Chapter08

주성분과 요인 분석

주성분 분석 (PCA)

FASTCAMPUS
ONLINE
금융공학/퀀트

강사. 장순용

키포인트

- 주성분 분석 (Principal Component Analysis, PCA).
- 주성분 (Principal Component, PC)의 해석.
- 특이값 분해와 고유값 분해.

주성분 분석: 목적

- 서로 상관관계가 있는 변수를 상관관계에서 자유로운 변수로의 변환.
- 서로 직교 (orthogonal)하는 새로운 좌표축으로의 변환.
- 변동 (분산) 크기에 따라서 변수 (주성분)를 정렬한다.
- 데이터 전처리, 모델링, 차원축소, 시각화 등 활용.

주성분 분석: 결과

- Loading: 정규화된 주성분 (PC).

  → 주성분의 개수 = 원 변수의 개수.

- Variance: 개개 주성분에 해당하는 분산 σ^2.

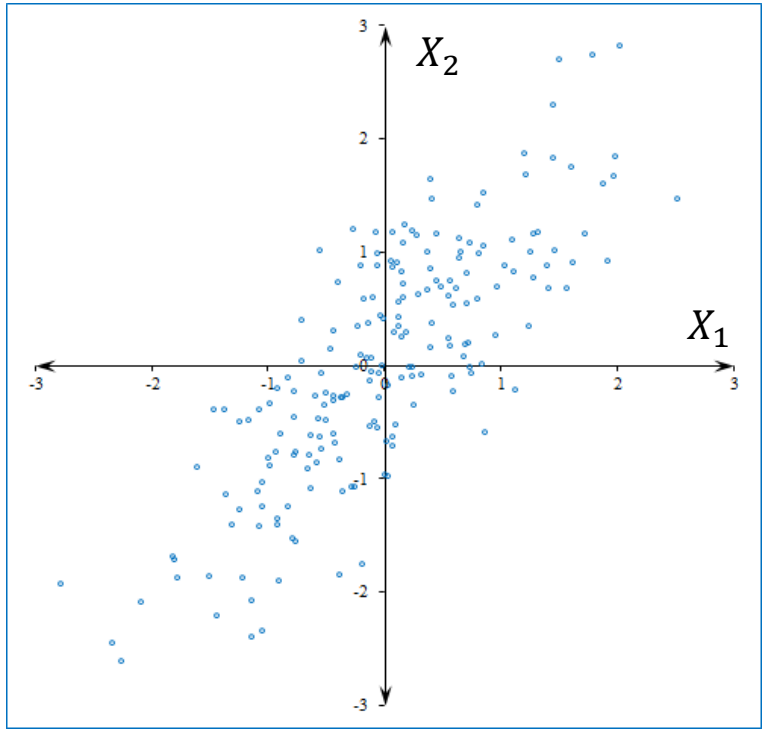
  → 표준편차 σ로 대체될 수도 있다.

  → 주성분 방향으로의 변동의 폭.

- Transformed score: 주성분을 새로운 좌표축으로 사용하여 데이터를 변환하여 나타낸 것.

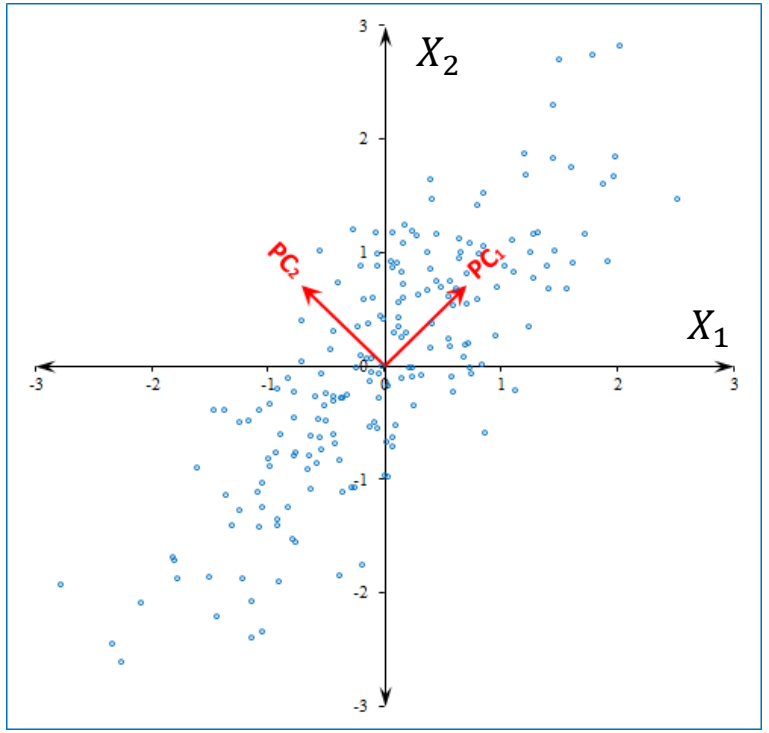
I 주성분 분석: 원리

- 다음과 같이 2차원 데이터가 분포되어 있다고 가정한다.



