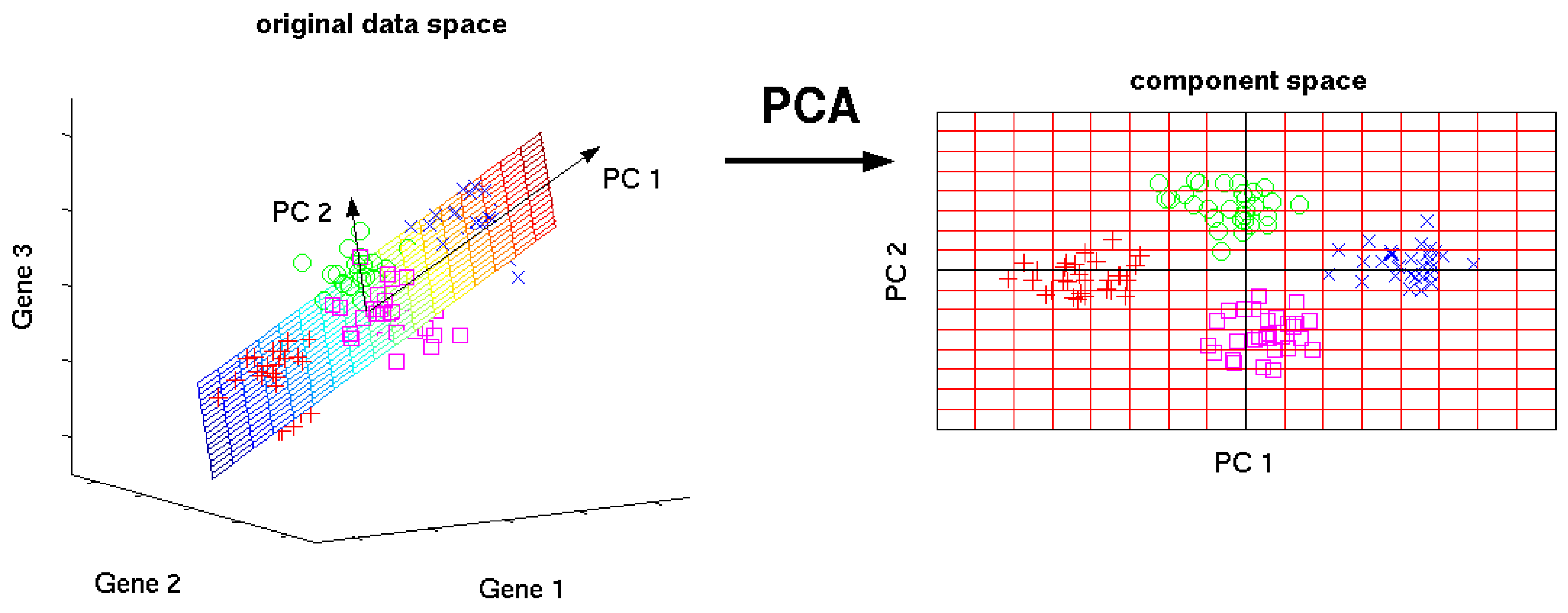
SVD(Singular Value Decomposition), PCA(Principle Components Analysis), and LSA(Latent Semantic Indexing)

Hyopil Shin(Seoul National University)
Computational Linguistics
#Semantics with Dense Vectors
Supplement Data

PCA(Principle Components Analysis)

- 데이터의 분산(variance)을 최대한 보존하면서 서로 직교하는 새 축을 찾아 고차원 공간의 표본들이 선형 연관성이 없는 저차원 공간으로 변환하는 기법



PCA(Principle Components Analysis)

