**Lab 7. Principal Component Analysis**

고급소프트웨어실습 3분반

20151591 이지현

**숙제: PCA에서 압축을 위한 최적 eigenvector 개수 구하기**

1. Eigenvector의 개수는 원래 데이터의 차원(dimension)과 같으므로 모든 eigenvector를 이용하여 원래 데이터를 투영하면 데이터 압축의 효과를 볼 수 없다. 또한 데이터의 분산 특성을 고려하지 않고 특정 몇 개의 eigenvector를 이용하여 투영시키면 데이터의 유용한 정보가 손실될 수 있다. 유용한 정보를 손실하지 않으면서 최대로 압축 효과를 얻을 수 있도록 eigenvector의 개수를 설정할 수 있는 방법이 있는 지 설명해 보자.

유용한 데이터를 손실되지 않으면서도 최대 압축 효과를 보기 위해서는 분산이 큰 eigenvector를 선택하는 것이 좋다. 분산이 큰 eigenvector라는 것은 원본 데이터의 모습을 잘 표현한다는 의미이기 때문이다. 따라서 분산이 큰 순서대로 eigenvector들을 선택하면 유용한 정보를 손실하지 않으면서 대푯값을 가지기 때문에 압축효과를 얻을 수 있다고 생각한다.

Eigenvector의 개수에 대한 선택은 사용자가 필요로 하는 섬세도에 따라 달라진다. 좀 더 디테일한 데이터를 원한다면 개수를 많이 선택하면 되는 것이고, 주요한 모습만 보고 싶다면 개수 선택을 줄인다면 될 것이다.