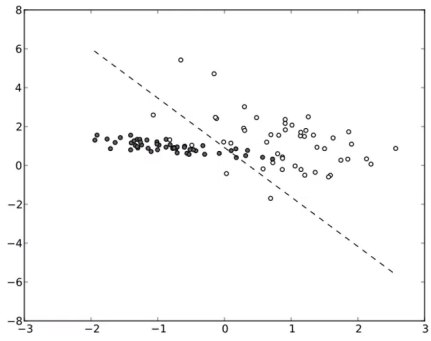
PCA 란?

- Principal Component Analysis
- 여러 고차원 데이터에서 패턴을 찾는 도구 중 하나
- 상관관계가 있는 변수들을 선형결합 하여 변수를 축약하는 기법
- 변수들 간에 내재하는 상관관계, 연관성을 이용해 소수의 주성분으로 차원 축소



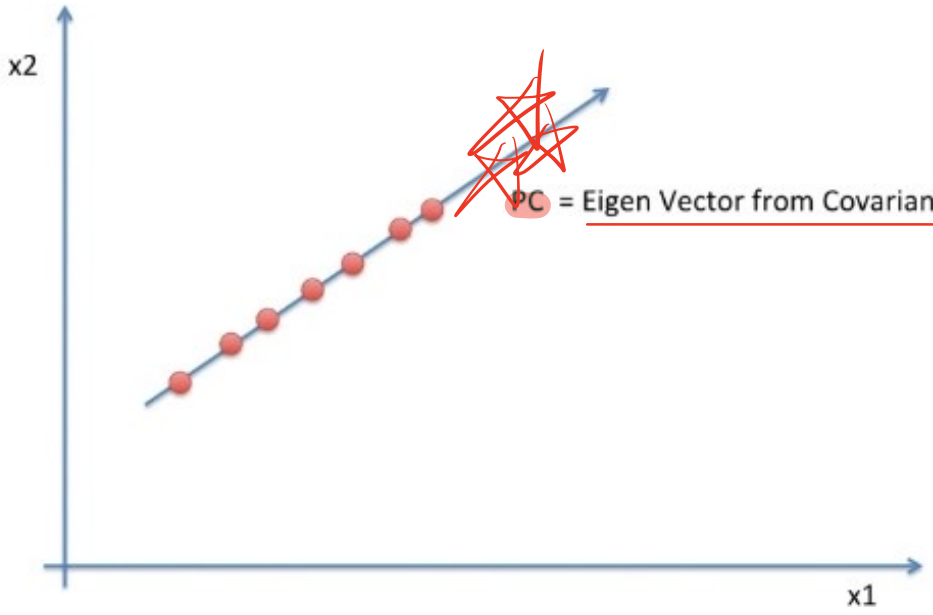
PCA 예제

- **PCA의 순서**
  1. 전체 데이터의 평균을 구한다.
  2. 전체 데이터에 평균을 뺀다.
  3. 공분산 행렬을 만든다.
     - 공분산 행렬은 모든 요소의 공통된 부분을 뺀 개별적 특징이 포함된 데이터
  4. 공분산 행렬을 이용해 Eigen value 와 Eigen vector를 구한다.

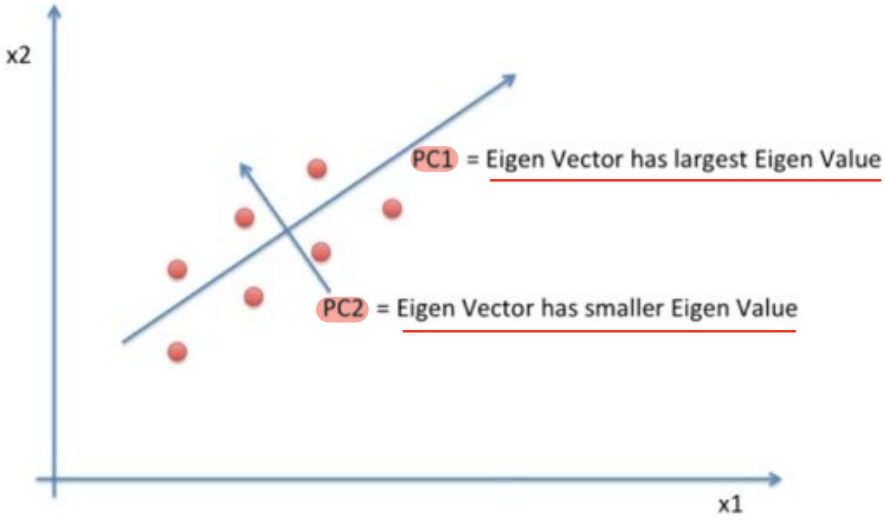
1. 차원 축소 (Dimensionality reduction)

