What if there is noise?

Abstract Topic Model

① unknown topic matrix A, distribution μ on simplex in $\mathbb{R}^r$

② For each document i

ⓐ sample W_i from μ

ⓑ Generate L words i.i.d. from AW_i

Can we recover A? Notice that

$$\mathbb{E}[\tilde{M} \mid W] = M = AW$$

empirical term-by-document

but $\tilde{M}$ and M are far apart. why? sparsity

This is a rich model, containing

supp on corners, dirichlet, lognormal, hierarchical

Pure, LDA, CTM, Pachinko

as special cases

Thm [Arora, Ge, Moitra] there is a polynomial time algorithm for learning A when it is separable, under nondegeneracy conditions

It's not obvious how to use NMF

<u>def</u>: The Gram matrix G is

$$G_{j,j'} \triangleq \mathbb{P}\left[\begin{array}{l}\text{in random document,} \\ 1^{st} \text{ word} = j \\ 2^{nd} \text{ word} = j'\end{array}\right]$$

<u>def</u>: The topic co-occurrence R is

$$R_{i,i'} \triangleq \mathbb{P}\left[\begin{array}{l}\text{in random document,} \\ 1^{st} \text{ word was sampled from topic } i \\ 2^{nd} \text{ word} \quad \text{"} \quad \text{topic } i'\end{array}\right]$$

By law of total expectation

$$G_{j,j'} = \sum_{i,i'} R_{i,i'} A_{ji} A_{j'i'}$$

what conditional independence is implicit here?

$$\Rightarrow G = ARA^T$$

Can we apply separable NMF to G?

Yes, but much better probabilistic approach that uses non-anchors too!

Let w_1 = 1^{st} word of random document, t_1 the topic it came from

<u>Fact</u>: j is an anchor word iff

$$\mathbb{P}[t_1 \mid w_1 = j] \text{ is a point mass}$$

Furthermore we can expand

$$\mathbb{P}[w_1 = j \mid w_2 = j'] =$$

$$\sum_{i'} \mathbb{P}[w_1 = j \mid w_2 = j', t_2 = i'] \, \mathbb{P}[t_2 = i' \mid w_2 = j']$$

By conditional independence

(*) $$= \sum_{i'} \mathbb{P}[w_1 = j \mid t_2 = i'] \, \mathbb{P}[t_2 = i' \mid w_2 = j']$$

<u>Claim</u>: $\mathbb{P}[w_1 = j \mid t_2 = i'] = \mathbb{P}[w_1 = j \mid w_2 = \pi(i')]$

This implies

(1) $$(*) = \sum_{i'} \mathbb{P}[w_1 = j \mid w_2 = \pi(i')] \, \underbrace{\mathbb{P}[t_2 = i' \mid w_2 = j']}_{\text{unknowns}}$$

This linear system has unique soln if R is full rank

<u>Note</u>: It uses all words, not just anchor words

Finally, by Bayes rule

(2) $$\mathbb{P}[w_2 = j' \mid t_2 = i'] = \frac{\mathbb{P}[t_2 = i' \mid w_2 = j'] \, \mathbb{P}[w_2 = j']}{\mathbb{P}[t_2 = i']}$$

and moreover

$$\mathbb{P}[t_2 = i'] = \sum_j \mathbb{P}[t_2 = i' \mid w_2 = j]\, \mathbb{P}[w_2 = j]$$

Thus our algorithm is

<u>Anchor-Bayes</u>

Compute Gram matrix G

Compute the anchorwords (via separable NMF)

Solve (1) for $\mathbb{P}[t_2 = i' \mid w_2 = j']$

Solve (2) to compute $\mathbb{P}[w_2 = j' \mid t_2 = i']$

$$= A_{j'i'}$$

<u>Proof of Claim</u> Let's expand

$$\mathbb{P}[w_1 = j \mid w_2 = \pi(i')] =$$

$$\sum_{i''} \mathbb{P}[w_1 = j \mid w_2 = \pi(i'), t_2 = i''] \underbrace{\mathbb{P}[t_2 = i'' \mid w_2 = \pi(i')]}_{= \begin{cases} 1 & \text{if } i'' = i' \\ 0 & \text{else} \end{cases}}$$

does this use uniqueness of anchor words? no

$$= \mathbb{P}[w_1 = j \mid w_2 = \pi(i'), t_2 = i']$$

by conditional independence

$$= \mathbb{P}[w_1 = j \mid t_2 = i'] \quad \square$$

Are natural topic models separable?

UCI Dataset
300k NYT articles $\rightsquigarrow$ MALLET $\rightsquigarrow$ $\hat{A}$

<u>Findings</u>: with $r = 200$, about 0.9 fraction of topics had a near anchor word
(posterior $\gtrsim 0.9$)

How do algorithms based on separability perform?

$\hat{A}$ $\rightsquigarrow$ synthetic documents $\rightsquigarrow$ MALLET
$\rightsquigarrow$ Anchor Bayes

<u>Findings</u>: More accurate, and a hundred times faster

What do we find on real data?

Show figures