

# 클린업 3주차

선형대수학팀(3팀)

박서영  
김민주  
이윤희  
이지연  
황정현

# INDEX

---

## 0. Review

1. 특이값 분해(SVD)

2. 주성분분석(PCA)

3. 요인분석(FA)

4. 잠재의미분석(LSA)

## 고유값과 고유벡터(Eigenvalue & Eigenvector)

- $n \times n$  정방행렬  $A$ 에서  $\lambda$ 는 스칼라,  $x$ 는 0벡터가 아닌 벡터일 때,

$$\underbrace{Ax}_{\text{고유벡터}} = \underbrace{\lambda x}_{\text{고유값}}$$

- $x$ 가  $Ax$ 로 선형변환 되었을 때,  $Ax$ 가  $x$ 의 스칼라곱( $\lambda x$ ), 즉  $x$ 와 평행한 방향으로 뻗어가는 벡터로 변환되는 경우
- 서로 다른 고유값에 대응하는 고유벡터들은 선형독립

## 고유값&고유벡터와 $Ax=0$

- $n \times n$ 의 정방행렬  $A$ 에 대하여

$(A - \lambda I)x = 0$ 의 해인 벡터  $x$ 들은  $\lambda$ 에 대응하는 고유벡터!

이때 벡터  $x$ 의 집합은 **eigenspace** 라고 하며,

$(A - \lambda I)$ 의 **Null space** 이자  $\mathbb{R}^n$  의 선형부분공간이 된다.

$$(A - \lambda I)x = 0$$

## 대각화(Diagonalization)

- $n \times n$  행렬  $A$ 가  $n$ 개의 선형독립인 고유벡터를 가질 때

$$A = P D P^{-1}$$

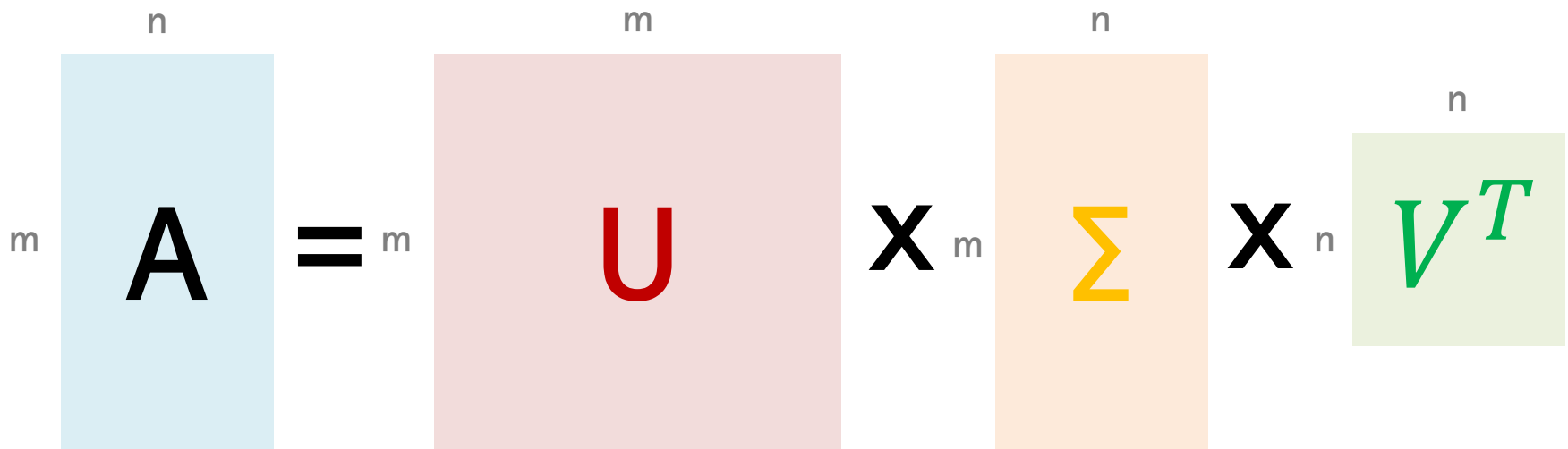
$P$ 의 열은  $n$ 개의 선형독립인 고유벡터들

$D$ 는  $P$ 의 고유벡터에 대응하는 고유값이 대각원소인 대각행렬

※ 대각화의 장점 :  $A$ 의 거듭제곱( $A^k = P D^k P^{-1}$ ) 을 쉽게 구할 수 있다!