대부분의 경우, 현실 세계의 문제는 가공되지 않은 데이터를 처리해야 한다. 예를 들어, 머신 러닝 모델을 학습하여 증명사진에 있는 인물의 성별을 맞추는 문제가 있을 때, 이 문제를 풀기 위해 우리는 성별이 표시된 증명사진을 머신 러닝 모델의 학습 데이터로 이용할 것이다. 하나의 사진이 200X200의 이미지라고 하면, 해당 사진은 총 40,000개의 feature를 갖는 벡터로 표현이 될 것이다. 그러나 대부분의 머신 러닝 모델은 입력 데이터의 차원이 클 경우, 차원의 저주와 학습 속도가 저하되는 문제를 갖고 있다. 이를 위해 생각해볼 수 있는 것은 이미지에서 인물에 대한 정보를 포함하지 않는 부분을 제거하여 입력 데이터의 차원을 낮추는 것이다. 예를 들어, 증명사진의 왼쪽과 오른쪽 상단은 단색의 배경이기 때문에 인물에 대한 정보를 포함하지 않으며, 배경 부분을 제거하여 모델을 학습하여도 성능에는 큰 차이가 없을 것이다. 이와 같이 데이터에서 불필요한 feature를 제거하는 작업을 차원 축소라고 한다.

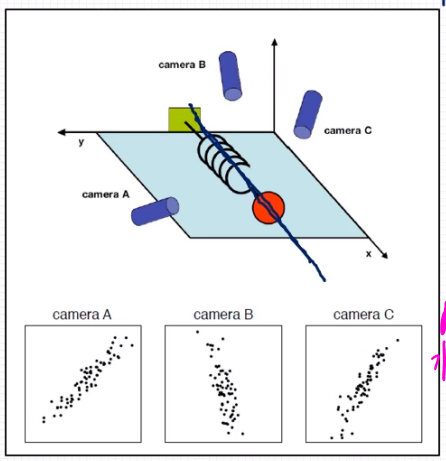
how math can achieve this?



1. We select Eigen vector has largest Eigen value from covariance matrix



32 PCA



A Tutorial on PCA, J. Shlens (2014)

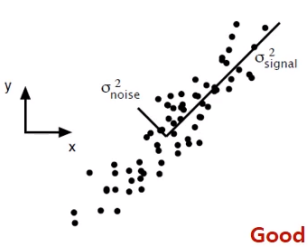
PCA의 가정·한계

I. *Linearity*
Linearity frames the problem as a change of basis. Several areas of research have explored how extending these notions to nonlinear regimes (see Discussion).

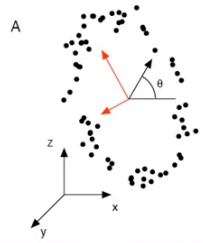
II. *Large variances have important structure.*
This assumption also encompasses the belief that the data has a high SNR. Hence, principal components with larger associated variances represent interesting structure, while those with lower variances represent noise. Note that this is a strong, and sometimes, incorrect assumption (see Discussion).

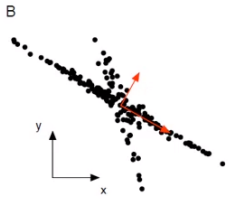
데이터 조건을 들은 직교하는 벡터들의 선형 결합으로 만들어진다.'라는 명제를 가정하는 것이다.

III. *The principal components are orthogonal.*
This assumption provides an intuitive simplification that makes PCA soluble with linear algebra decomposition techniques. These techniques are highlighted in the two following sections.



Good





Bad

- 위의 한계점이 있다고 하더라도, 주어진 데이터 셋에 PCA를 실시하는 것이 좋다.

※ Unsupervised learning의 시작은 dimensionality reduction이다.

↑ 1. 차원을 축소시키는 알고리즘들의 공통목적은 '데이터들이 살고있는 진짜 공간을 찾는 것'이다. → 'latent space(잠재하는)를 찾는 것'이 목적이다.

- 또한, 차원을 축소시키는 알고리즘들은 공통적으로 불필요한 차원을 버리고 필요한 차원을 모으는 작업을 실시한다.

- PCA를 실시할 때 모든 독립변수 쌍의 상관관계가 높으면 한개의 pc로 축소할 수 있고, 모든 독립변수 쌍의 상관관계가 높지 않으면 두 개 이상의 pc로 축소할 수 없다.