확률과 통계

섹션 - 7

강사 : James 쌤

주성분 분석

키포인트

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

주성분 분석의 목적

- 서로 상관성이 있는 변수들을 상관성이 없는 (≈ 0) 변수들로 변환.
- 서로 직교 (orthogonal)하는 새로운 좌표축의 좌표계로 변환.
- 변동 (분산) 크기에 따라서 변수 (주성분)를 정렬한다.
- 데이터 전처리, 모델링, 차원축소, 시각화 등의 용도로 활용.

주성분 원리

- 먼저 k 개의변수 또는 feature를 전제한다 $X_1, X_2, \ldots, X_k$.
- 주성분의 개수는 원래 차원의 개수와 같다.
- 그러면 주성분 $PC_1, PC_2, \ldots, PC_k$는 원 변수의 선형조합으로 표현할 수 있다.

$$PC_i = \alpha_{1,i}X_1 + \alpha_{2,i}X_2 + \cdots + \alpha_{k,i}X_k$$

- 반대로, 원 변수도 주성분들의 선형조합으로 표현할 수 있다.

$$X_i = \beta_{1,i}PC_1 + \beta_{2,i}PC_2 + \cdots + \beta_{k,i}PC_k$$

주성분 분석 결과

- Loading: 정규화된 주성분 (PC) 벡터.

  $\Rightarrow$ 주성분의 개수 = 원 변수의 개수.

- Variance: 개개 주성분에 해당하는 분산 σ^2.

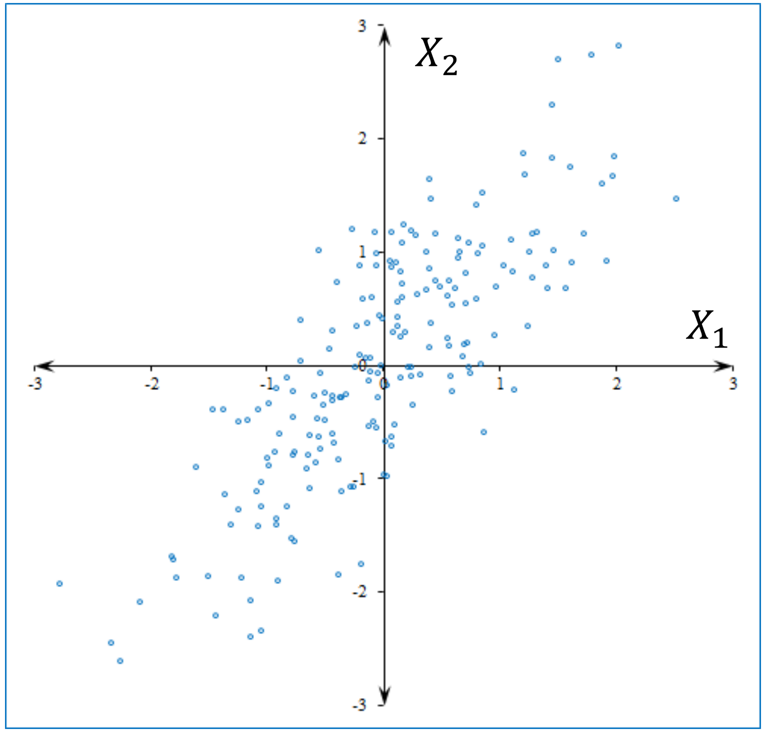
  $\Rightarrow$ 표준편차 σ로 대체될 수도 있다.

  $\Rightarrow$ 주성분 방향으로의 변동의 폭.

- Transformed score: 주성분 벡터들을 새로운 좌표축으로 삼고 데이터를 표현한 것.

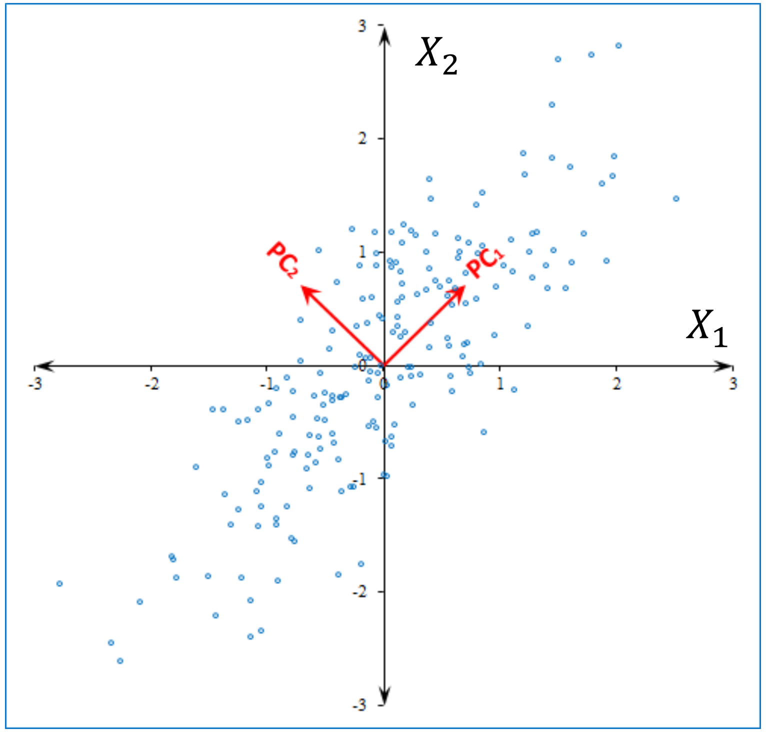
주성분 분석 결과 예시

- 다음과 같이 2차원 데이터가 분포되어 있다고 가정한다.



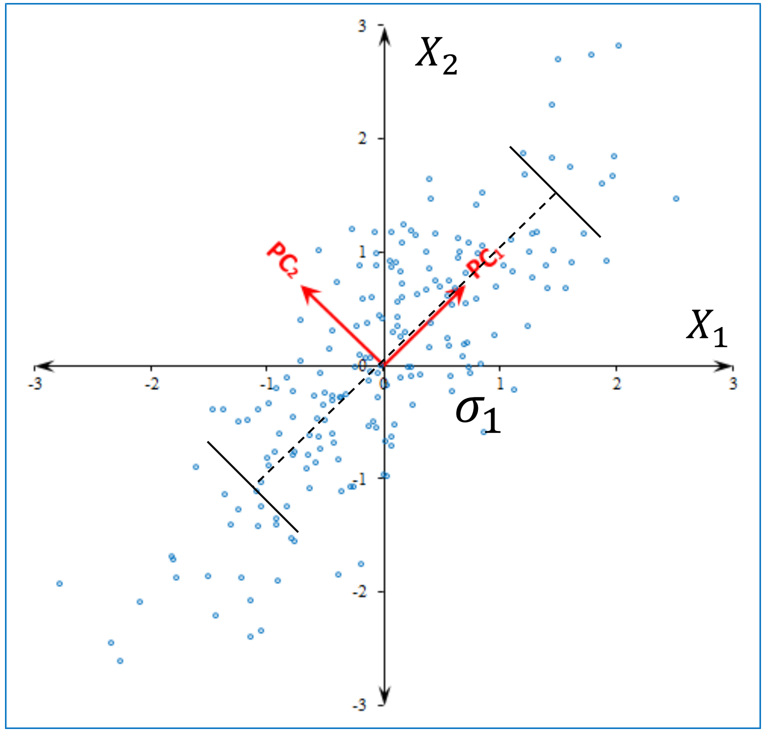
주성분 분석 결과 예시

- 다음 PC_1과 PC_2는 서로 직교한다.



