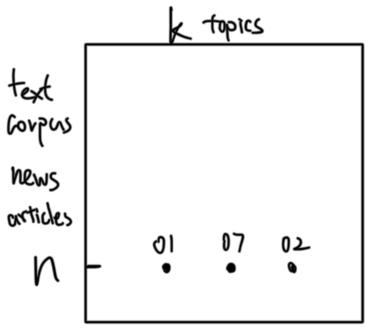
Lecture 7

Topic models



Given: documents $D_1, \cdots, D_n$

Unknown: topics $A_1^0, \cdots, A_k^0$

topic proportions $W_1^0, \cdots, W_n^0 \in \mathbb{R}^k_{\geq 0}$ summing to 1

Representation: bag-of-words * $W_\ell(j)$ is weight of topic j in document ℓ.

- $D_\ell(i) = \dfrac{\text{\# occurences of the word in doc. } \ell}{\text{\# word in doc. } \ell}$
- $D_\ell \in \mathbb{R}^d_{\geq 0}$ summing to 1

an article might have multiple topics.

- $A_j^0 \in \mathbb{R}^d_{\geq 0}$, $\mathbb{1}^T A_j^0 = 1$, $A_j^0(i) \propto$ significance of word i for topic j

Statistical model

- $M_\ell = \sum_{j=1}^{k} W_\ell^0(j) \cdot A_j^0$
- $D_\ell \leftarrow$ avg. of L-size sample from M_ℓ
- $\mathbb{E}[D_\ell] = M_\ell$

matrix view

$$\left.\begin{array}{l} - M = [M_1 \cdots M_n] \in \mathbb{R}^{d\times n} \\ - A = [A_1^0 \cdots A_k^0] \in \mathbb{R}^{d\times k} \\ - W = [W_1^0 \cdots W_n^0] \in \mathbb{R}^{k\times n} \end{array}\right\} M = A^0 W^0$$ *low-rank

- $D = [D_1, \cdots, D_n] \longrightarrow \mathbb{E}[D] = M$

Rotation problem

If $M = AW$, then for invertible $R \in \mathbb{R}^{k\times k}$, $M = (AR)(R^{-1}W)$ * SVD $Y = \sum \sigma_i u_i v_i^T$ $(U\Sigma)V^T$

rank-k non-negative matrix factorization (NMF) (for M)

Any pair (A, W) with $A \in \mathbb{R}^{d\times k}_{\geq 0}$, $W \in \mathbb{R}^{k\times n}_{\geq 0}$, and $M = AW$

colum stochastic matrix: All columns sum up to 1

row stochastic matrix: All rows sum up to 1

Claim: If M has rank-k NMF and M is column stochastic, then exists rank-k NMF for M with both factors column stochastic.

cols of M $\subseteq$ convex hull (cols of A)

rank-k NMF exists iff $\exists$ polytope $\in \mathbb{R}^d_{\geq 0}$ with k corners, contains all columns of M.

NP-Hard!!!

Separable NMF

word i is an anchor for topic j:

row i of A is multiple of e_j

Separable A:

every topic has anchor word, i.e., the rows of A contain $e_1, \cdots, e_k$ up to scaling.

Theorem

Suppose $M = AW$ for separable A and full-rank W. Then, given M, can compute

columns of A in poly time.
Here we assume M, A, W are (row) column stochastic, fixed scaling.

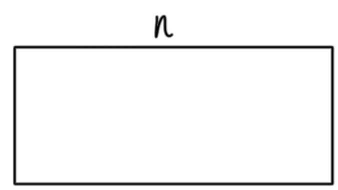
rescale rows of M to sum to 1,
→ may assume A, W to be row stochastic
→ separable A contains $e_1, \cdots, e_k$ as rows (without scaling)
→ $\{\text{rows of } M\} \supseteq \{\text{rows of } W\}$

$$\begin{bmatrix} W \\ \end{bmatrix} = \begin{bmatrix} I \\ A \end{bmatrix} \begin{bmatrix} W \end{bmatrix}$$

Corners of convex hull of rows of M must be rows of W
claim: $\{\text{corners}\} = \{\text{rows of } W\}$
W has rank $k \longrightarrow$ rows of W linearly independent
Convex-hull(rows of M) = convex-hull(rows of W) * Simply because rows of W are lin. indep.

