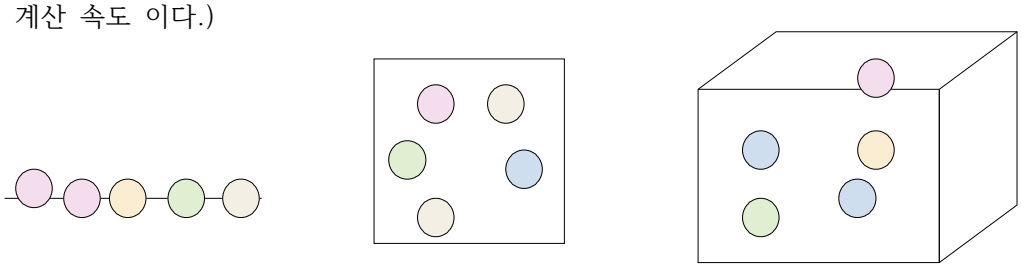
Principle Component Analysis (PCA, 주성분분석)

Principle Component Analysis (PCA, 주성분분석)은 고차원 데이터를 저차원으로 환원하는 방법이다. 이는 대중적인 분석법으로 여러 블로그 또는 검색을 통해 손쉽게 접할 수 있다. 다른 정보들과는 차별성을 주기 위하여 PCA 기법을 사용하게 된 배경, 예시를 통한 설명, 일상생활에서 볼 수 있는 예제, 수학적으로 Principle Component Analysis 증명 등을 정리하여 처음 배우더라도 쉽게 접근 가능하도록 하였다.

1, curse of dimensionality

차원이 증가할수록 필요한 샘플 데이터가 많아진다. (예를 들어, 변수가 10개일 때 보다 100개 일 때 필요한 데이터가 더 많은 것을 의미한다.) 따라서, 차원의 수보다 데이터의 수가 적게 되면 성능이 떨어진다. 이러한 어려움을 표현한 용어가 '차원의 저주'이다.
따라서 데이터 분석에서 핵심이 되는 변수를 선별하여 문제의 차원을 낮추고자 한다. 이것이 차원 축소이며 Principal Component Analysis (PCA)는 데이터의 차원을 낮추는 기계적인 방법이다. (차원축소로 얻는 장점은 1, 데이터를 다루기 편함 2, 시각화하기 쉬움 3, 계산 속도 이다.)



-> 다각형이 커질수록 빈공간이 많아진다. 즉, 필요한 데이터가 많아진다.

2, 일상생활에서 볼 수 있는 차원축소

일상생활에서 차원을 줄여 특정한 대상의 내부적인 특징을 알아내는 예시를 살펴 볼 것이다. '착함'이라는 개념을 생각해보자. 우리는 사람들의 착한 정도를 측정하는 방법이 없으며, '착함'은 실제로 관측되지도 않고, 측정할 수 없는 잠재특징이다. 하지만 우리는 사람들의 서로 다른 행동패턴을 분석하여 마음속에서 '착함'을 일차원으로 사상하거나 차원축소를 한다.

이 알고리즘은 잠재적인 특징 중 중요한 것을 알아내도록 한다. 여기에서 '중요함'은 다양한 질문에 대한 반응들의 분산값을 의미한다. 즉, 반응을 효율적으로 모형화 하겠다는 것이다. 우리는 잠재적인 특징들 중 증요한 것들을 모아서 이들을 저차원의 부분공간에 표현하는 모형을 만들 것이다.

3, Principle Component Analysis (PCA, 주성분분석)

1) Principle Component Analysis (PCA, 주성분분석) 이란?

: 여러 변수가 있는 데이터들 중에서 상관성이 높은 변수들을 요약, 축소하는 기법이다.
: 첫 번째 주성분으로 전체 변동을 가장 많이 설명하고, 두 번째 주성분으로는 첫 번째 주성분이 설명하지 못하는 나머지 변동을 정보의 손실 없이 가장 많이 설명할 수 있도록 변수들의 선형조합을 만든다.

* 필요한 선형대수 개념 (rank, orthogonal, transpose, base, span, matrix decompositon)

2) 평균과 공분산

; sample mean (표본평균)

[x1 ... xn] 을 p × N 의 관측행렬이라고 할 때, 백터 x1 ... xn 들의 sample mean

$$M=\frac{1}{N}(x_1 + \dots + x_n)$$ 이다.

: 공분산

$$\widehat{X_K} = X_K - M \quad (K = 1\sim N)$$

다음으로 정의된 P × N 행렬 B = $[\widehat{X_1}\ \widehat{X_2}\ \dots\ \widehat{X_N}]$ 의 각 열의 표본평균이 0인 관계로, B를 평균편차 형식이라 한다.

covariance matrix (공분산행렬) 은 아래의 p×p 행렬 S로 정의된다.

$$S=\frac{1}{N-1} B B^T$$ (모든 BB^T 형태의 행렬은 항상 양의 준정부호)

3) Principle Component Analysis (PCA, 주성분분석)

행렬 [x1 ... xn] 이 이미 평균편차 형식으로 되어 있다고 하자.