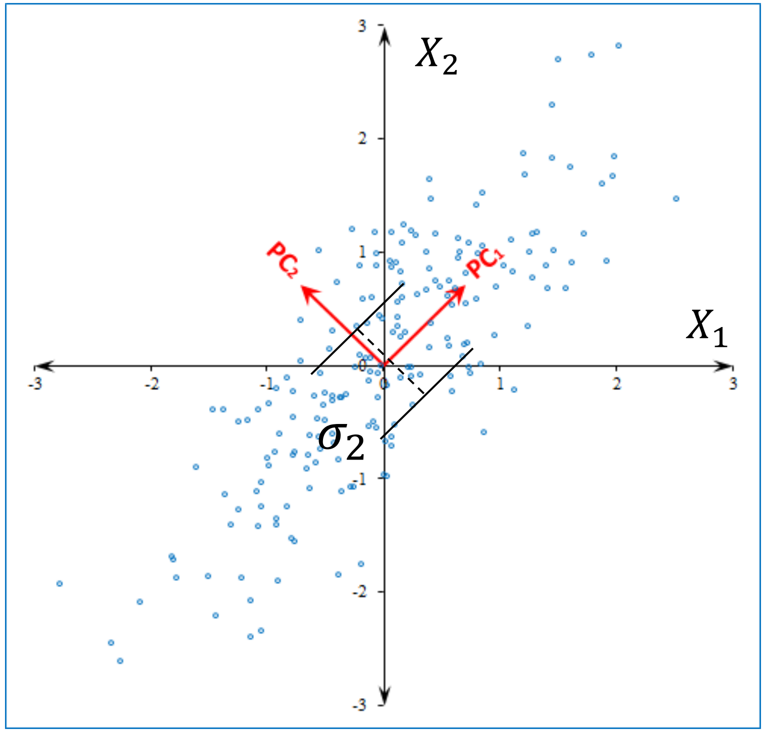
주성분 분석 결과 예시

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



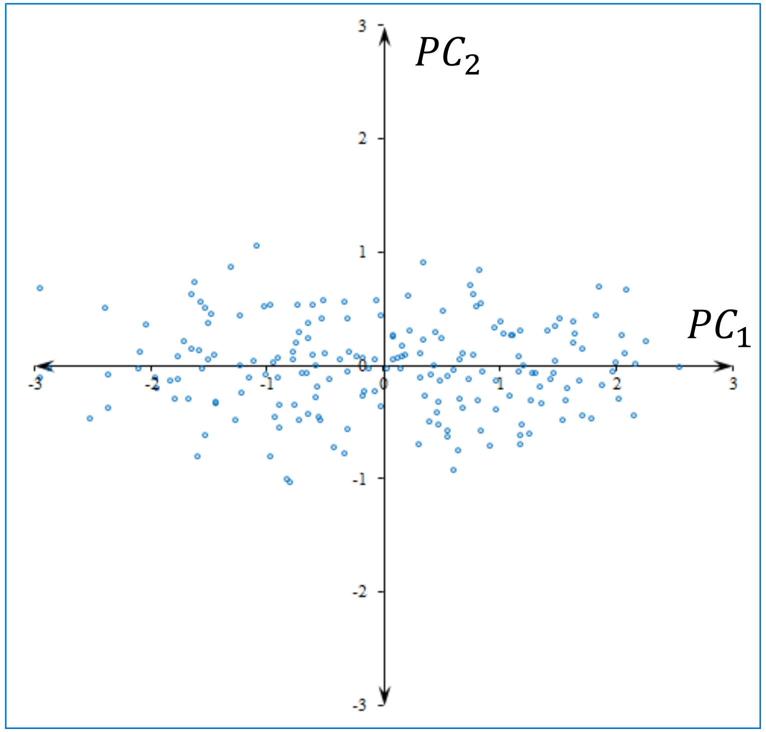
주성분 분석 결과 예시

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



주성분 분석 결과 예시

- 주성분 PC_1 과 PC_2를 새로운 좌표축으로 사용하였을 때.



주성분 계산 방법

- 주성분은 다음 두 가지 방법으로 구할 수 있다.

1. 데이터 행렬에 특이값 분해 (Singular Value Decomposition, SVD)를 적용한다.
2. 공분산 행렬 또는 상관계수 행렬에 고유값분해 (Eigenvalue Decomposition, ED)를 적용한다.

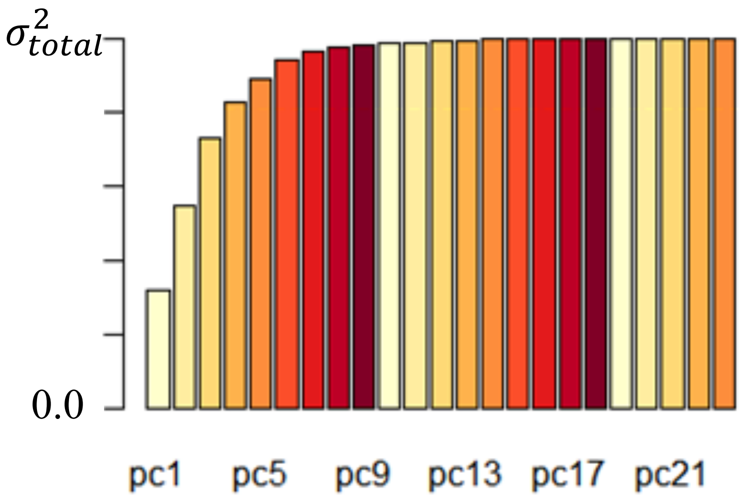
⇒ 만약에 변수를 표준화 했다면 공분산행렬 = 상관계수행렬과 같다.

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

누적 분산

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

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



주성분 활용

키포인트

- 차원 축소.
- 고차원 클러스터의 시각화.

차원축소의 원리

- 주성분의 개수는 원래 차원의 수 (변수의 개수)와 같다.
- 주성분을 새로운 좌표축으로 사용할 수 있다.
- 원 변수 X_i는 주성분들의 선형조합으로 표현할 수 있다.

$$X_i = \beta_{1,i}PC_1 + \beta_{2,i}PC_2 + \cdots + \beta_{k,i}PC_k$$

- 그런데 주성분은 분산의 크기 순서대로 정렬되어 있다: $\sigma_1^2 > \sigma_2^2 > \sigma_3^2 > \ldots$
- 분산이 작은 주성분을 우선적으로 없앨 수 있다, $(q < k)$:

$$X_i \approx \beta_{1,i}PC_1 + \beta_{2,i}PC_2 + \cdots + \beta_{q,i}PC_q$$ "Reduced Dimensional Input"

차원축소의 장단점

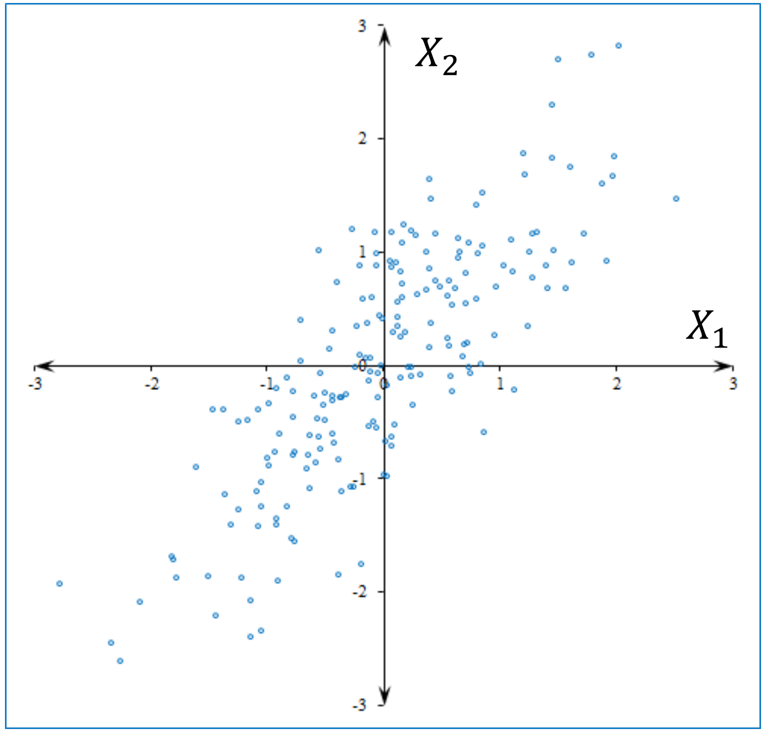
장점	단점
✓ 가장 뚜렷한 특징만 뽑아낸다. ✓ 데이터 표현의 간소화. ✓ 연산, 메모리 부하 감소.	✓ 직관적인 해석은 어렵다. ✓ 작은 디테일은 손상된다.

차원축소 결과

- Reduced dimensional input: 차원축소 후의 데이터 좌표.
- Loading: 정규화된 주성분 벡터.
- Communality: 차원축소 후 원래의 변수가 어느정도 복원 되었는가.