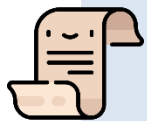
## 특이값 분해(SVD\_Singular Value Decomposition)

- 임의의  $m \times n$  행렬  $A$ 에 대하여 다음과 같이 분해하는 방법을 '특이값 분해'라고 한다.


$$\begin{matrix} & n \\ m & \begin{matrix} A \end{matrix} \end{matrix} = \begin{matrix} & m \\ m & \begin{matrix} U \end{matrix} \end{matrix} \times \begin{matrix} & m \\ m & \begin{matrix} X \end{matrix} \end{matrix} \begin{matrix} & n \\ n & \begin{matrix} \Sigma \end{matrix} \end{matrix} \times \begin{matrix} & n \\ n & \begin{matrix} X \end{matrix} \end{matrix} \begin{matrix} & n \\ n & \begin{matrix} V^T \end{matrix} \end{matrix}$$

## 특이값 분해(SVD)

- 특이값 분해의 기하학적 의미



**관점1** : 길이가 1인  $A$ 의 고유벡터  $x$  중에서  
 $\|Ax\|$  를 최대화하는  $x$ 와 그 때의  $\|Ax\|$  값 찾기



**관점2** : 직교하는 벡터집합 중 선형변환 후에도  
직교가 유지되는 직교집합 찾기

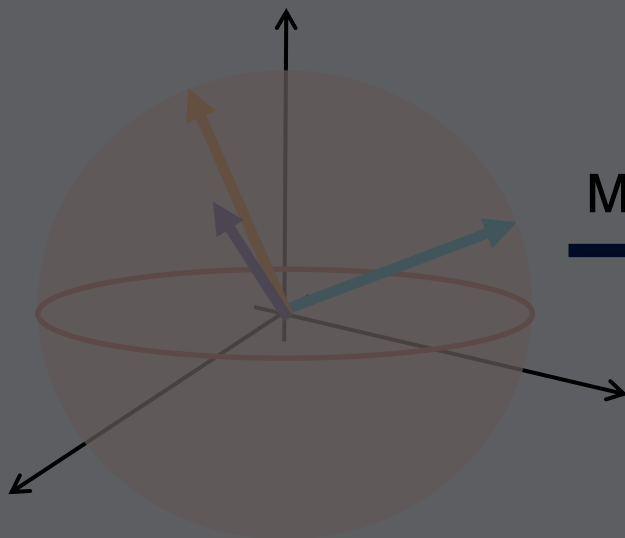
## 특이값 분해(SVD)의 기하학적 의미 - 관점1

- $m \times n$  행렬  $A$ 에 대해  $\|Ax\|$ 를 최대화하는 값과 벡터를 찾기 위해  $\|Ax\|^2$ 를 최대화 하는 값과 벡터를 찾음

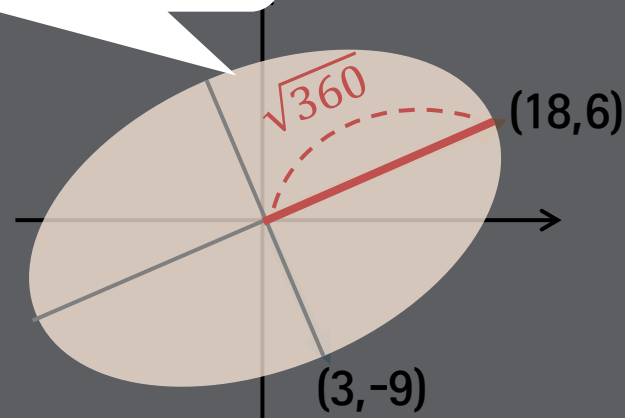
예시

$A = \begin{bmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{bmatrix}$  이고,  $x$ 가  $\mathbb{R}^3$ 의 단위벡터일 때  $\|Ax\|$ 를 최대화하는 값과 벡터 찾기

선형변환된  $Ax$ 의 최대 길이는  
 $\sqrt{\lambda}$ , 즉  $\sqrt{360}$ 이다!