Principle Component Analysis (PCA, 주성분분석법) 은 새로운 변수 $y_1, .. y_p$가 서로 상관되어 있지 않고, 그 분산이 내림차순으로 정렬되도로 하는 변수변환 X = pY

$$\begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_p \end{bmatrix} = \begin{bmatrix} u_1 u_2 \dots u_p \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_p \end{bmatrix}$$

를 결정하는 p × p 직교행렬 $p=\begin{bmatrix} u_1 u_2 \dots u_p \end{bmatrix}$ 를 찾는 것을 목표로 한다.

직교 변수변환 X = pY는 각 관측벡터 X_k가 X_k= pY_k를 만족하는 '새로운 명칭 Y_k를 부여받는 것을 의미한다. Y_K는 P의 열들에 대한 X_K의 좌표벡터이고
$Y_k = p^{-1}X_k = p^TX_k$ (k=1,2,3,..,N)을 만족한다.

임의의 직교행렬 p에 대하여, Y_1,....,Y_N 의 공분산행렬은 P^TSP 임을 쉽게 보일 수 있다. 따라서 구하고자 하는 직교행렬 P는 P^TSP를 대각화하는 행렬이다. S 의 고윳값 λ_1,....,λ_P 가 $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_P \geq 0$으로 정렬되어 대각행렬 D의 대각선상에 놓인다 하자. 또한 그에 대응되는 단위고유벡터 $u_1 u_2 \ldots u_p$ 를 (관측행렬 내) 데이터의 principal components (주성분)이라 한다. S의 가장 큰 고육값에 대응되는 고유벡터를 첫 번째 주성분, 두 번째로 큰 고윳값에 되응되는 고유벡터를 두 번째 주성분 등과 같이 말한다.

첫 번째 주성분 u_1은 새 변수 y_1을 다음의 방식으로 결정한다. $c_1 c_2 \ldots c_p$를 u_1의 성분이라 하면, u_1^T는 P^T의 첫 번째 행이므로, 방정식 Y=P^TX로부터

$$y_1 = u_1^TX = c_1x_1 + c_2x_2 + \ldots + c_px_p$$

임을 알 수 있다. 따라서, y_1 은 원래 변수 x_1,....,x_p 의 일차결합이고, 이때 고유벡터 u_1의 성분을 가중치로 사용한다. 유사한 방식으로 u_2 가 y_2 를 결정한다.

예제) 다중스펙트럼 영상에 대한 초기 데이터는 R^3의 4백만 개의 벡터로 구성되어 있으며, 이에 대한 공분산행렬은

$$S = \begin{vmatrix} 2382.78 & 2611.84 & 2136.20 \\ 2611.84 & 3160.47 & 2553.90 \\ 2136.20 & 2553.90 & 2650.71 \end{vmatrix}$$ 이다.

데이터의 주성분을 구하고, 첫 번째 주성분에 의하여 결정되는 새로운 변수를 열거하라.

sol) S의 고윳값과 그에 대응되는 주성분(단위고유벡터)은

λ_1=7614.23 λ_2=427.63 λ_3=98.20

$$u_1 = \begin{pmatrix} .5417 \\ .6295 \\ .5570 \end{pmatrix} \quad u_2 = \begin{pmatrix} -.4894 \\ -.3026 \\ .8179 \end{pmatrix} \quad u_3 = \begin{pmatrix} .6834 \\ -.7157 \\ .1441 \end{pmatrix}$$

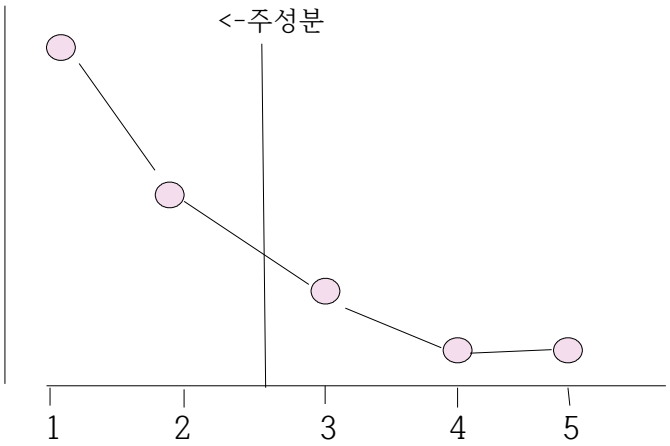
이다, 첫 번째 주성분에 관한 변수를 소수점 이하 두 자리까지 표시하면

$$y_1 = .54x_1 + .63x_2 + .56x_3$$ 이다.

4) 주성분의 선택법

: Principle Component Analysis (PCA, 주성분분석)의 결과에서 cumulative proportion(누적기여율)이 85%이상이면 주성분의 수로 결정할 수 있다.

: scree plot을 활요하여 eigenvalue(고유값)이 수평을 유지하기 전단계로 주성분의 수를 선택한다.



*Reference

[1] 2019 ADsP 데이터 분석 준전문가, 윤종식, 2019

[2] Linear Algebra and Its Applications (5th ed), David C. Lay, Stephen R. Lay and Judi J. McDonald, 2016

[3] Doing Data Science 데이터 과학 입문, Rachel Schutt, Cathy O'Neil, 2013