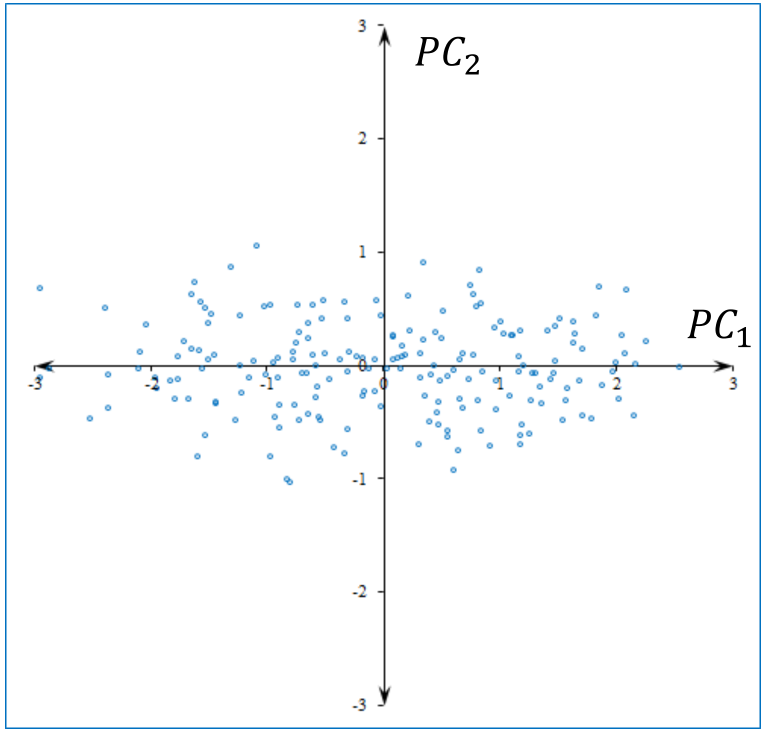
차원축소 예시

- 다음과 같이 2차원 데이터가 분포되어 있다고 가정한다.



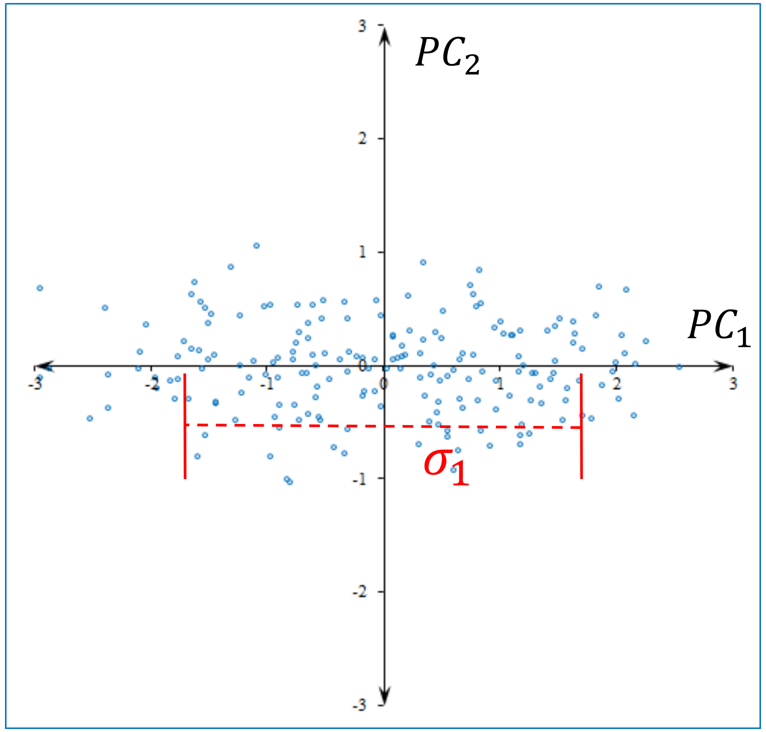
차원축소 예시

- 주성분 PC_1 과 PC_2를 새로운 좌표축으로 사용하였을 때.



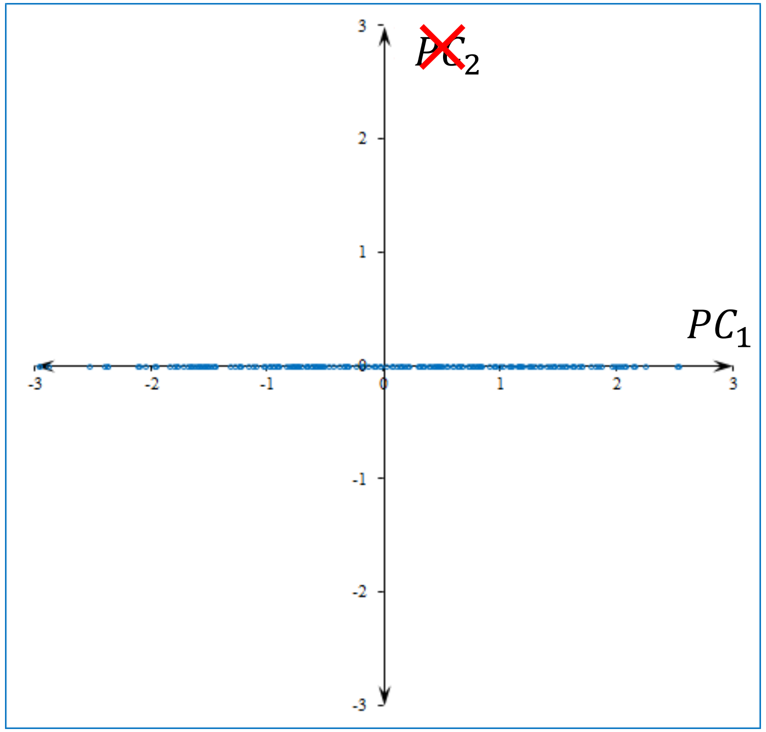
차원축소 예시

- PC_1는 변동이 가장 큰 방향.



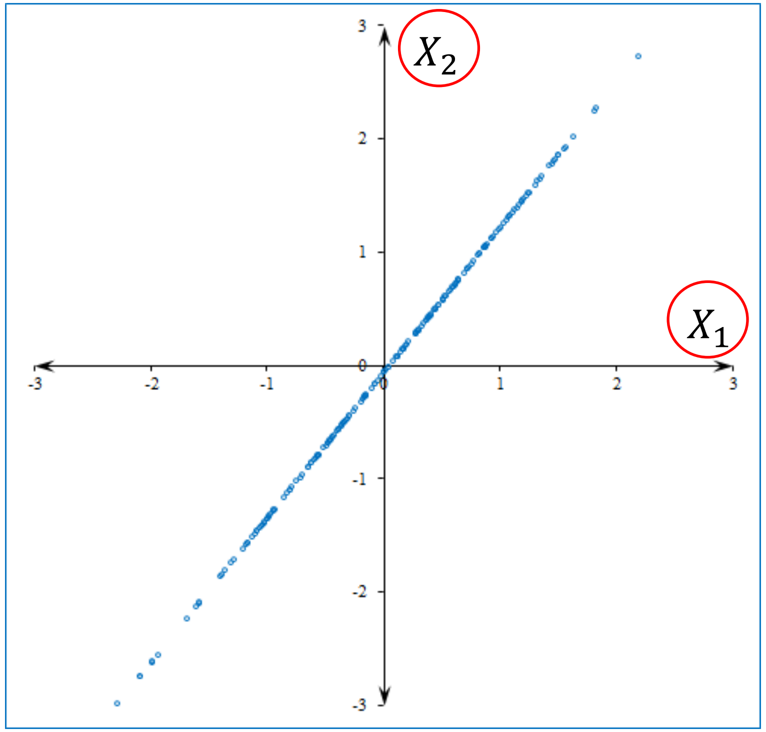
차원축소 예시

- PC_2방향으로 차원 축소.



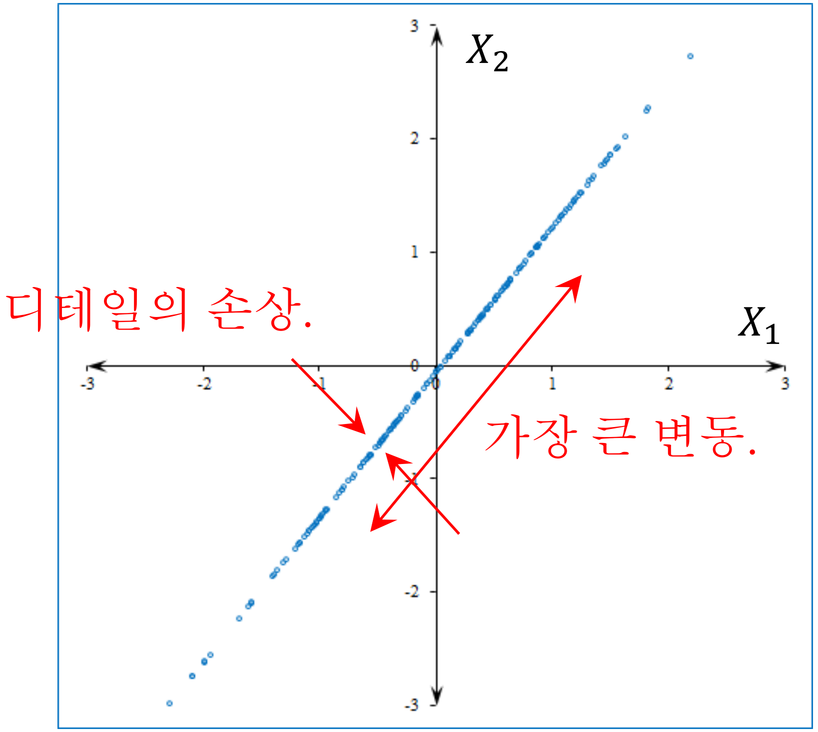
차원축소 예시

- 원 좌표축으로 돌아와 본다 (reduced dimensional input).