- 공분산 행렬
  - 원 데이터의 분산을 최대화하는 새로운 기저를 찾는 목표를 달성하려면 우선 데이터 행렬 A 의 공분산 행렬부터 구해야 함.
  - 데이터가 각 변수별로 평균이 0으로 맞춰져 있을 때(centering 작업 이미 수행되어 있다고 가정) 공분산 행렬은 아래와 같이 구함

$$cov(A) = \frac{1}{n-1} A^T A \propto A^T A$$

  - 새로운 축을 찾기 위해 고유분해(eigen decomposition)을 수행해야 함

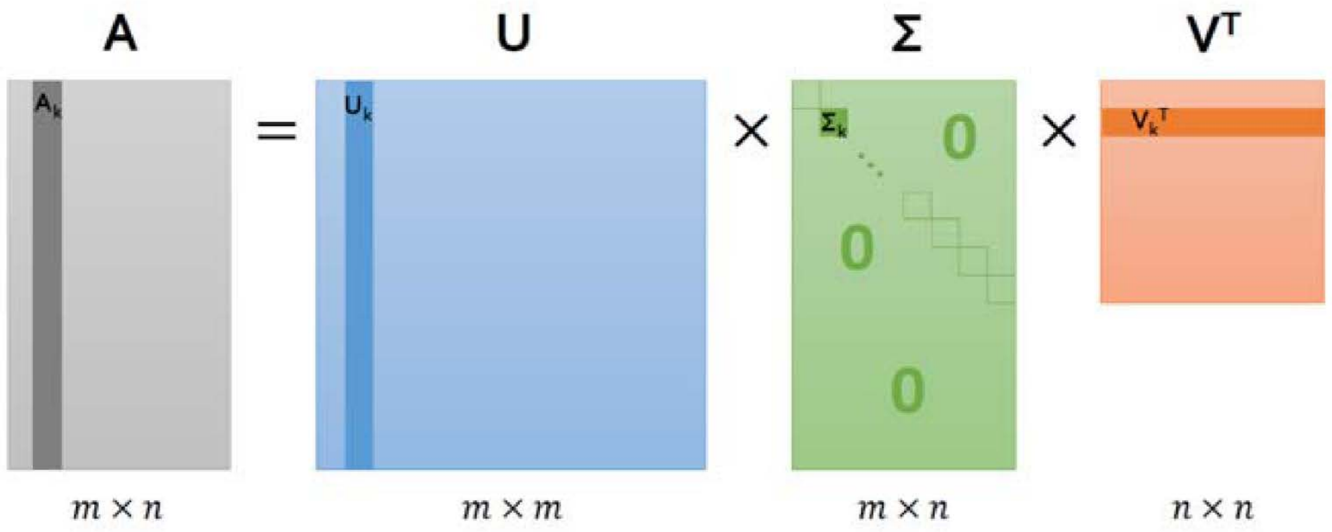
$$A^T A = V \Lambda V^T$$

  - Λ 의 대각성분은 데이터 행렬 A 의 각 변수에 해당하는 분산을 의미
  - 원 데이터의 분산을 최대화하는 새로운 기저를 찾는 것이 PCA 목표인 만큼 가장 큰 고유값 몇 개를 고르고, 그에 해당하는 고유벡터를 새로운 기저로 하여 원데이터를 사영(선형변환)
  - 예를 들어변수가 100개(100차원)인 데이터에 PCA를 적용한 후 가장 큰 고유값 두 개에 해당하는 고유벡터로 원 데이터를 사영시키면 원데이터의 분산을 최대한 보존하면서도 그 차원수를 100차원에서 2차원으로 줄일 수 있음

특이값 분해(singular value decomposition)

- m x n 크기의 데이터 행렬 A를 다음과 같이 분해

$$A = U\Sigma V^T$$



특이값 분해(singular value decomposition)

- 특이벡터(singualar vector) – 행렬 U와 V에 속한 열벡터
- 모든 특이벡터는 서로 직교

$$U = [\overrightarrow{u_1} \quad \overrightarrow{u_2} \quad \dots \quad \overrightarrow{u_m}]$$

$$V = [\overrightarrow{v_1} \quad \overrightarrow{v_2} \quad \dots \quad \overrightarrow{v_n}]$$

$$\overrightarrow{u_k} = \begin{bmatrix} u_{k1} \\ u_{k2} \\ \dots \\ u_{km} \end{bmatrix} \qquad \overrightarrow{v_k} = \begin{bmatrix} v_{k1} \\ v_{k2} \\ \dots \\ v_{kn} \end{bmatrix}$$

$$\overrightarrow{u_k}^T \overrightarrow{u_k} = 1, \quad U^T U = I$$

$$\overrightarrow{v_k}^T \overrightarrow{v_k} = 1, \quad V^T V = I$$

특이값 분해(singular value decomposition)

