4월23일~4월28일

∙ 저 번주 연구노트에 이어서 논문 주제로 선정한 Principal Components Analysis(PCA)에 대한 mathematical concepts을 공부합니다. Covariance의 개념으로 Covariance Matrix에 대해 공부한다.

**1.4 The covariance Matrix**

Covariance는 항상 2차원 사이에서 측정된다. 2차원 이상의 데이터 세트가 있는 경우 계산할 수 있는 공분산 측정이 두개 이상 있다. 예를 들어, 3차원 데이터 세트(차원 x, y, z)에서 cov(x, y), cov(x, z) 및 cov(y, z)를 계산할 수 있다. 사실, n차원 데이터 세트의 경우 개의 다른 공분산 값을 계산할 수 있다.

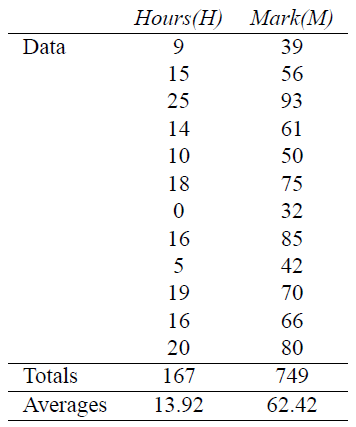
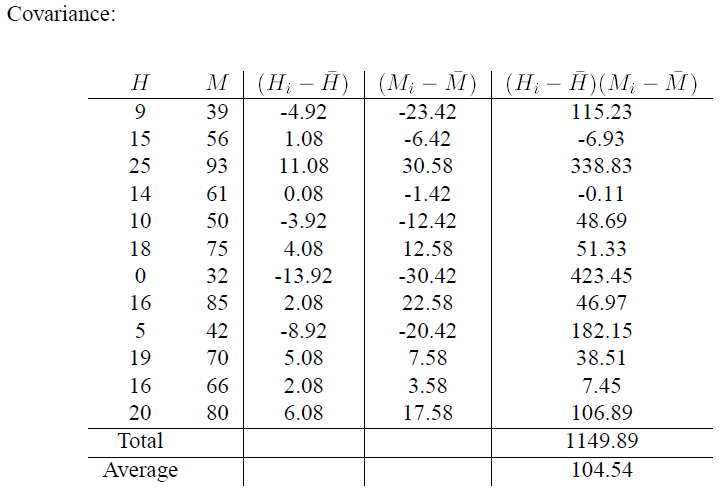
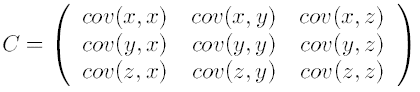


Figure 1. 2-dimensional data set and covariance calculation

다른 차원사이에 가능한 모든 공분산 값을 얻는 유용한 방법은 모두를 계산하여 행렬에 넣는 것이다. 이 과정에서는 우리가 행렬에 익숙하고 어떻게 정의할 수 있는지를 가정한다. 따라서 n차원의 데이터 집합에 대한 공분산 행렬의 정의는 다음과 같다.

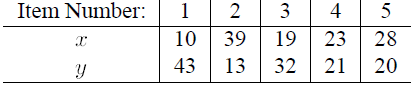


여기서 는 n개의 행과 n개의 열이 있는 행렬이고, 는 x번째 차원이다. 이 공식은 n차원 데이터 세트가 있는 경우 행렬에 n개의 행과 열이 있으므로 행렬의 각 항목이 두개의 개별 차원 사이의 공분산을 계산한 결과이다. 예를 들어, 행2 열3의 항목은 2차원과 3차원 사이에서 계산된 공분산 값 입니다. 또한, 일반적인 차원 x, y, z를 사용하여 가상의 3차원 데이터 세트에 대한 공분산 행렬을 만든다. 그런 다음, 공분산 행렬은 3행 3열을 가지며 값은 다음과 같다.



몇가지 주의해야 할 점이 있다. 주요 대각선 아래로 공분산 값이 차원 중 하나와 그 자체 사이에 있음을 알 수 있다. 이것들은 해당 차원의 차이이다. 다른 점은 cov(a,b) = cov(b,a)이기 때문에 행렬은 주 대각선에 대해 대칭이라는 것이다.

예를 들어, 다음 2차원 데이터 세트에서 x와 y차원 사이의 공분산을 계산하고 데이터에 나타나는 결과를 설명한다.



이 3차원 데이터 세트에 대한 공분산 행렬을 계산한다.