차원축소 예시

- 원 좌표축으로 돌아와 본다 (reduced dimensional input).



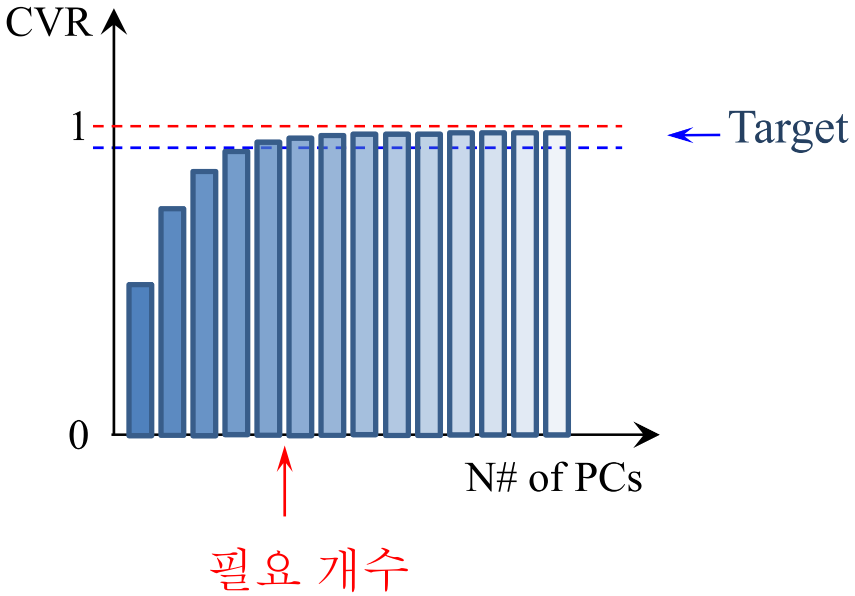
차원 축소 기준

- 전체 분산: $\sigma_{total}^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \cdots + \sigma_k^2$
- 누적 분산의 비율 (Cumulative Variance Ratio):

$$CVR_1 = \frac{\sigma_1^2}{\sigma_{total}^2}$$

$$CVR_2 = \frac{\sigma_1^2 + \sigma_2^2}{\sigma_{total}^2}$$

$$\vdots$$



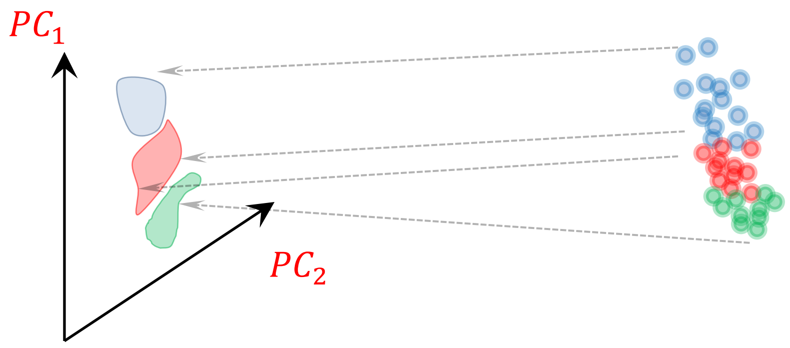
- CVR 목표치에 해당하는 만큼의 주성분의 개수를 유지한다.

고차원 클러스터의 시각화

- 고차원에서의 클러스터/군집을 평면 (2D)에 투영하여 시각화하려고 한다.
- 분산이 큰 순서대로 두개의 주성분 (PC_1 과 PC_2)을 사용한다.
- 이 두 주성분은 새로운 2D 좌표계를 정의한다.
- 고차원 데이터 좌표를 새로운 좌표계로 투영한다.

고차원 클러스터의 시각화

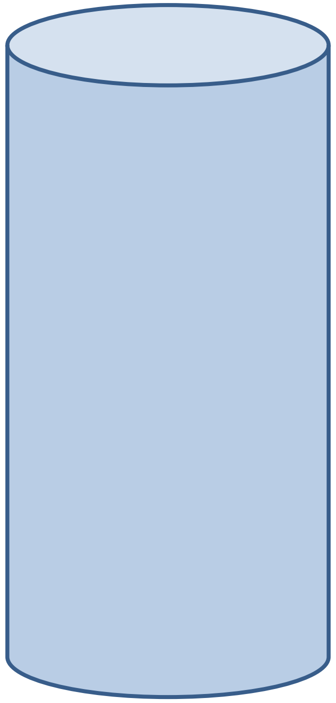
- PC_1 와 PC_2는 변동 (분산) 크기로는 각각 1등과 2등이다.
- PC_1 와 PC_2는 고차원 좌표를 투사하기에 적합한 평면을 정의한다.



$\Rightarrow$ Transformed Scores를 가져다 그대로 사용하면 되기 때문에 적용이 쉽다!

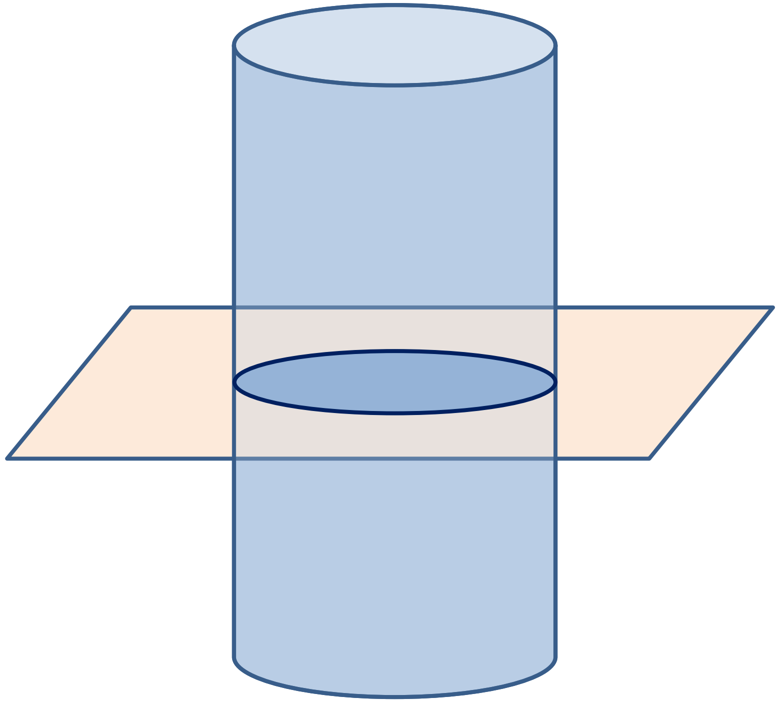
고차원 클러스터의 시각화 원리

- 데이터 좌표가 실린더 형상으로 분포해 있다고 가정해 본다.



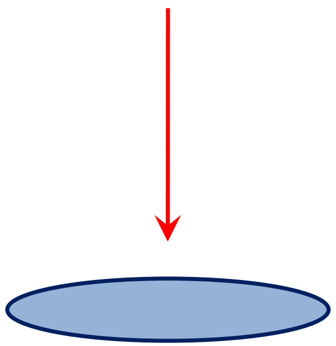
고차원 클러스터의 시각화 원리

- 2D 단면을 다음과 같이 자른다.



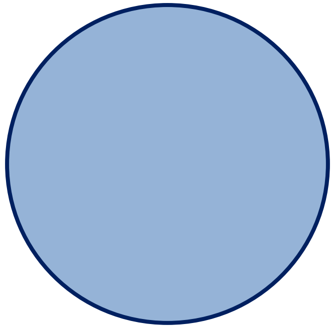
고차원 클러스터의 시각화 원리

- 단면을 정면에서 바라 본다.