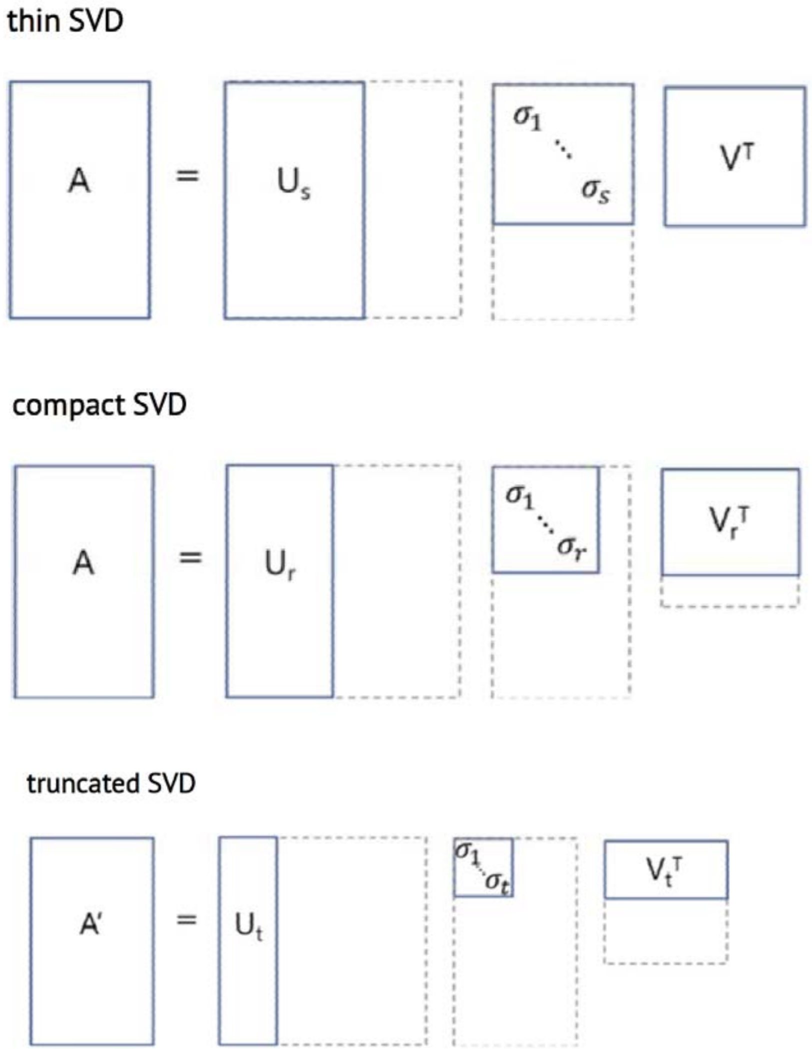
- 행렬 ∑의 특이값은 모두 0보다 크거나 같으며 내림차순으로 정렬
- 행렬 ∑의 k번째 대각원소 해당하는 ∑k는 행렬 A의 k번째 고유값에 제곱근을 취한 값과 같음
- 특이값 분해를 주성분 분석과 비교하기 위해 행렬 A를 제곱
  - 대각행렬의 거듭제곱은 대각원소들만 거듭제곱 해 준 결과와 동일
  - ∑의 제곱은 대각원소, 즉 행렬 A의 특이값들을 제곱해 준 것과 동일
  - 행렬 A의 특이값은 같은 행렬이 고유값에 제곱근을 취한 값과 동일하므로 ∑을 제곱한 행렬은 A^{T}A의 고유값으로 이루어진 Λ가 되고, 주성분 분석과 동일

$$\Sigma_k = \sqrt{\lambda_k}$$

$$\begin{aligned} A^T A &= (U\Sigma V^T)^T U\Sigma V^T \\ &= V\Sigma U^T U\Sigma V^T \\ &= V\Sigma^2 V^T \\ &= V\Lambda V^T \end{aligned}$$

