5월6일~5월12일

∙ 연구에 있어서 Principal Components Analysis에 대한 깊은 내용과 이해가 필요하여 집중적으로 공부 하였습니다.

**3. Principal Components Analysis(PCA)**

이는 데이터의 패턴을 식별하고 유사점과 차이점을 강조하는 방식으로 데이터를 표현하는 한 방법이다. 그래픽 표현의 고급스러움이 없는 높은 차원의 데이터에서 데이터 패턴을 찾기가 어려울 수 있으므로 PCA는 데이터 분석을 위한 강력한 도구이다. PCA의 다른 주요 이점은 일단 데이터에서 패턴을 발견하고 데이터를 압축한다는 것이다. 많은 정보 손실없이 차수의 수를 줄임으로써 이 기술은 이미지 압축에 사용된다. 이 장에서는 구성 요소 데이터 집합에 대한 분석. 이 기술을 직접 사용하려고 할 때 어떤 일이 일어나는지에 대한 설명을 제공한다.

**3. 1 Method**

**Step 1: Get some data**

간단한 예에서는 제공된 data set를 사용합니다. 이것을 선택한 이유는 PCA분석이 각 단계에서 무엇을 하는지 보여주기 위해 데이터 플롯을 제공 할 수 있기 때문이다. 여기서 사용될 데이터는 데이터 플롯과 함께 Figure 3.1에 나와있다.

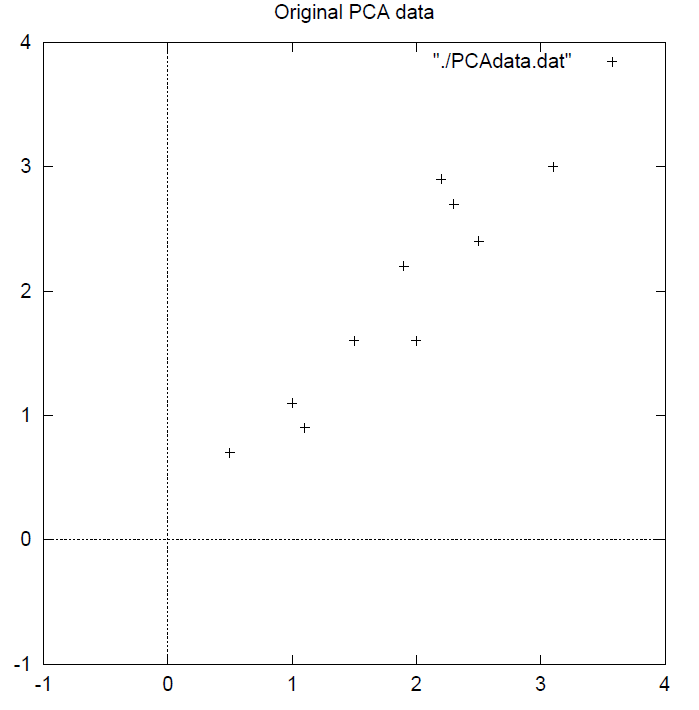
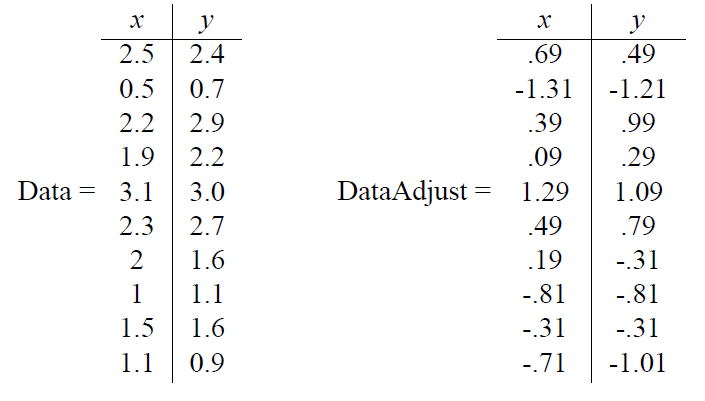


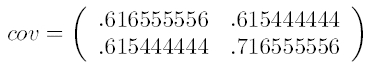
Figure 3.1: PCA example data, original data on the left, data with the means subtracted on the right, and a plot of the data

**Step 2: Subtract the mean**

PCA가 제대로 작동하려면 각 데이터 차원에서 평균을 빼야한다. 평균은 각 차원의 평균이다. 따라서 모든 x값은 (모든 데이터 포인트의 x값의 평균)를 뺀 값이고 모든 y 값은 에서 뺀다. 그러면 평균이 0인 data set가 생성된다.

**Step 3: Calculate the covariance matrix**

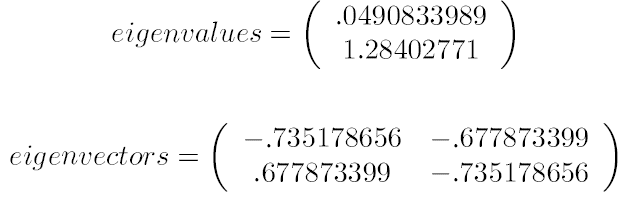
이는 1.4절에서 논의 된 것과 똑 같은 방식으로 수행된다. 데이터가 2차원이므로 공분산 행렬은 22가 된다. 결과는 아래와 같다.



따라서 이 공분산 행렬의 각 요소가 양수이기 때문에 x와 y변수가 함께 증가 할 것으로 예상한다.

**Step 4: Calculate the eigenvectors and eigenvalues of the covariance matrix**

공분산 행렬이 정사각이므로 이 행렬에 대한 고유 벡터와 고유값을 계산할 수 있다. 이 정보는 데이터에 대한 유용한 정보를 제공하므로 중요하다. 고유 벡터와 고유값은 다음과 같다.



이러한 고유벡터는 모두 단위 고유 벡터이다. 길이는 모두 1이다. 이것은 PCA에서 매우 중요하다. 그러나 대부분의 maths packages는 고유벡터를 요구할 때 단위 고유벡터를 줄 것이다.

그래서 그들이 의미하는게 무엇일까? Figure 3.2의 데이터 플롯을 보면 데이터가 매우 강한 패턴을 가지고 있음을 알 수 있다. 공분산 행렬에서 예상 한 바와 같이 두 변수는 실제로 함께 증가한다. 데이터의 꼭대기에 두 고유벡터를 도식화 했다. 플롯에 대각선 점선으로 나타난다. 고유벡터 섹션에서 언급했듯이, 이들은 서로 수직이다. 그러나 더 중요한 것은 데이터의 패턴에 대한 정보를 제공한다는 것이다. 고유벡터 중 하나가 최적의 라인을 그리는 것과 같은 포인트의 중간을 어떻게 지나가는지 확인해야한다. 이 고유벡터는 두 데이터 세트가 어떻게 그 라인과 관련이 되는지 보여준다. 두 번째 고유벡터는 데이터에서 덜 중요한 다른 패턴을 우리에게 제공한다. 모든 점은 기본 선을 따르지만 기본 선의 측면에는 어느 정도 떨어져 있다.

따라서, 공분산 행렬의 고유 벡터를 취하는 과정에 의해, 우리는 데이터를 특성화 하는 선을 추출 할 수 있다. 나머지 단계에서는 데이터를 변환하여 해당 행으로 표현한다.