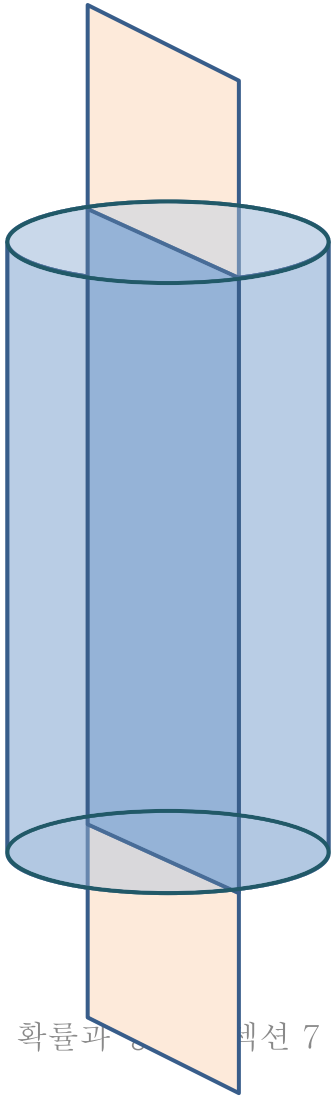
고차원 클러스터의 시각화 원리

- 그러면 이렇게 보이겠죠^^



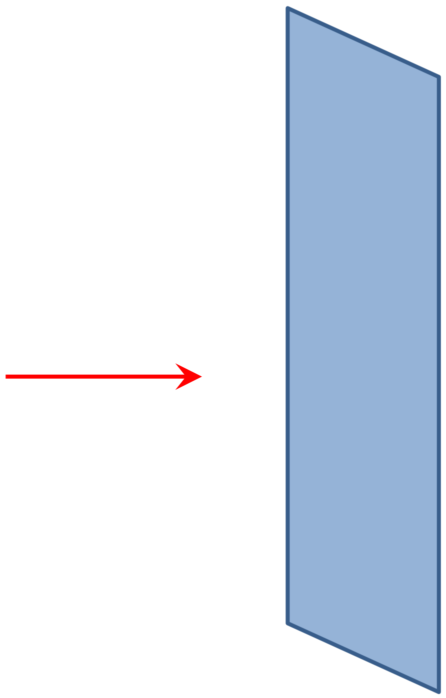
고차원 클러스터의 시각화 원리

- 또다른 2D 단면을 자른다.



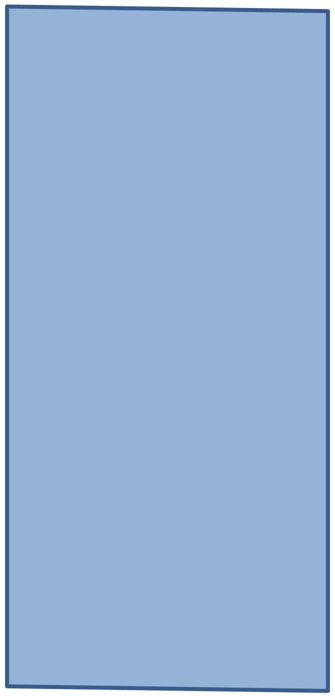
고차원 클러스터의 시각화 원리

- 단면을 정면에서 바라 본다.

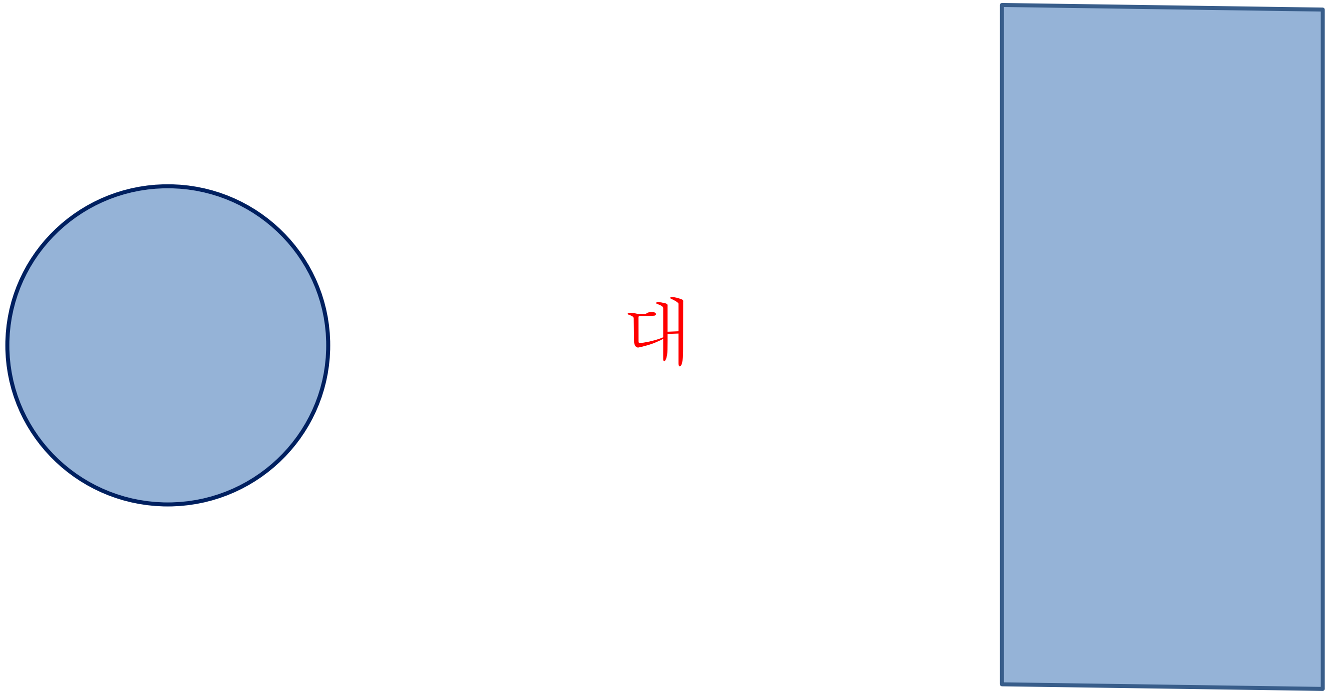


고차원 클러스터의 시각화 원리

- 그러면 이렇게 보이겠죠^^



고차원 클러스터의 시각화 원리



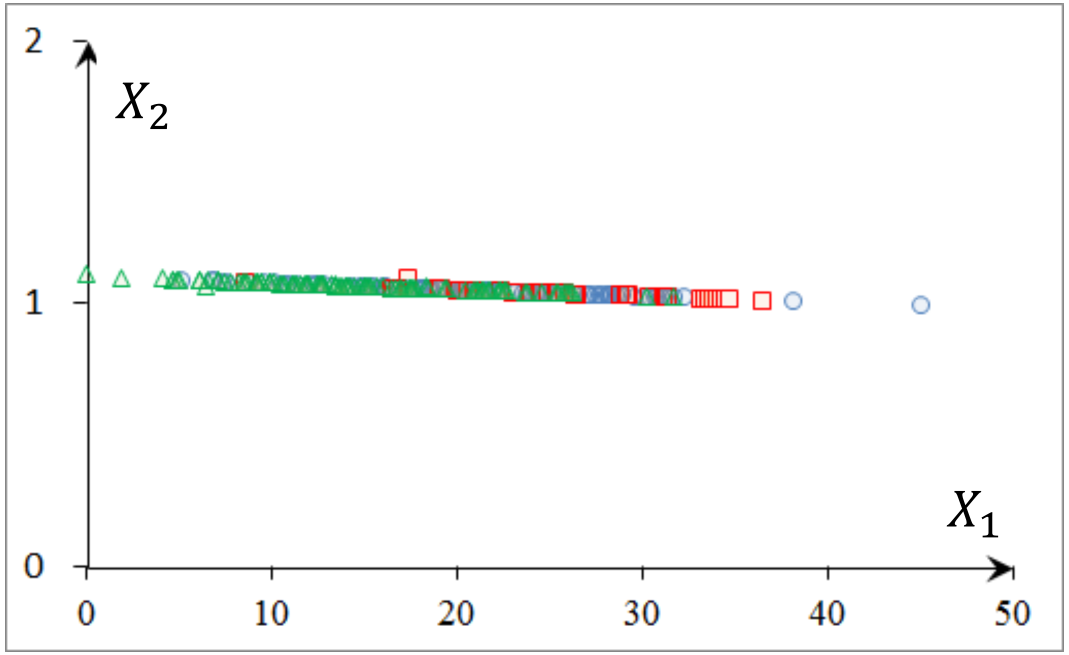
- ✓ 더 넓게 퍼져있는 단면.
- ✓ 투영하기에 좋음.

고차원 클리스터의 시각화 원리

- 단면을 잘 선택하면 시각화에 유리하다.
- PC_1 과 PC_2는 데이터가 가장 넓게 퍼져있는 단면을 정의한다!