Multiplication by  $A$

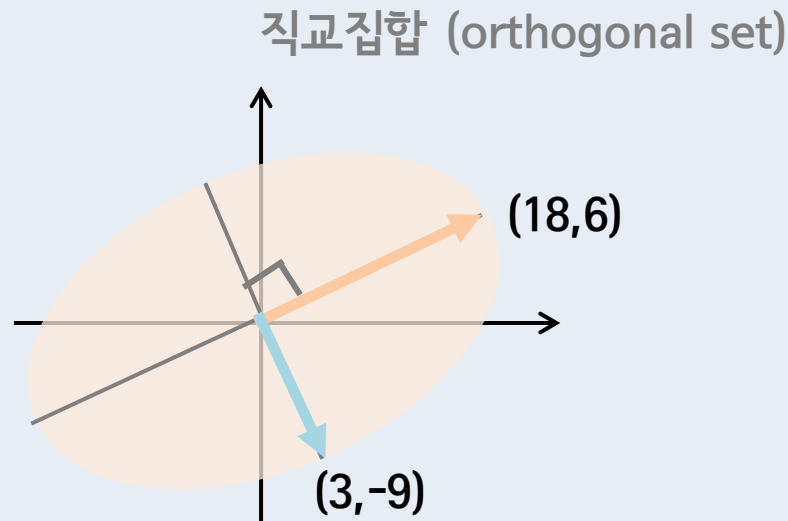




## 특이값 분해(SVD)의 기하학적 의미 - 관점2

- 선형변환 후에 크기는 변해도 직교는 유지되는 직교 집합

예시 )  $A = \begin{bmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{bmatrix}$  일 때, 선형변환 후의 고유벡터



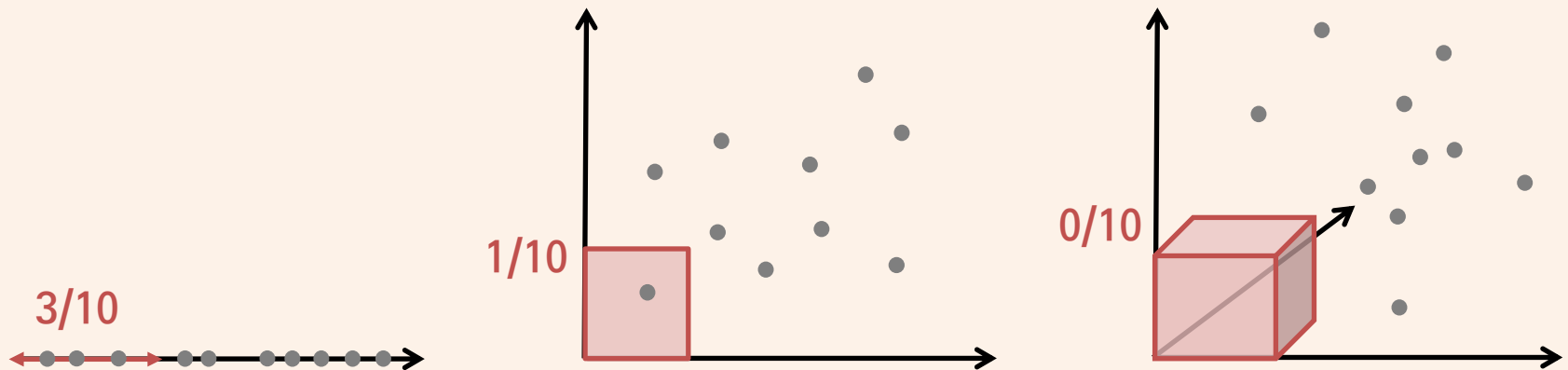
## 특이값 분해(SVD)

- 임의의  $m \times n$  행렬  $A$ 에 대하여
- 몇 개의 singular value를 이용해  $A'$  라는 행렬로 부분 복원시킬 수 있다.

$$\begin{matrix} & p \\ m & U' \end{matrix} \times \begin{matrix} & p \\ p & \Sigma' \end{matrix} \times \begin{matrix} & n \\ p & V'^T \end{matrix} = \begin{matrix} & n \\ m & A' \end{matrix}$$

## 차원의 저주

- 변수의 수가 늘어나는 문제, 즉 차원이 늘어나는 문제를 **차원의 저주(Curse of dimensionality)** 라고 한다.
- 데이터를 표현할 공간의 차원이 늘어날수록 주어진 데이터, 즉 표본이 모집단에서 차지하는 비중이 급격히 감소하기 때문에 과적합(overfitting)이 발생할 수 있다.