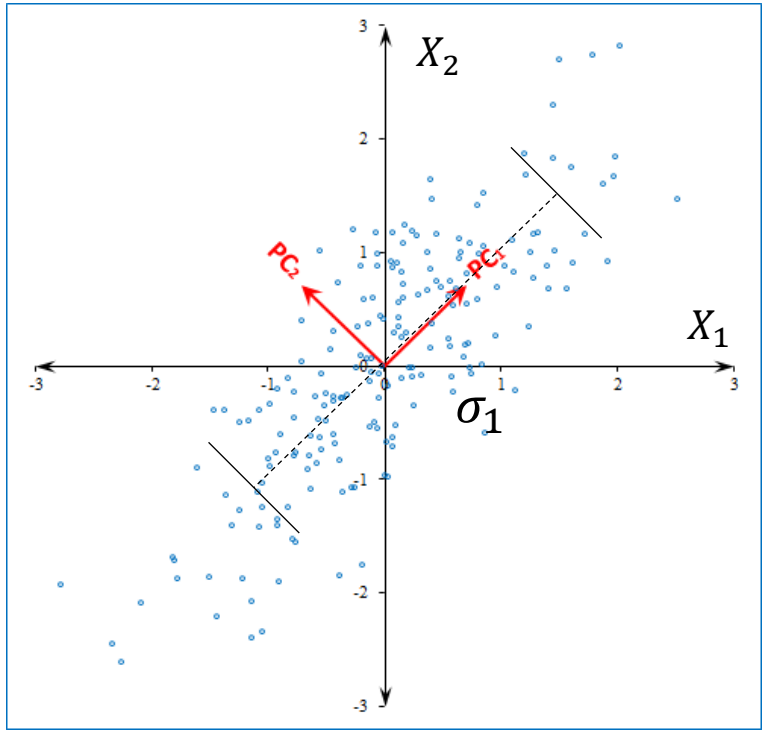
I 주성분 분석: 원리

- PC_1과 PC_2는 서로 직교함.



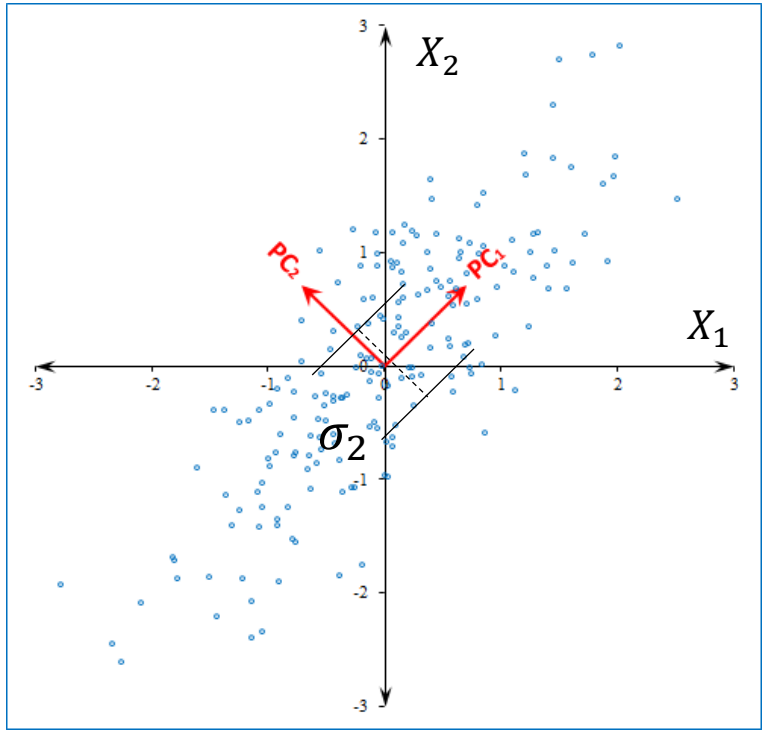
I 주성분 분석: 원리

- PC_1에 해당하는 변동은 σ_1으로 나타낼 수 있다. 가장 큰 변동에 해당한다.