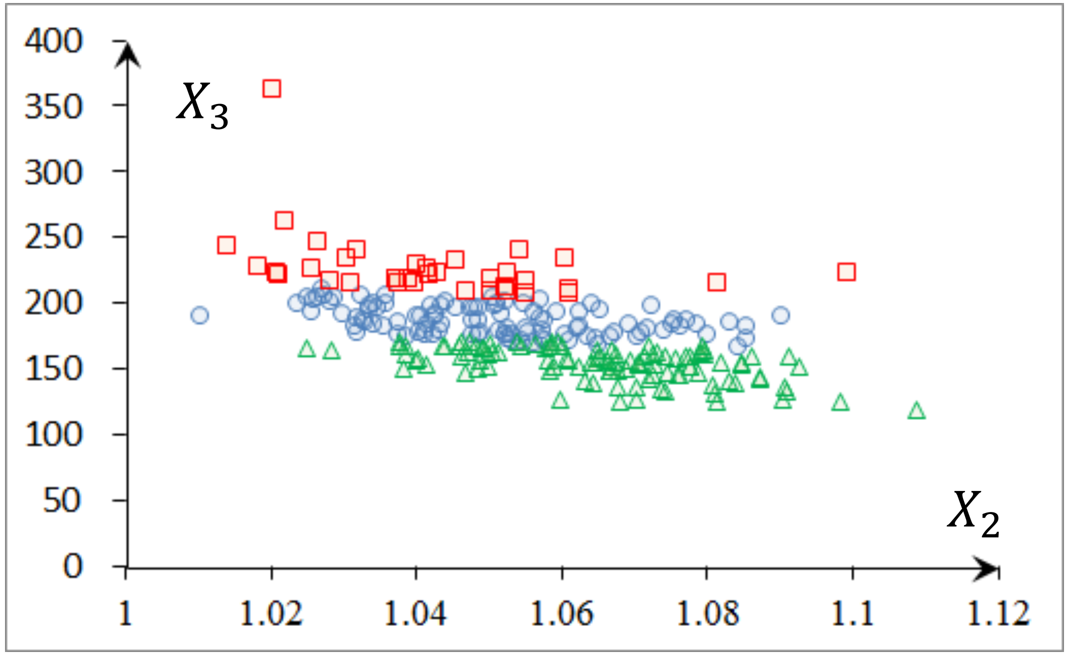
고차원 클러스터의 시각화 예시

- 원 좌표 X_1과 X_2 사용.



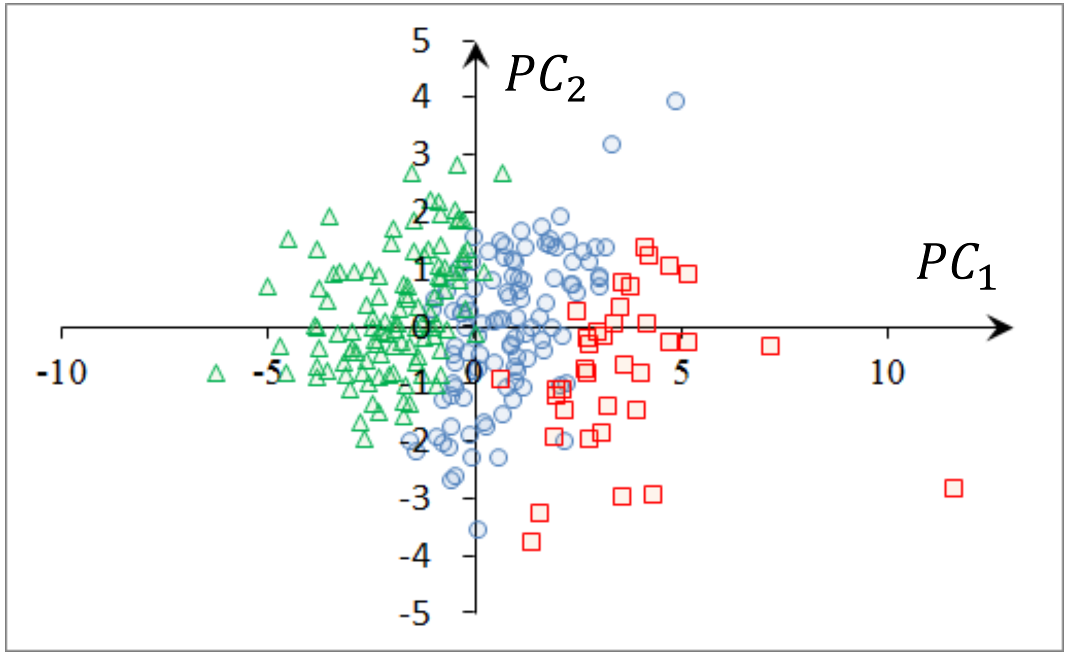
고차원 클러스터의 시각화 예시

- 원 좌표 X_2과 X_3 사용.



고차원 클러스터의 시각화 예시

- PC_1과 PC_2로 정의되는 단면에 투영.



요인분석

키포인트

- 요인분석 (Factor Analysis, FA).
- 주성분과 요인의 비교.
- 직교회전과 사교회전의 원리.

요인분석 & 주성분 분석 비교

- 먼저 $X_1, X_2, \ldots, X_p$와 같이 p개의 변수가 있다고 가정한다.
- 주성분 분석으로 p개의 주성분 $PC_1, PC_2, \ldots, PC_p$을 얻을 수 있다.

$\Rightarrow$ 개개 주성분은 원 변수의 선형 조합이다.

$$PC_i = \alpha_{1,i}X_1 + \alpha_{2,i}X_2 + \cdots + \alpha_{p,i}X_p$$

$\Rightarrow$ 반대로 원 변수는 주성분의 선형 조합으로 표현할 수 있다.

$$X_i = \beta_{1,i}PC_1 + \beta_{2,i}PC_2 + \cdots + \beta_{p,i}PC_p$$

요인분석 & 주성분 분석 비교

$\Rightarrow$ 차원축소할 때에는 적은 개수 $(q < p)$ 의 주성분만을 사용한다.

$$X_i \approx \beta_{1,i}PC_1 + \beta_{2,i}PC_2 + \cdots + \beta_{q,i}PC_q$$

$\Rightarrow$ 주성분은 분산의 크기로 정렬되어 있으니 첫 몇개만 사용한다.

$$\sigma_1^2 > \sigma_2^2 > \sigma_3^2 > \ldots$$

or

$$\sigma_1 > \sigma_2 > \sigma_3 > \cdots$$

요인분석 & 주성분 분석 비교

- 요인분석도 적의 수의 잠재 요인으로 원 변수를 설명하는 것이 목표이다.

$$X_i = \beta_{0,i} + \beta_{1,i}F_1 + \beta_{2,i}F_2 + \cdots + \beta_{q,i}F_q + \varepsilon_i$$

$\Rightarrow$ $F_1, F_2, \ldots, F_q$ 는 공통적으로 나타나는 잠재 요인.

$\Rightarrow$ 주성분과는 다르게 요인의 개수 (q) 를 먼저 정해놓고 구한다.

$\Rightarrow$ ε_i는 변수 X_i의 개별 오차이며 $F_1, F_2, \ldots, F_q$와는 독립적이다.

$\Rightarrow$ 가중치 $\beta_{1,i}, \beta_{2,i}, \ldots, \beta_{q,i}$를 loading (요인 부하량) 이라 부른다.

$\Rightarrow$ 요인으로 설명 가능한 변수의 변동 비율을 communality (공통성) 이라 부른다.

요인분석 & 주성분 분석 비교

- 다음과 같은 유사성이 있다.

  ⇒ 주성분 (PC) ≈ 요인 (factor).

  ⇒ 차원축소를 주된 목표로 한다. 차원 축소 후 주성분의 개수 ≈ 요인의 개수.

  ⇒ 원 변수 사이의 공유되는 구조 발견을 목적으로 한다.

요인분석 & 주성분 분석 비교

- 하지만 다음과 같은 차이점이 있다.

요인에는 주관적인 의미 부여 ⇔ 주성분은 객관적 해석 우선시

요인은 다 동등한 취급을 받음 ⇔ 주성분은 분산의 크기로 차등을 둠

요인의 개수는 분석 전에 정해 둠 ⇔ 주성분의 개수 = 원 변수의 개수

요인분석 & 주성분 분석 비교

- 하지만 다음과 같은 차이점이 있다.

요인에는 주관적인 의미 부여 ⇔ 주성분은 객관적 해석 우선시

요인은 다 동등한 취급을 받음 ⇔ 주성분은 분산의 크기로 차등을 둠

요인의 개수는 분석 전에 정해 둠 ⇔ 주성분의 개수 = 원 변수의 개수

주관적 또는 상관계수
행렬의 고유값 크기를
분석하여 정한다.

요인분석 & 주성분 분석 비교

- 하지만 다음과 같은 차이점이 있다.

요인에는 주관적인 의미 부여 ⇔ 주성분은 객관적 해석 우선시

요인은 다 동등한 취급을 받음 ⇔ 주성분은 분산의 크기로 차등을 둠

요인의 개수는 분석 전에 정해 둠 ⇔ 주성분의 개수 = 원 변수의 개수

차원 축소시에는 누적분산 비율을 고려한다.

