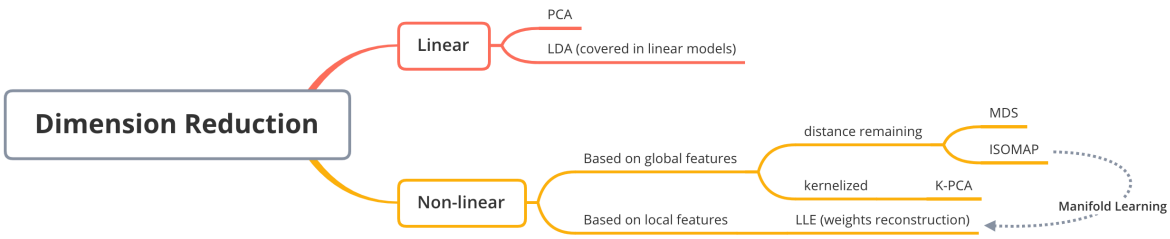
Dimension Reduction

In high dimension, we often have the issue called curse of dimensionality bacause samples are sparse or it's hard to calculate distances. Dimension reduction can not only used to solve this problem, but also extract features.



Note: Linear methods apply linear transformation on the original dataset. That is to say, $Z = W^T X$ where W is the transformation matrix which is orthogonal.

Principal Component Analysis (PCA)

The idea of PCA is to find some principal components to represent original samples. How to find these components? Basically, we expect these principal components that:

- **Pass the closest to a cloud of samples**, in terms of squared Euclidean distance.
- The projection onto them is **the ones with high variances**.

Both goals could derive the consistent solution of PCA.

1. Objective Function & Solution

- Goal 1: Pass the closest to a cloud of samples

  As for this goal, we hope that the restructed sample is close to the original sample as possible. We know that the restructed sample could be represented as $\hat{x}_i = \Sigma_{j=1}^{d'} \mathbf{w_j} z_{ij}$. Therefore this goal would be written as

  $$\Sigma_{i=1}^{m} (\Sigma_{j=1}^{d'} \mathbf{w_j} z_{ij} - x_i)^2 \propto -tr(W^T X X^T W)$$

  Thus, the objective function is

  $$min: -tr(W^T X X^T W)$$
  $$s.t.: W^T W = I$$

- Goal 2: The projection onto them is the ones with high variances

  The projection of sample x_i onto the principal components is $z_i = W^T x_i$ and the variance of samples after projection is $\Sigma_{i=1} W^T x_i x_i^T W$ (centerized samples). Therefore, the objective function is

  $$max: tr(W^T X X^T W)$$
  $$s.t.: W^T W = I$$

The solution of goal 1 and 2 are equivalent. Then we could apply Lagrange multiplier on the objective function:

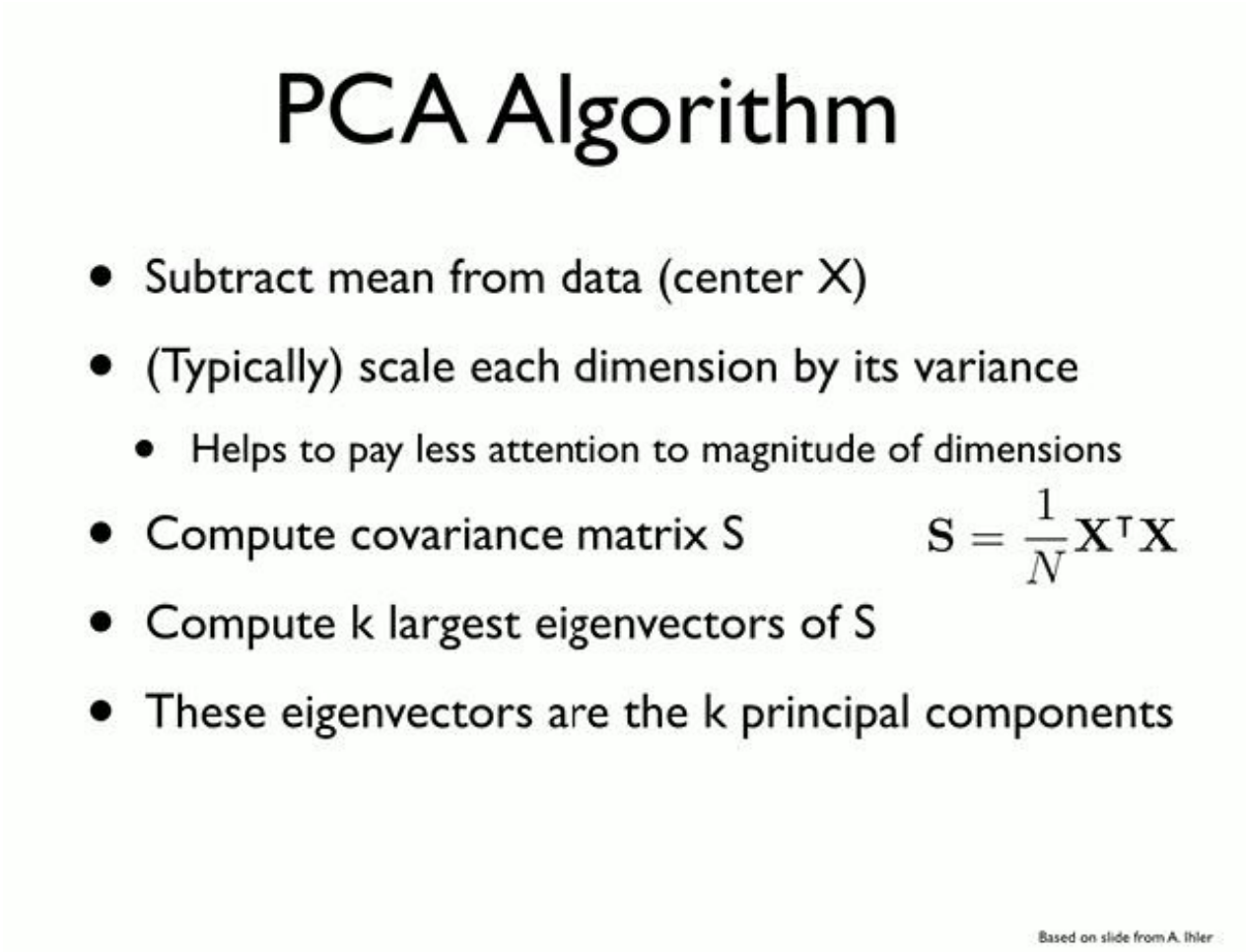
$$L = -tr(W^T X X^T) + \lambda(W^T W - I)$$

Then we could take the derivative of L over w_i and make it to be 0. And we'll get the solution

$$XX^T w_i = \lambda_i w_i$$

This is exactly the form of eigenvectors and eigenvalues.

2. Algorithm



3. Additionals

- Data used in PCA should be centerized first because when we take it as an assumption in derivation.
- How to get eigenvectors and eigenvalues of XX^T ?

  Of course, eigen decomposition is the right answer. However, in practise, we usually use **SVD (singular value decomposition)** because it could decompose arbitrary matrix (does not necessarily to be square).
- w_i stands is the principal component vector and λ_i is actually the variance explained by the principal component. This could help us choose the best d'.