## 차원축소

변수선택  
(feature selection)

변수추출  
(feature extraction)

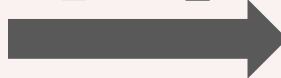


- 모든 변수를 조합해 데이터를 잘 표현할 수 있는 중요 성분을 가진 새로운 변수를 추출하는 것
- 대표적 기법으로 주성분 분석이 있음.

다차원 변수

키  
몸무게  
머리길이  
눈 크기  
코 크기  
신발사이즈  
손톱 길이

변수추출



새로운 변수

가  
나  
다

## 주성분분석(Principal Component Analysis)

※ 주성분의 개수는 기존 변수의 개수보다 작기 때문에  
기존 데이터의 **정보**를 잃게 된다.

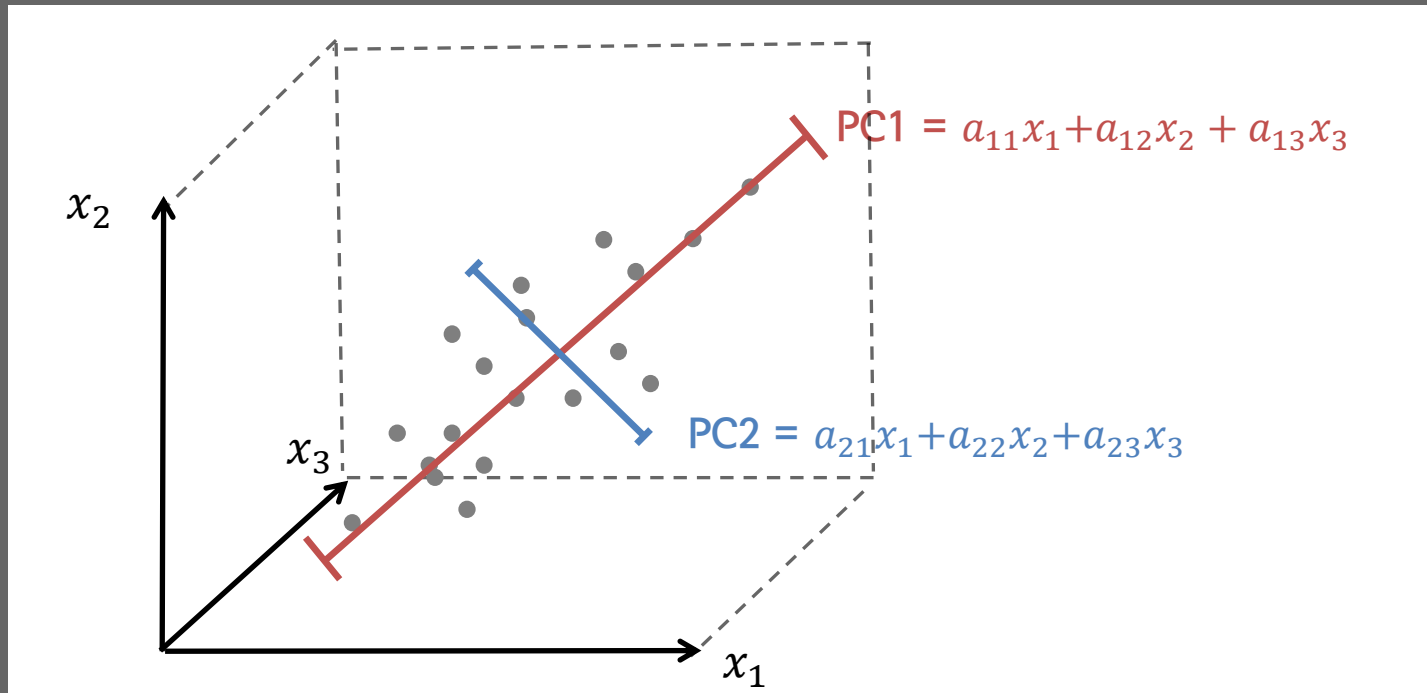
따라서 **가능한 한 가장 적게 정보를 잃도록**  
**주성분과 그 개수를 선택**해야 한다!



다중공선성 문제를 **해결**하여 회귀분석의 조건 만족!

