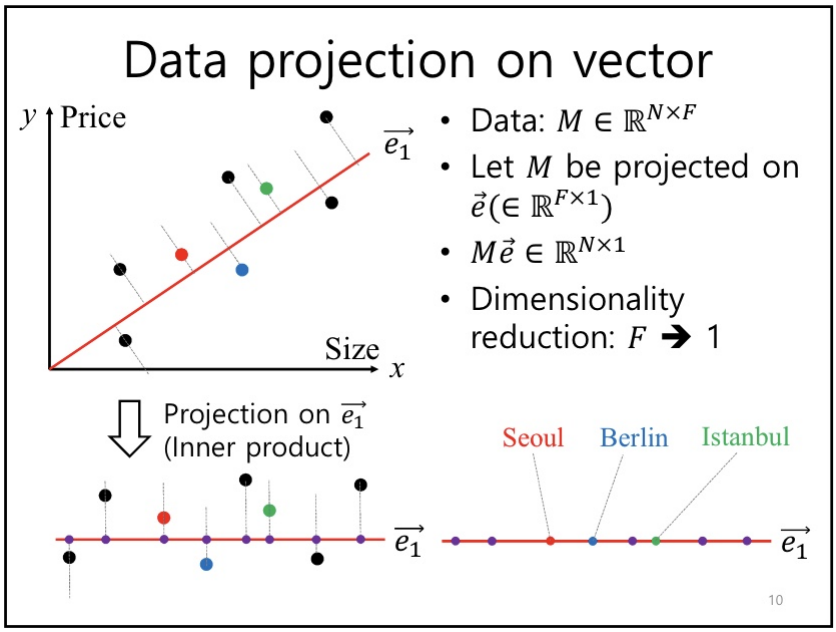
보통, k-means clustering 전에 노이즈 감소를 위해 PCA (principal component analysis)를 적용한다.

So k-means can be seen as a super-sparse PCA.

1) Projection (투영) 의 의미



$$\vec{e}^* = argmax_{\vec{e}} Var(M\vec{e})$$

Projection 시, **데이터set인 M**을 가장 잘 설명하는 (데이터가 골고루 분산되도록 하는) **vector(axis) e** 를 찾고자 한다.

왜냐하면 데이터의 차원 축소 시 정보 손실을 최소화 하기 위해서이다.

다시 말해, *Var(Me)* 분산식 을 최대로 하는 *eigen vector e* 를 찾고자 한다.

분산식은 *공분산 행렬* $ \sigma=b $ 로 나타낼 수 있으며, 분산식은 곧 *eigen value* 를 의미한다.

$$Var(M\vec{e}) = \vec{e}^T \Sigma \vec{e} = \lambda$$

이는 다음과 같은 풀이과정으로 도출된다.

$$Var(M\vec{e}) = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} (M\vec{e} - E(M\vec{e}))^2$$

$$Var(M\vec{e}) = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} (M\vec{e})^2 \quad s.t.\, E(M\vec{e}) = 0$$

$$Var(M\vec{e}) = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} (M\vec{e})(M\vec{e}) = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} (M\vec{e})(M\vec{e})$$

참조

https://stats.stackexchange.com/questions/183236/what-is-the-relation-between-k-means-clustering-and-pca

[]