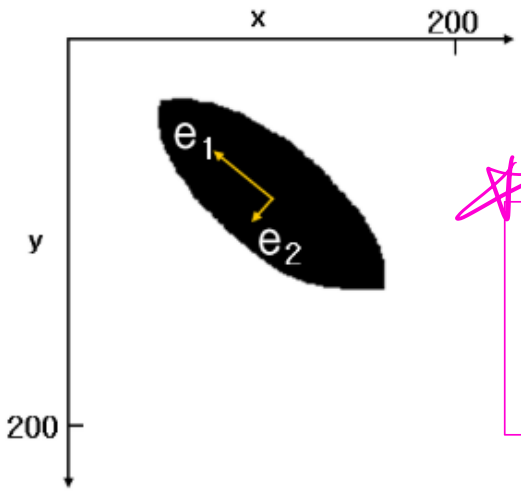
1. PCA(Principal Component Analysis)란?

PC : 특정 데이터셋의 공분산 행렬에 대한 고유벡터.

PCA는 분포된 데이터들의 주성분(Principal Component)를 찾아주는 방법이다. 좀더 구체적으로 보면 아래 그림과 같이 2차원 좌표평면에 n개의 점 데이터 (x1,y1), (x2,y2), ..., (xn,yn)들이 타원형으로 분포되어 있을 때



<그림 1> 2D에서의 PCA 예

• 대칭행렬의 고유벡터 행렬은 직교행렬이다.

• 분산행렬은 대칭행렬이다.

이 데이터들의 분포 특성을 2개의 벡터로 가장 잘 설명할 수 있는 방법은 무엇일까? 그건 바로, 그림에서와 같이 e1, e2 두 개의 벡터로 데이터 분포를 설명하는 것이다. e1의 방향과 크기, 그리고 e2의 방향과 크기를 알면 이 데이터 분포가 어떤 형태인지를 가장 단순하면서도 효과적으로 파악할 수 있다.

PCA는 데이터 하나 하나에 대한 성분을 분석하는 것이 아니라, 여러 데이터들이 모여 하나의 분포를 이룰 때 이 분포의 주 성분을 분석해 주는 방법이다.

여기서 주성분이라 함은 그 방향으로 데이터들의 분산이 가장 큰 방향벡터를 의미한다. <그림 1>에서 e1 방향을 따라 데이터들의 분산(흩어진 정도)이 가장 크다. 그리고 e1에 수직이면서 그 다음으로 데이터들의 분산이 가장 큰 방향은 e2이다.

PCA는 2차원 데이터 집합에 대해 PCA를 수행하면 2개의 서로 수직인 주성분 벡터를 반환하고, 3차원 점들에 대해 PCA를 수행하면 3개의 서로 수직인 주성분 벡터들을 반환한다. 예를 들어 3차원 데이터의 경우는 아래 그림과 같이 3개의 서로 수직인 주성분 벡터를 찾아준다.