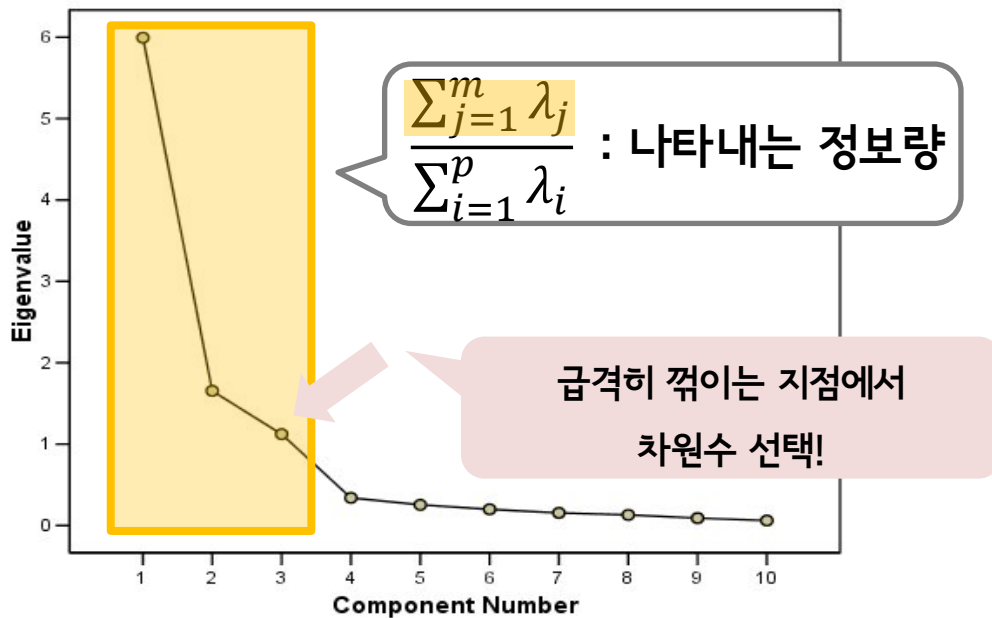
## 공분산 행렬(Covariance Matrix)

차원을 축소할 때, 공분산행렬의 **큰 고유값에 대응하는 고유벡터로 데이터를 projection 시킨 것을 주성분으로 택하면**  
정보손실을 최소화할 수 있다!



## 주성분 개수 결정

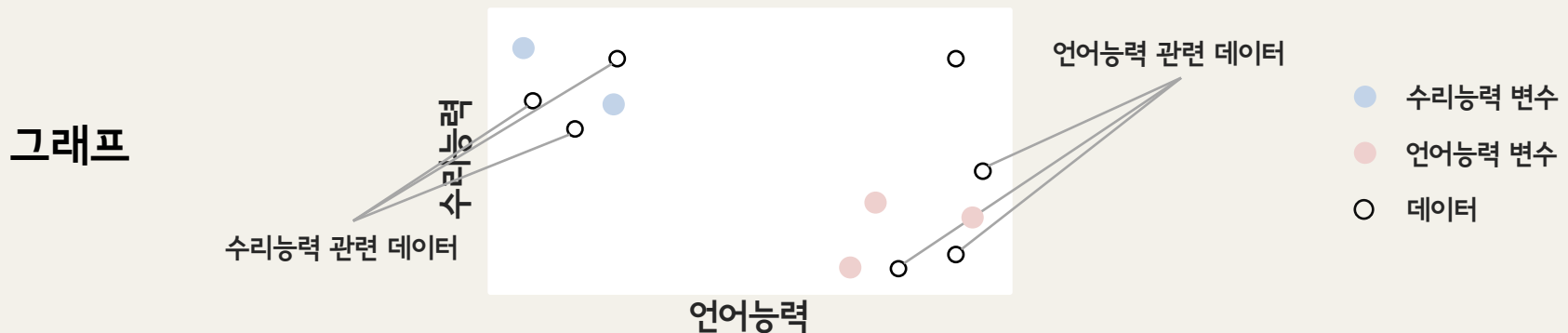
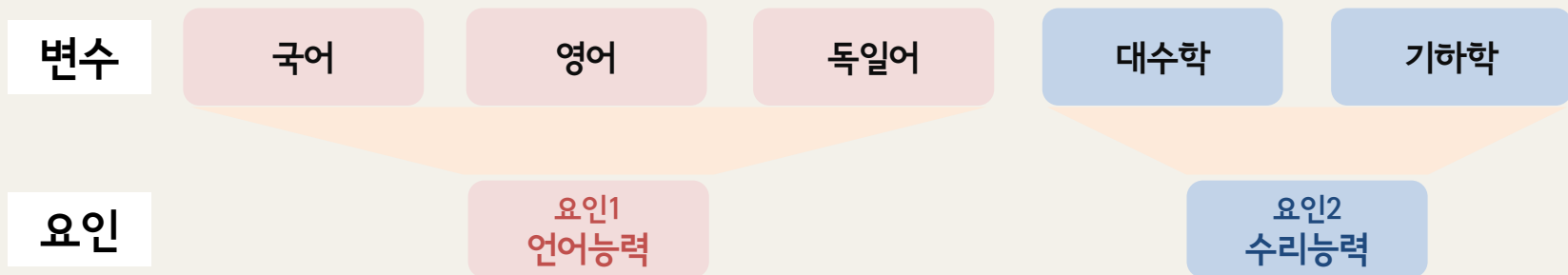
- 주성분 개수 선택 기준: 사용할 주성분이 데이터의 정보(=분산)를 얼마나 갖고 있는가?
- 선택할 주성분들의 총 분산이 크면서도 차원을 축소할 수 있도록 개수 선택
- 일반적으로 **Scree plot**을 그려서 시각적으로 결정!



