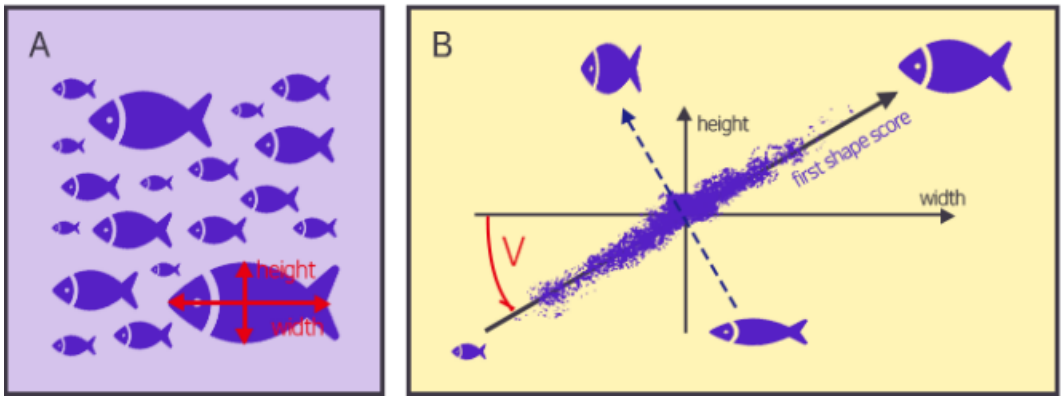
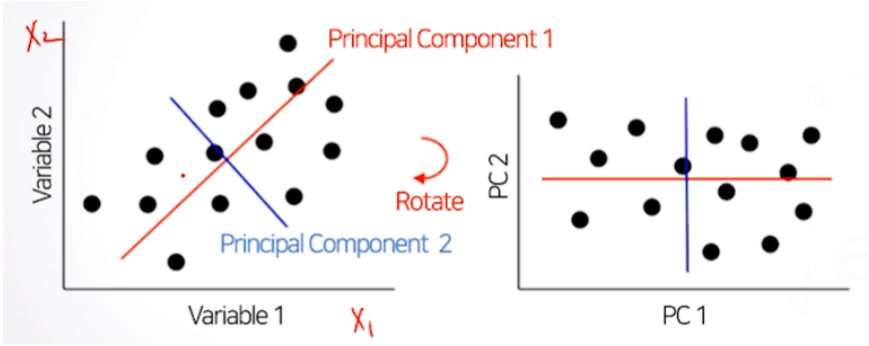
07.특성 추출 방법론

- 특성 추출(Feature extraction) : 가지고 있는 특성을 결합하여 더 유용한 특성을 생성
- 주요 특성 추출법
  - PCA(Principal component analysis)
    - 서로 연관되어 있는 변수들이 관찰되었을 때, 이 변수들이 전체적으로 가지고 있는 정보들을 최대한 확보하는 적은 수의 새로운 변수(주성분, PC)를 생성하는 방법



  - 주성분 분석의 목적
    - 자료에서 변동이 큰축을 탐색함
    - 변수들에 담긴 정보의 손실을 최소화하면서 차원을 축소함
    - 서로 상관이 없거나 독립적인 새로운 변수인 주성분을 통해 데이터의 해석을 용이하게함
  - 주성분 분석에 관한 기하학적 의미
    - 주성분 축은 원래 변수들의 좌표축이 직교 회전 변환 된 것으로 해석

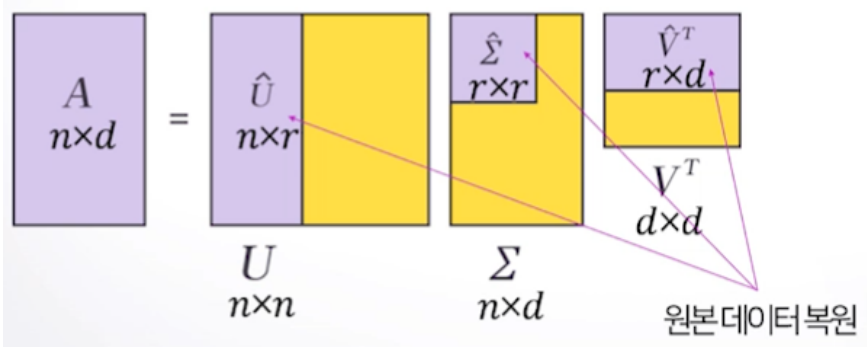


- 첫번째 주성분 축은 데이터의 변동이 가장 커지는 축
- 두번째 주성분 축은 첫번째 주성분 축과 직교하며 첫번째 주성분 축 다음으로 데이터의 변동이 큰 축을 나타냄
- 각 관찰지 별 주성분 점수는 대응하는 원 자료 값들의 주성분 좌표축에서의 좌표 값에 해당함
- 자료들의 공분산 행렬이 대각행렬이 되도록 회전한 것으로 해석가능

◦ SVD(Singular Value Decomposition)

▪ 특성값 분해 이론

- 특이값 분해 : 임의의 $n \times d$ 행렬 A는 $A = U \sum V^T$로 분해 가능
- U와 V는 직교행렬 : $U^T U = I_n \times'_n VV^T = I_d \times_d$
- U의 각 열을 A의 왼쪽 특성벡터, V의 각 열을 A의 오른쪽 특성벡터라고 함
- $\sum$는 $n \times d$의 대각행렬 : 대각원소를 A의 특성값이라고함

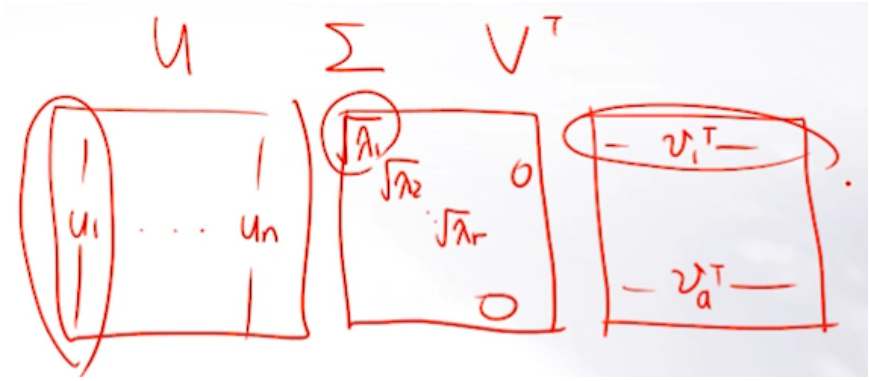


▪ 특이값 분해와 차원 축소

- U의 각 열을 $u_i, i = 1, ..., n$
- V^T의 각 행을 $v_i^t, i = 1, ..., d$
- Σ의 0이 아닌 대각원소를 $\lambda_i, i = 1, ..., r(\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r)$ 이라고 할 때,

$$A = U\Sigma V^T = \sqrt{\lambda_1} u_1 v_1^t + \sqrt{\lambda_2} u_2 v_2^t + \cdots + \sqrt{\lambda_m} u_m v_m^t + \cdots + \sqrt{\lambda_r} u_r v_r^t$$

정보가 많은 순서대로 m개만 이용하여 근사하는 경우 m계수 근사라고 함



- 주성분분석(PCA)와 특성값 분해의 관계
  - A의 오른쪽 특성벡터는 A의 공분산행렬의 고유벡터와 동일함.
  - 자료 행렬에 대한 특성값 분해로 주성분을 도출가능

- LDA(Linear discriminant analysis)
- NMF(Non-negative matrix factorization)