$M = AW$, given M, W, can find A as unique to this linear system.

Generate document i:
- Draw L-size sample M_i
- Set $D_i(j) = \frac{1}{L}$ (count of word j in doc. i)



Assumption:
- A, W full rank (rank-k)
- $d \ll n$
- $k \ll d$
- A is separable
- $L \ll d$

Strategy:
- Estimate $\frac{1}{n} M M^T \in \mathbb{R}^{d \times d}$

$$M M^T = \underbrace{A}_{\text{separable}} \underbrace{(W W^T A^T)}_{\text{full rank}} \quad \text{separable NMF!}$$

- Apply (robust version) of algorithm for separable NMF to this estimate

Gram matrix
- $G = \frac{1}{n} D D^T = \frac{1}{n} \sum_{k=1}^{n} D_k D_k^T$ (dxd) * Since $n \gg d$, can expect G close to $\mathbb{E}[G]$
- $Q = \mathbb{E}[G] = \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}[D_k D_k^T]$
- $D_{k,i,l}$ = draw l from M_k is word i

$$\rightarrow D_k(i) = \frac{1}{L} \sum_{l=1}^{L} D_{k,i,l}$$

$$\mathbb{E}[D_k(i) D_k(i')] = \frac{1}{L^2} \sum_{l,l'} \mathbb{E}[D_{k,i,l} \cdot D_{k,i',l'}]$$

$$= \frac{1}{L^2} \sum_{l} \underbrace{\mathbb{E}[D_{k,i,l} \cdot D_{k,i',l}]}_{= \begin{cases} 0, & \text{if } i \neq i' \\ M_k^\circ(i), & \text{if } i = i' \end{cases}} + \frac{1}{L^2} \sum_{l \neq l'} \mathbb{E}[\underbrace{D_{k,i,l}}_{M_k(i)=0} \cdot \underbrace{D_{k,i',l'}}_{M_k(i')=0}]$$

$$\rightarrow \mathbb{E}[D_k(i) D_k(i')] = \frac{1}{L} \cdot M_k^\circ(i) \cdot [\![i = i']\!] + \left(1 - \frac{1}{L}\right) \cdot M_k^\circ(i) \cdot M_k^\circ(i')$$

$$\rightarrow \mathbb{E}[D_k D_k^T] = \frac{1}{L} \cdot \mathrm{diag}(M_k^\circ) + \frac{L-1}{L} \cdot M_k^\circ (M_k^\circ)^T$$

* estimator for $\frac{1}{n} M^\circ M^{\circ T} = A^\circ \left(\frac{1}{n} W^\circ (W^\circ)^T (A^\circ)^T\right)$

full rank NN

- $\underline{\underline{Y}}_k = \frac{L}{L-1}\left(\underline{D}_k \underline{D}_k^T - \frac{1}{L}\cdot \mathrm{diag}(\underline{D}_k)\right)$

  $\rightarrow \mathbb{E}[\underline{\underline{Y}}_k] = M_k M_k^T$

  $\rightarrow \underline{\underline{Y}}_k = M_k M_k^T + \underline{\underline{Z}}_k$, $\mathbb{E}[\underline{\underline{Z}}_k] = 0$, $\underline{\underline{Z}}_1, \cdots, \underline{\underline{Z}}_k$ independent * not necessarily identical

  $\rightarrow$ for $n \gg d$, can use concentration bounds to show that $\frac{1}{n}\underline{\underline{Z}}_k$ is small w.h.p.

  $\rightarrow \left\| \frac{1}{n} \Sigma \underline{\underline{Y}}_k - \frac{1}{n} M^0 M^{0T} \right\| = \left\| \frac{1}{n} \Sigma \underline{\underline{Z}}_k \right\|$ is small w.h.p.