PCA

1. 서론: 데이터 변환

데이터를 불러와서, 결합하고, 결측값과 특이값을 확인 후 처리하고, 필요한 부분의 데이터만 선별적으로 선택 혹은 제거한 후, 분석의 목적에 따라, **데이터의 형태에 따라 데이터 변환 작업을 수행한다.**

**개념적 정의, 조작적 정의를 통해 파생변수를 개발하여 Input으로 넣는 것이 굉장히 중요하다.**

1. **본론: PCA란**

서로 상관성이 높은 여러 변수들의 선형 조합으로 만든 새로운 변수들(주성분)로 요약, 축소하는 기법이다.

첫번째 주성분으로 전체 변동(variance)을 가장 잘 설명하고, 두번째 주성분으로 나머지 부분을 설명하여 정보의 손실없이 가장 많이 설명할 수 있도록 변수들을 선형조합한다.

* **주성분분석, 왜 사용할까**

1. 여러 변수들간에 내재하는 상관관계, 연관성을 이용해 소수의 주성분 또는 요인으로 차원을 축소해서 데이터이해 & 관리가 쉽게 해줌 (약간 충분통계량 같은 느낌?)
2. 다중 공선성 문제 해결
3. 연관성이 높은 변수간 주성분 or 요인분석을 통해 차원을 축소한 후, 군집분석을 수행하면 군집화 결과, 연산 속도 개선에 기여가능.

R 프로그래밍

* **주성분 분석에는 3가지에 대한 결정이 필요하다.**

1. 상관 행렬과 공분산 행렬 중에 어느 것을 사용?
2. 주성분의 개수를 몇개로 설정할 것인가?
3. 어떤 주성분을 사용할 것인가?

* **프로그래밍 step**

1. Scaling을 통해서 Z변환 (min-max scaling도 존재함)
2. 같은 방향으로 맞춰주기

(V1: 총자본순이익율 // V2: 자기자본순이익율 // V3: 자기자본비율 // V4: 부채비율 // V5: 자기자본회전율)

**(V4는 부채비율로 낮을 수록 좋은 것이기 때문)**

1. 공분산 확인

* 상관관계가 높은 변수들간의 선형결합이 PCA의 핵심이기 때문이다.

1. PCA진행하고 해석하기

![텍스트이(가) 표시된 사진

자동 생성된 설명]()

Standard deviation은 각 주성분에 대한 표준편차를 의미한다.

Proportion of variance는 각 주성분이 전체 데이터에 대한 기여율을 의미한다.

**Question: 몇개의 주성분을 사용할 것인가?**

1. 누적기여율이 0.8이상
2. 평균분산보다 큰 PC만 선별하기

**(표준화한 데이터에 대한 상관관계 행렬을 사용할 경우, 고유값이 최소 1보다 큰 PC 사용하기)**

1. Screen plot을 그려봤을 때, 꺽이는 부분이 있다면, elbow 지점 (x축에는 pca 변수의 개수를, y축에는 eigen value를 잡았다)

* 그 다음부터는 R로 보여주기
* PCA 로 차원축소 할 때 몇 개의 PCs 를 선택할 것인가에 대해 학문적으로 정의된 정답(universal rule)은 없으며, 많이 사용되는 '경험에서 나온 법칙(rule of thumb)'은 3가지 있습니다.
* **(1) 누적기여율(설명된 분산의 누적 비율)이 최소 (at least) 0.8 이상일 것.**
* **(2) 단지 평균 분산보다 큰 PC만 선별할 것.**
* **(만약 표준화한 데이터에 대한 상관관계행렬을 사용할 경우 고유값(eigenvalue)이 최소 1보다 큰 PC)**
* **(3) Scree plot 을 그려봤을 때 꺽이는 부분 (elbow)이 있다면 elbow 지점 앞의 PC 개수 선택.**
* 것입니다.
* 아래 그래프로 보면 주성분 4개째에서 수평으로 드러누웠으므로 한개를 뺀 (4-1 = 3) 3개 주성분이 적합해 보입니다.  
    
  출처: <https://rfriend.tistory.com/61> [R, Python 분석과 프로그래밍의 친구 (by R Friend)]

출처: <https://rfriend.tistory.com/61> [R, Python 분석과 프로그래밍의 친구 (by R Friend)]

1. 결론

* PCA 정리