특이값 분해의 종류

- thin SVD
  - Σ 행렬의 아랫부분(비대각 파트, 모두 0)과 U 에서 여기에 해당하는 부분을 모두 제거
  - 이렇게 U 와 Σ 를 줄여도 $U_s \Sigma_s V^T$ 로 A 를 복원가능
- compact SVD
  - Σ 행렬에서 비대각파트뿐 아니라 대각원소(특이값)가 0인 부분도 모두 제거
  - 여기에 대응하는 U 와 V 의 요소 또한 제거
  - 특이값이 양수인 부분만 골라냄
  - 이렇게 U 와 Σ, V 를 줄여도 $U_r \Sigma_r V^T$ 로 A 를 복원 가능
- truncated SVD
  - Σ 행렬의 대각원소(특이값) 가운데 상위 t 개만 고름
  - 이렇게 하면 행렬 A 를 복원 수 없게 되지만, 데이터 정보를 상당히 압축했음에도 행렬 A 를 근사할 수 있게 됨
  - 잠재의미분석은 바로 이 방법을 사용합니다.



특이값 분해 예시

- SVD 예

$$A = U\Sigma V^T$$

$$\begin{bmatrix} 2 & 3 \\ 1 & 4 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0.82 & -0.58 & 0 & 0 \\ 0.58 & 0.82 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 5.47 & 0 & 0 & 0 \\ 0 & 0.37 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0.40 & 0.91 \\ -0.91 & 0.40 \end{bmatrix}$$

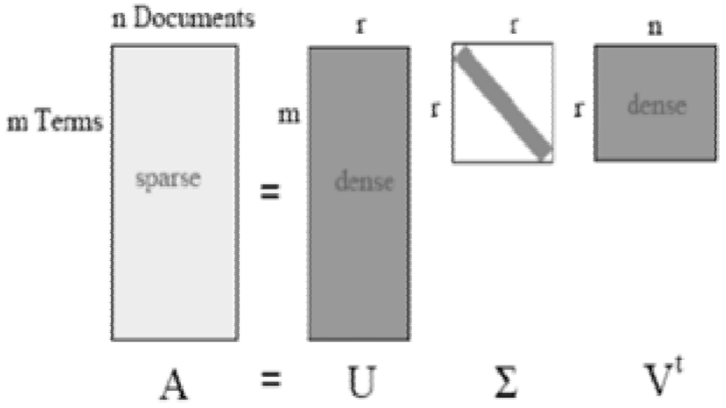
- Truncated SVD 예

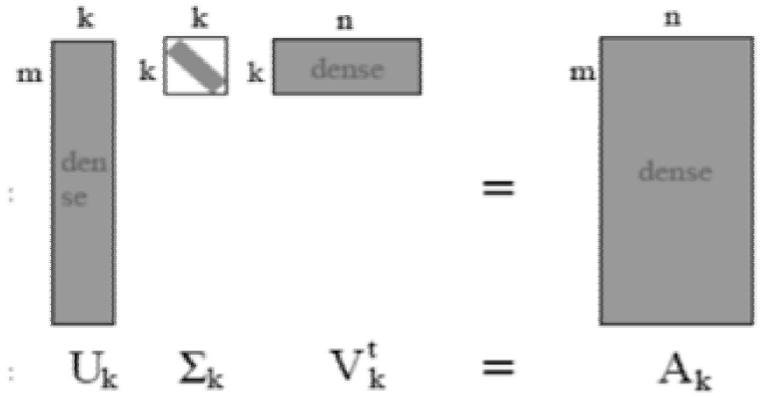
$$A' = U_1 \Sigma_1 V_1^T$$

$$\begin{bmatrix} 1.79 & 4.08 \\ 1.27 & 2.89 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0.82 \\ 0.58 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 5.47 \end{bmatrix} \begin{bmatrix} 0.40 & 0.91 \end{bmatrix}$$

Latent Semantic Analysis

- 단어-문서행렬(word-document matrix), 단어-문맥행렬(window-based co-occurrence matrix) 등 입력 데이터에 특이값 분해를 수행해 데이터의 차원수를 줄여 계산 효율성을 키우는 한편 행간에 숨어있는 의미를 이끌어 내기 위한 방법





