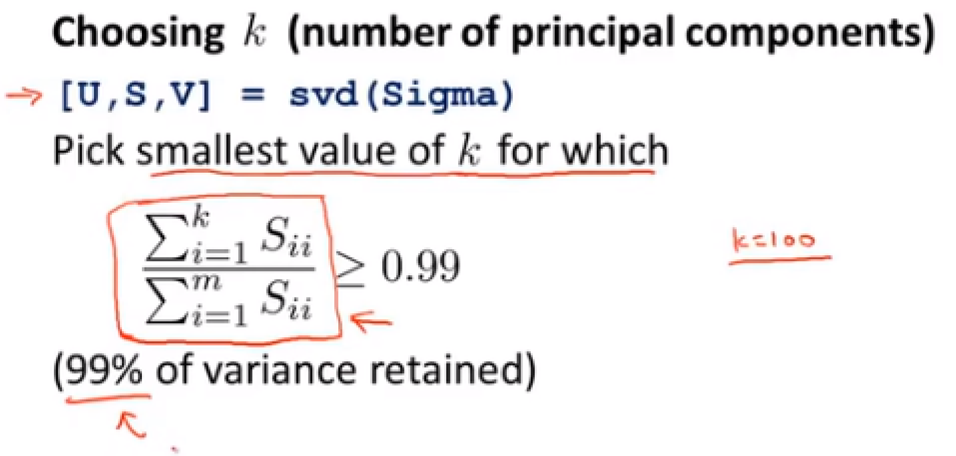


其中一种选择k的方法是从1开始增加k，直到找到满足公式的选择。

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 $\cancel{k=1}$ $\cancel{k=2}$ $k=3$ $k=4$
Compute $U_{reduce}, z^{(1)}, z^{(2)}, \ldots, z^{(m)}, x^{(1)}_{approx}, \ldots, x^{(m)}_{approx}$
Check if
$\frac{\frac{1}{m}\sum_{i=1}^{m}\|x^{(i)} - x^{(i)}_{approx}\|^2}{\frac{1}{m}\sum_{i=1}^{m}\|x^{(i)}\|^2} \le 0.01?$
k=17
[U,S,V] = svd(Sigma)
$S = \begin{bmatrix} S_{11} & & & 0 \\ & S_{22} & & \\ & & S_{33} & \\ 0 & & & S_{nn} \end{bmatrix}$
k=3
For given k
$1 - \frac{\sum_{i=1}^{k} S_{ii}}{\sum_{i=1}^{n} S_{ii}} \le 0.01$
$\frac{\sum_{i=1}^{k} S_{ii}}{\sum_{i=1}^{n} S_{ii}} \ge 0.99$
Andrew Ng

直接选择k


Choosing k (number of principal components)
[U,S,V] = svd(Sigma)
Pick smallest value of k for which
$\frac{\sum_{i=1}^{k} S_{ii}}{\sum_{i=1}^{m} S_{ii}} \ge 0.99$
k=100
(99% of variance retained)