요인분석 & 주성분 분석 비교

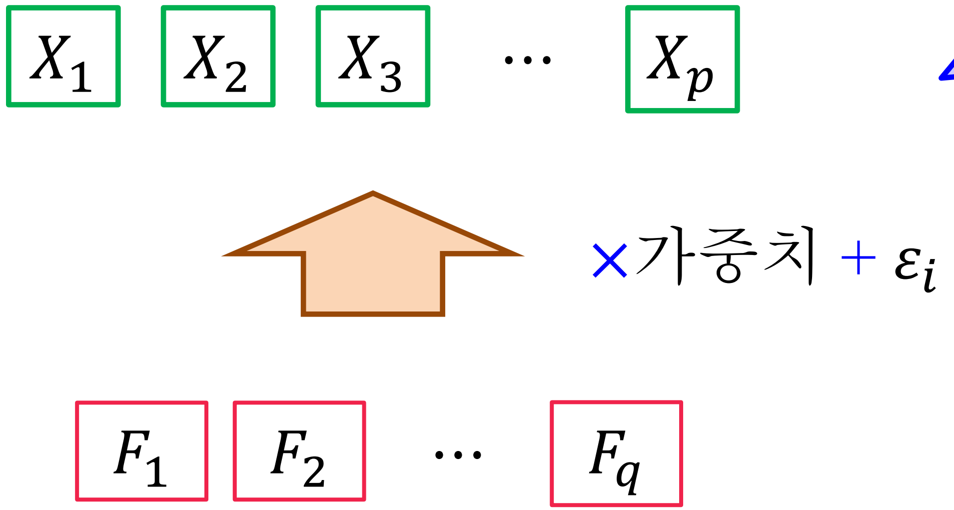
PCA는 원 변수의 선형 조합으로 주성분 계산을 우선시.

X_1 X_2 X_3 $\cdots$ X_p

×가중치

PC_1 PC_2 PC_3 $\cdots$ PC_p

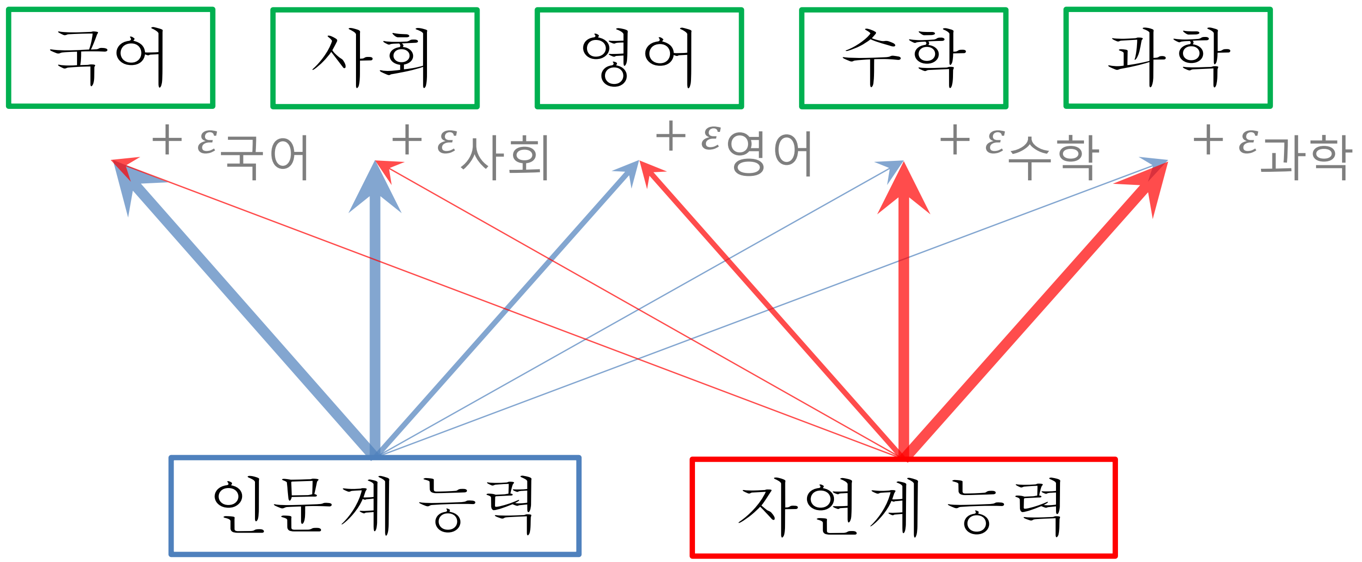
요인분석 & 주성분 분석 비교



FA는 잠재 요인을 발굴해서 원 변수의 주관적 설명을 우선시.

요인분석 예시

예). 다섯 개의 관측 변수를 설명하는 두 개의 요인.



회전

- 요인 부하량을 가지고 요인의 주관적 해석과 naming을 한다.

$$X_i = \beta_{0,i} + \beta_{1,i}F_1 + \beta_{2,i}F_2 + \cdots + \beta_{q,i}F_q + \varepsilon_i$$

예). 국어 $\cong 0.9\,F_1 + 0.1\,F_2$

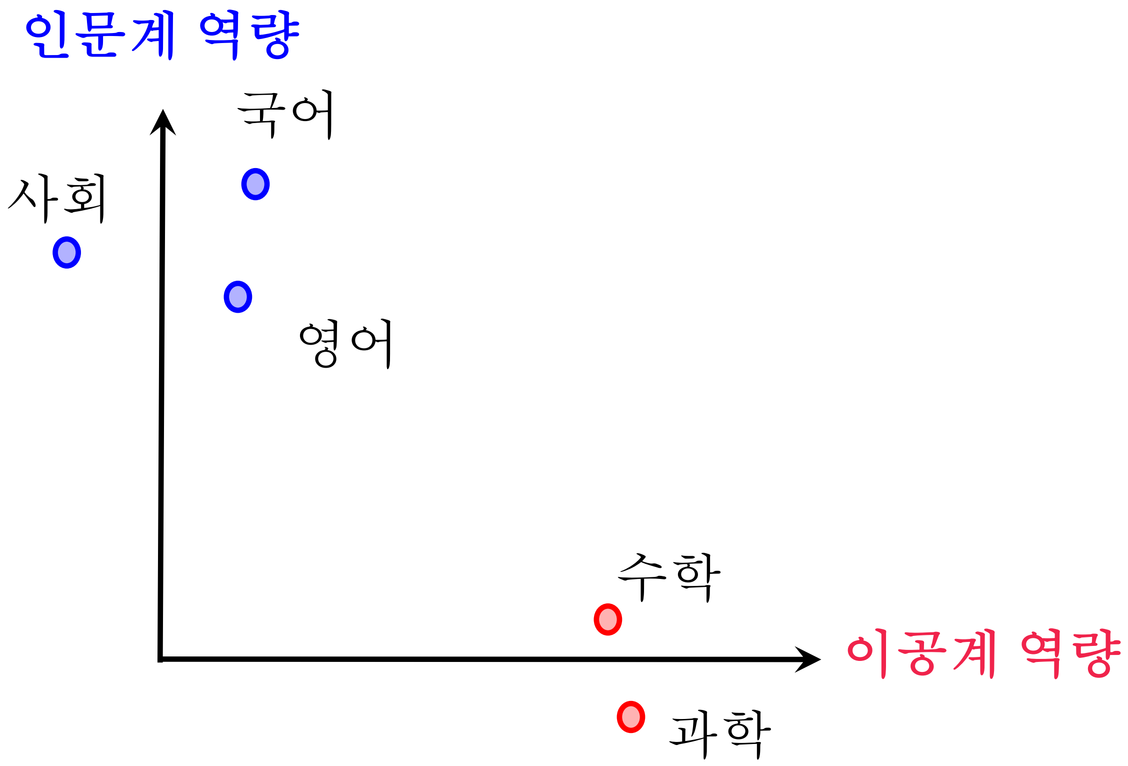
과학 $\cong 0.1\,F_1 + 0.9\,F_2$

F_1는 "인문계 역량" 이라 해석하고 naming.

F_2는 "이공계 역량" 이라 해석하고 naming.

회전

- 요인을 새로운 좌표축으로 정하고 부하량을 사용하여 원 변수를 plot할 수 있다.