꺾이는 지점 기준 왼쪽 주성분들의  
 분산의 합( $\sum_{j=1}^m \lambda_j$ ),  
 즉 해당 주성분들이 나타내는  
 데이터의 정보량이 충분하다고  
 판단!

## 요인분석(Factor Analysis)

- 변수들의 상관관계를 고려하여 **내재된 요인을 추출**해 요인 별로 변수를 묶어주는 방법
- 변수들이 몇 개의 요인에 영향을 받는가를 알아보는 것







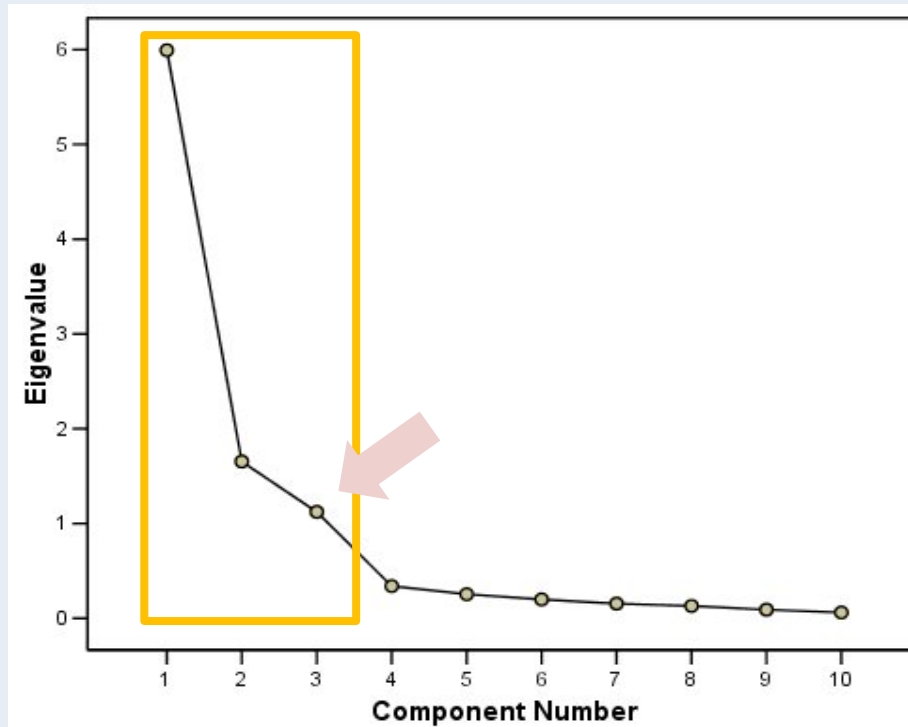
## 주성분분석 vs 요인분석

※ 정리

	주성분분석	요인분석
변수	변수로 주성분 설명	요인으로 변수 설명
분석방법	종속변수 고려 O	종속변수 고려 X
분석목적	모델의 정확성	변수의 의미와 특성 파악
생성 변수 이름	의미 X	의미 O
생성 변수 간 관계	중요도 다름	대등한 관계
통계 모형	non-parametric	parametric

## 요인개수 결정하기

- PCA의 경우처럼 요인분석도 **scree plot**을 통해 변수 개수를 결정한다.
- 원 데이터의 공분산행렬의 고유값  $\lambda_j$ 는  $F_j$ 의 분산을 의미한다.



