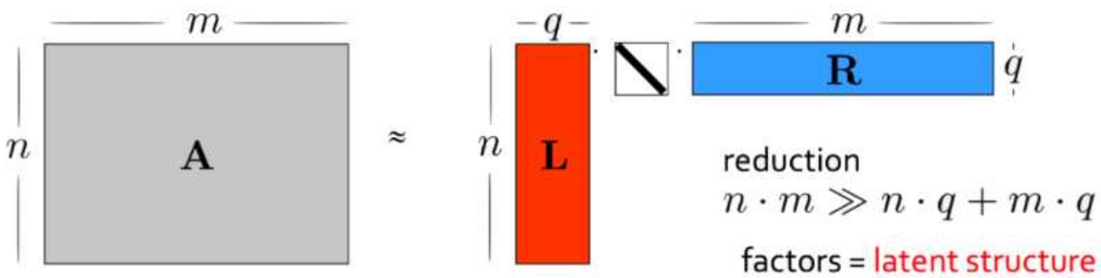
LSA & PLSA

- **LSA(Latent Semantic Analysis)**

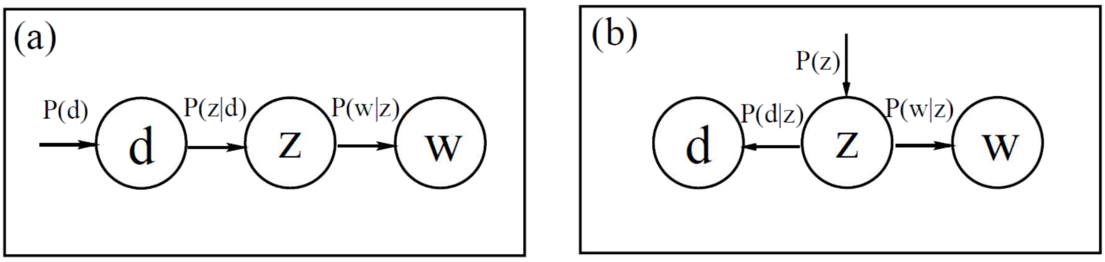
- 해당 matrix를 matrix factorization함으로써, 행과 열에 관련된 잠재 의미를 형성하고 이를 이용하여 행과 열 사이의 관계를 분석하는 기법
- 이때, 행과 열은 user, item이 될 수도, word와 document 등 다양한 것이 될 수도 있다.
- matrix factorization 방법으로 **SVD(singular Value Decomposition)**를 이용하여 본래 nxm의 matrix를 nxq, qxm의 matrix로 분해. 특이값은 각 q에 대한 중요도 가중치와 같은 역할을 할 것이다.



- LSA를 이용하여 modeling/similarity 계산 등을 할 수 있음
- LSA를 이용하면 noise, sparsity 감소와 같은 효과가 있기 때문에 모델의 성능 향상에 도움을 줄 수 있다. 하지만, 새로운 데이터가 추가되면 작업을 새로 시작하여야 한다는 단점이 있다.

- **PLSA(Probabilistic Latent Semantic Analysis)**

- 해당 matrix의 행과 열 사이에 "**확률적 잠재 구조**"를 찾는 기법
- d : document(n개), z : latent concept(k개), w : word(m개)



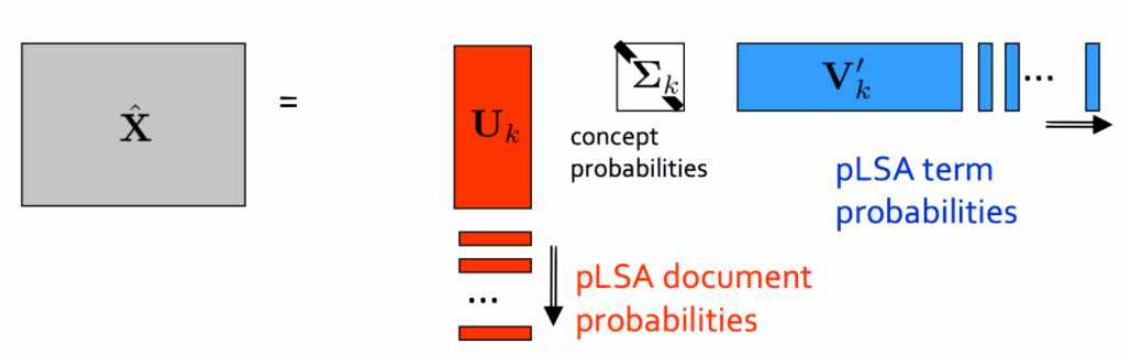
(a) 어떤 d가 주어졌을 때 z topic이 발생하고, 해당 z topic에서 w 단어가 발생할 확률을 계산

(b) 어떤 주제를 먼저 뽑은 뒤, 이 주제가 나타났을 때 해당 단어와 문서가 나타날 확률을 계산.

-> PLSA는 (b)방법으로 구성

▶ **추정 방법**

$$\widehat{X_{d,w}} = \widehat{P_{LSA}}(d,w) = \sum_z P(d|z)P(z)P(w|z)$$



- $P(d|z)$: z가 주어졌을 때 d가 나타날 확률, LSA의 U_k에 대응
- $P(w|z)$: z가 주어졌을 때 w가 나타날 확률, LSA의 V'_k에 대응
- $P(z)$: z가 나타날 확률(얼마나 중요한지를 나타내는 가중치), LSA의 $\sum_k$에 대응
- $\hat{X}$: d와 w가 같이 나타날 확률, 확률값이기 때문에 모두 0이상 1이하의 값, 행렬 전체 합 = 1
- 이 때, $P(d|z)$와 $P(w|z)$는 서로 독립이라고 가정.

▶ PLSA의 목적식

- likelihood function(weight로 부터 dataset이 나올 확률)을 최대화 하고자 한다.

$Max\,L$

$$L = \Pi_{i=1}^{m} \Pi_{j=1}^{n} p(w_i, d_j)^{n(w_i, d_j)}$$

$$L = \Pi_{i=1}^{m} \Pi_{j=1}^{n} [\sum_{l=1}^{k} p(d_j|z_l) p(z_l) p(w_i|z_l)]^{n(w_i, d_j)}$$

- $n(w_i, d_j)$: j번째 document에 i번째 word가 등장한 횟수

▶ 학습 : EM 알고리즘

- 파라미터를 번갈아 가면서 고정시키고 업데이트 하는 방식

<E-Step : Posterior probability of latent variables(concepts)>

$$P(z|d,w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z' \in Z} P(z')P(d|z')P(w|z')}$$

- 해당 document와 word가 발생하였을 때, concept z에 의해 설명되는 정도

<M-step : Parameter estimation based on “completed” statistics>

$$P(w|z) = \frac{\sum_{d \in D} n(d,w)P(z|d,w)}{\sum_{d \in D, w' \in W} n(d,w')P(z|d,w')}$$

- 해당 concept일 때, 해당 word가 발생할 확률

$$P(d|z) = \frac{\sum_{w \in W} n(d,w) P(z|d,w)}{\sum_{d' \in D, w \in W} n(d',w) P(z|d',w)}$$

-해당 concept일 때, 해당 document가 발생할 확률

$$P(z) = \frac{\sum_{d \in D, w \in W} n(d,w) P(z|d,w)}{\sum_{d \in D, w \in W} n(d,w)}$$

- 해당 concept이 발생할 확률

- 3개의 식 모두, 분모 분자에 존재하는 $\frac{1}{\sum_{d \in D, w \in W} n(d,w)}$가 약분된 형식