Latent Semantic Analysis

- n 개의 문서를 *m* 개의 단어로 표현된 입력데이터 행렬 *A* 가 주어졌을 때
  - *A* 의 0보다 큰 고유값의 개수를 *r* 이라고 할 때, *r* 보다 작은 *k* 를 연구자가 임의로 설정하고 Σ_k 를 만듬
  - 이후 *U* 와 *V*행렬에서 여기에 대응하는 부분만 남겨 U_k 와 V_k 를 만들어 *A*와 비슷한 A_k 행렬을 구축
  - 이 식 양변에 U_k 의 전치행렬을 곱해준 것을 X_1 , V_k 를 곱해준 것을 X_2라 하면
  - X_1 의 경우 n개의 문서는 원래 단어수 *m*보다 훨씬 작은 *k* 개 변주로 표현됨
  - X_2 는 *m*개의 단어가 원래 문서 수 *n*보다 작은 *k*개 변수로 표현됨

$$A_k = U_k \Sigma_k V_k^T$$

$$U_k^T A_k = U_k^T U_k \Sigma_k V_k^T = I \Sigma_k V_k^T = \Sigma_k V_k^T = X_1$$

$$A_k V_k = U_k \Sigma_k V_k^T V_k = U_k \Sigma_k^T I = U_k V_k^T = X_2$$

Latent Semantic Analysis 예시

doc1 : 나,는,학교,에,가,ㄴ,다

doc2 : 학교,에,가,는,영희

doc3 : 나,는,영희,는,좋,다

-	doc1	doc2	doc3
나	1	0	0
는	1	1	2
학교	1	1	0
에	1	1	0
가	1	1	0
ㄴ	1	0	0
다	1	0	1
영희	0	1	1
좋	0	0	1

Latent Semantic Analysis 예시

$$A = U\Sigma V^T$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 2 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} -0.17 & 0.27 & -0.40 \\ -0.63 & -0.41 & -0.03 \\ -0.32 & 0.37 & 0.21 \\ -0.32 & 0.37 & 0.21 \\ -0.32 & 0.37 & 0.21 \\ -0.17 & 0.27 & -0.40 \\ -0.33 & -0.12 & -0.52 \\ -0.30 & -0.29 & 0.49 \\ -0.15 & -0.39 & -0.13 \end{bmatrix} \begin{bmatrix} 3.61 & 0 & 0 \\ 0 & 2.04 & 0 \\ 0 & 0 & 1.34 \end{bmatrix} \begin{bmatrix} -0.63 & -0.53 & -0.57 \\ 0.56 & 0.20 & -0.80 \\ -0.54 & 0.83 & -0.17 \end{bmatrix}$$

$$A' = U_2 \Sigma_2 V_2^T$$

$$\begin{bmatrix} 0.71 & 0.44 & -0.09 \\ 0.97 & 1.04 & 1.99 \\ 1.15 & 0.76 & 0.04 \\ 1.15 & 0.76 & 0.04 \\ 1.15 & 0.76 & 0.04 \\ 0.71 & 0.45 & -0.09 \\ 0.62 & 0.58 & 0.88 \\ 0.36 & 0.45 & 1.11 \\ -0.09 & 0.14 & 0.97 \end{bmatrix} = \begin{bmatrix} -0.17 & 0.27 \\ -0.63 & -0.41 \\ -0.32 & 0.37 \\ -0.32 & 0.37 \\ -0.32 & 0.37 \\ -0.17 & 0.27 \\ -0.33 & -0.12 \\ -0.30 & -0.29 \\ -0.15 & -0.39 \end{bmatrix} \begin{bmatrix} 3.61 & 0 \\ 0 & 2.04 \end{bmatrix} \begin{bmatrix} -0.63 & -0.53 & -0.57 \\ 0.56 & 0.20 & -0.80 \end{bmatrix}$$

Latent Semantic Analysis 예시

$$X_1 = \begin{bmatrix} -2.28 & -1.90 & -2.07 \\ 1.14 & 0.42 & -1.64 \end{bmatrix}$$

$$round(A') = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 2 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 2 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$X_2 = \begin{bmatrix} -0.63 & 0.56 \\ -2.30 & -0.84 \\ -1.16 & 0.76 \\ -1.16 & 0.76 \\ -1.16 & 0.76 \\ -0.63 & 0.56 \\ -1.20 & -0.24 \\ -1.10 & -0.60 \\ -0.57 & -0.80 \end{bmatrix}$$