## 잠재의미분석(Latent Semantic Analysis)



자연어처리 기법 중 하나인 **토픽 모델링**을 수행하기 위한 분석 방법



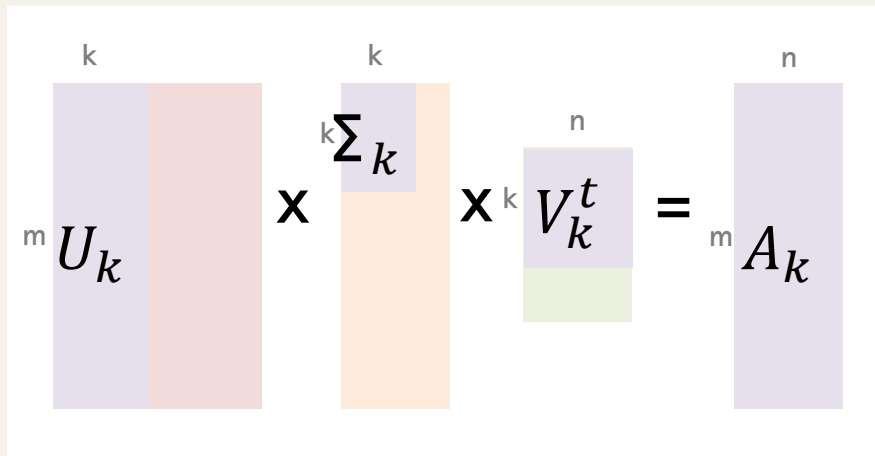
기존 토픽 모델링은 **단어의 빈도수**만을 고려하여  
문서의 주제를 찾는 방식



빈도수만으로는 단어의 의미를 정확히 고려 불가능하다는 단점 해결

## 잠재의미분석(Latent Semantic Analysis)

- $m \times n$  행렬  $A$ 는  $n$ 개의 문서가  $m$ 개의 단어로 표현된 입력 데이터
- $A = U \Sigma V^t$  로 특이값분해한 뒤 고유값의 개수를 임의로 설정해 부분복원 가능
- 고유값의 개수 = 찾고자 하는 토픽의 수를 반영한 모수



$U_k \Sigma_k V_k^t = A_k$  의 양변에

각각  $U_k^t$ ,  $V_k$ 를 곱하면

$X_1$ 과  $X_2$ 를 만들 수 있다.

$$U_k^t A_k = U_k^t U_k \Sigma_k V_k^t = \Sigma_k V_k^t = X_1$$

$$A_k V_k = U_k \Sigma_k V_k^t V_k = U_k \Sigma_k = X_2$$

## 잠재 의미 분석 예시

- 단어-문서 행렬을 특이값 분해하고 상위 2개의 특이값을 선택해 **부분복원**한다.

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} -0.40 & 0.30 & -0.19 \\ -0.40 & 0.30 & -0.19 \\ -0.40 & 0.30 & -0.19 \\ -0.28 & -0.30 & -0.63 \\ -0.25 & 0.15 & 0.38 \\ -0.38 & -0.30 & 0.31 \\ -0.38 & -0.30 & 0.31 \\ -0.25 & 0.15 & 0.38 \\ -0.13 & -0.45 & -0.06 \\ -0.13 & -0.45 & -0.06 \end{pmatrix} \begin{pmatrix} 3.14 & 0 & 0 \\ 0 & 2.00 & 0 \\ 0 & 0 & 1.46 \end{pmatrix} \begin{pmatrix} -0.47 & -0.78 & -0.42 \\ 0.30 & 0.30 & -0.90 \\ -0.83 & 0.55 & -0.09 \end{pmatrix}$$

부분복원행렬

$$U_2 \times \Sigma_2 \times V_2^t = \begin{pmatrix} 0.77 & 1.15 & -0.03 \\ 0.77 & 1.15 & -0.03 \\ 0.77 & 1.15 & -0.03 \\ 0.23 & 0.51 & 0.91 \\ 0.46 & 0.70 & 0.05 \\ 0.38 & 0.75 & 1.04 \\ 0.38 & 0.75 & 1.04 \\ 0.46 & 0.70 & 0.05 \\ -0.08 & 0.05 & 0.99 \\ -0.08 & 0.05 & 0.99 \end{pmatrix}$$