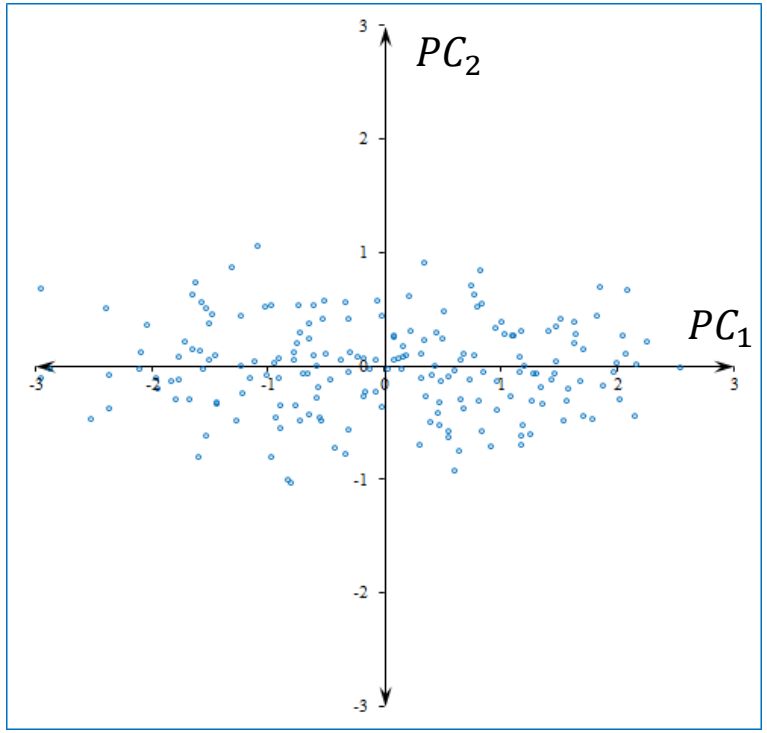
I 주성분 분석: 원리

- PC_2에 해당하는 변동은 σ_2이며 $\sigma_2 < \sigma_1$이다.



I 주성분 분석: 원리

- 주성분 PC_1 과 PC_2를 새로운 좌표축으로 사용하였을 때.



I 주성분 분석: 계산 원리

1. 데이터 행렬에 특이값 분해 (Singular Value Decomposition, SVD) 적용.

2. 공분산 행렬 또는 상관계수 행렬에 고유값분해 (Eigenvalue Decomposition, ED) 적용.

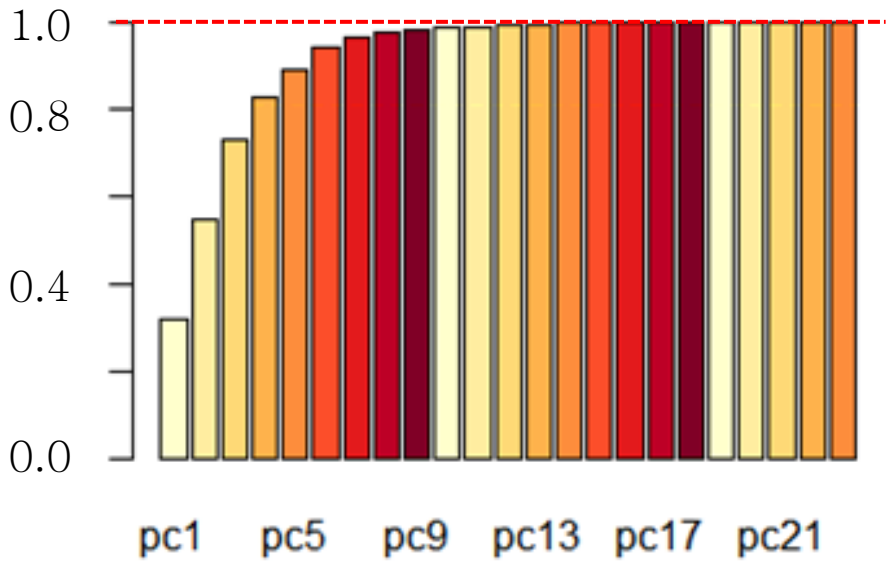
→ 만약에 변수를 표준화했다면 상관계수행렬을 사용하는 것과 같다.

$$\text{표준화} = \frac{X - \bar{X}}{\sigma}$$

주성분 분석: 누적 분산

- 주성분은 서로 독립적인 확률변수 이므로, 전체 분산은 개개 주성분 방향의 분산의 합으로 쉽게 구할 수 있다.

$$\sigma_{total}^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \cdots$$



끝.

감사합니다.

FAST CAMPUS
ONLINE
장순용 강사.