회전

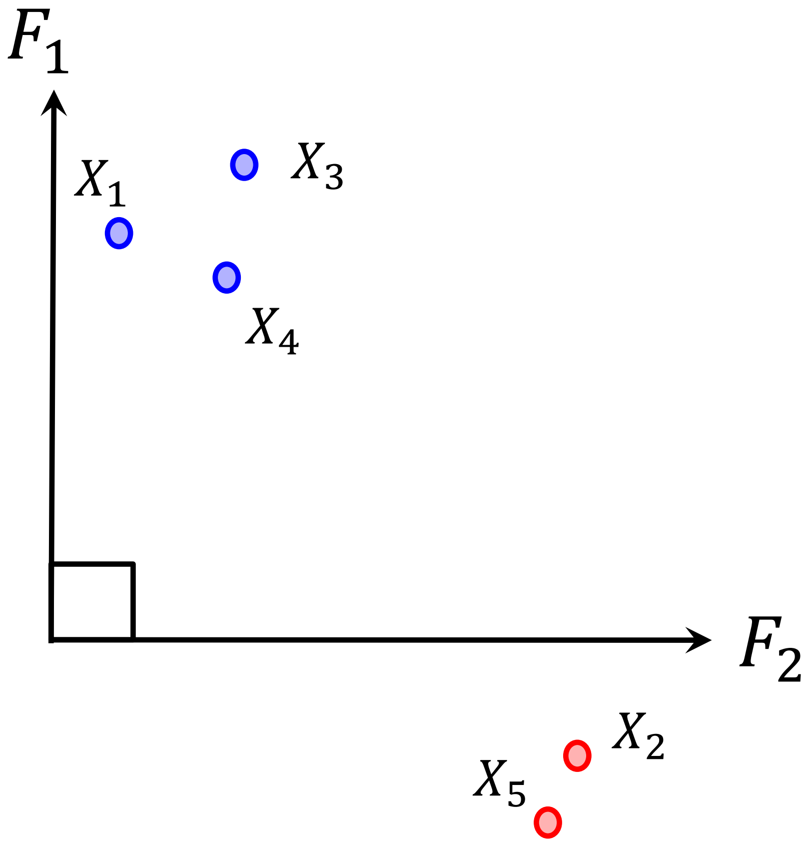
- 그런데 대다수의 경우 최적화된 해석을 위해서 좌표축의 회전이 필요하게 된다.

  ⇒ 직교회전 : 좌표축의 직교관계 유지. Varimax와 같은 방법.

  ⇒ 사교회전 : 좌표축이 별개로 회전. Promax, Oblimin, 등의 방법.

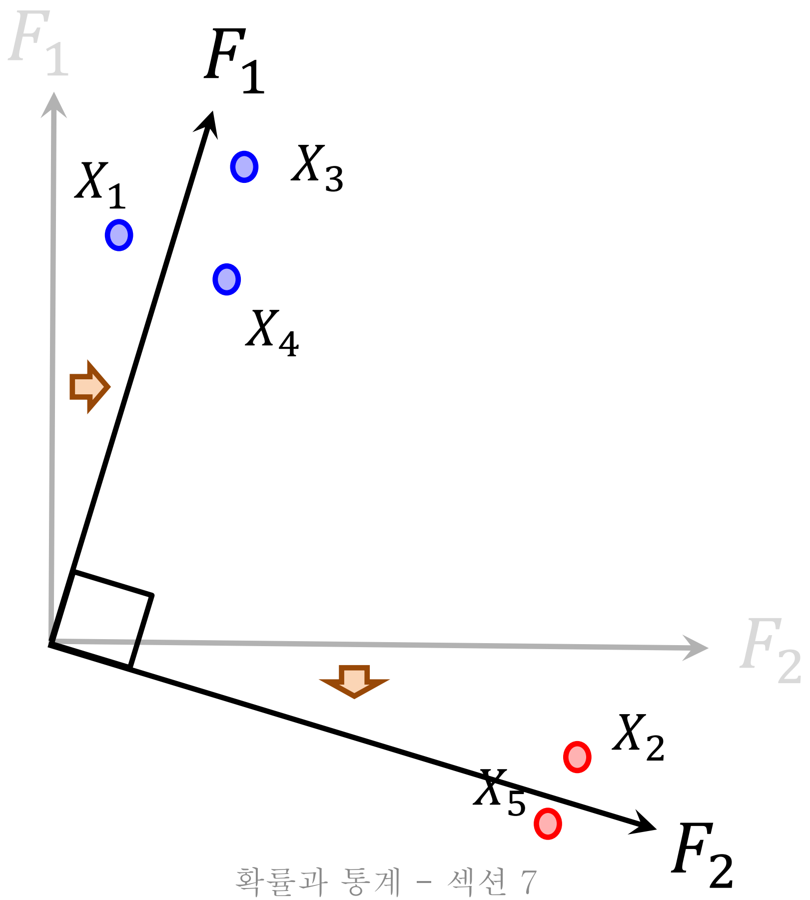
직교회전의 원리

- 다음과 같이 좌표축 (요인)의 방향이 최적화 되어 있지 않은 상황을 전제해 본다.



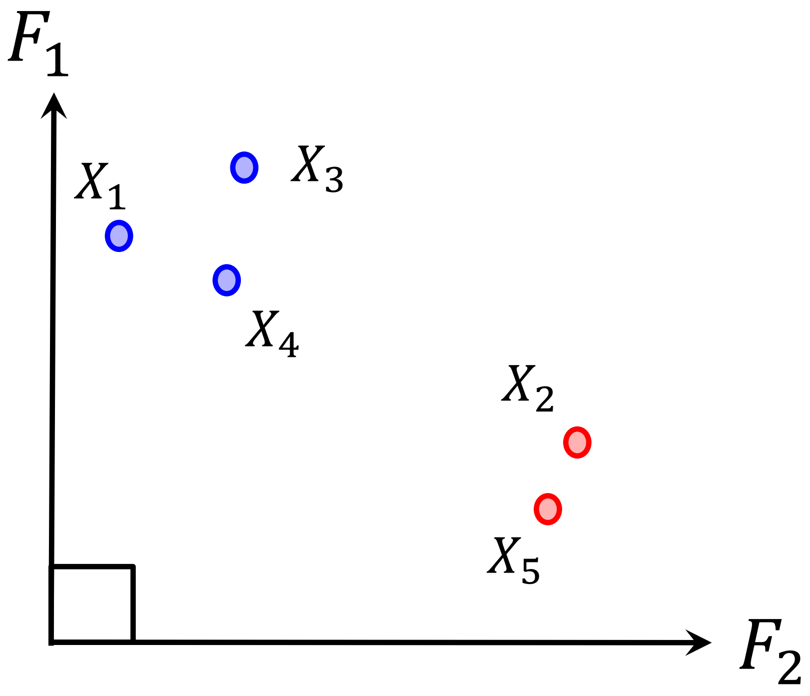
직교회전의 원리

- 직교회전 후 요인을 해석하고 naming하기 수월해 진다.



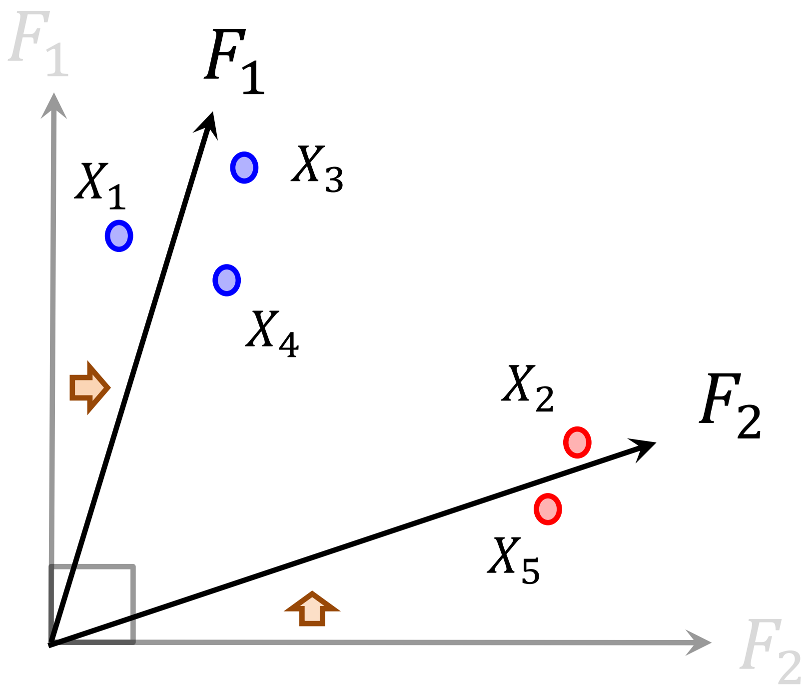
사교회전의 원리

- 또다시 좌표축 (요인)의 방향이 최적화 되어 있지 않은 상황을 전제해 본다.



사교회전의 원리

- 사교회전 후 요인을 해석하고 naming하기 수월해 진다.



요인분석 순서

- R과 같은 전문 툴을 사용하여 쉽게 실행할 수 있다.

요인수 결정: 상관계수 행렬의 고유값, 등.

↓

계산 실행: 최대 우도법, 최소 제곱법, 주성분, 등.

↓

요인의 회전: 직교, 사교, 등.

↓

요인의 해석 및 naming.

끝