Topic Models

Goal: uncover thematic structure
Approach: Bayesian, probabilistic model.

- Probabilistic Model
  - Def. (informal) A prob. model is a class of distributions
    \( D(\theta) \) (\( \theta \in \Theta \)), \( \Theta \) is called the parameter.
    \( D(\theta) \) is a distribution.

- Def. Parameter estimation: Given samples from \( D(\theta) \),
  estimate \( \theta \).

- Example: Gaussian distribution \( \mathcal{N}(\mu, \sigma^2) \)

  - Parameter estimation: easy \((\text{sample mean})\)

    Note: cannot get exact \( \mu \) only hope for \( \frac{1}{\sqrt{\text{#samples}}} \)

- Topic models
  - Plan: Define a topic model \((\text{hard!})\)
    \((\text{distribution of documents, parameters specify topics})\)

    Estimate the parameters.

  - Simplifying assumptions. \((\text{for the model})\)
    \- bag of words: ordering does not matter.

    Consequence: \( \text{doc} \Leftrightarrow \text{frequency of words} \)
- Top topic structure
  - \text{topic} := \text{distribution over words}
  - \text{document} = (\text{linear}) \text{ mixture of topics.}

- "Algorithm" for generating documents
  - Given: \( A \in \mathbb{R}^{n \times k} \) "word-topic" matrix

- Pick \( w \sim D_{\text{topic}} \)
  - \( w \in \mathbb{R}^k, \sum_{i=1}^k w_i = 1 \)
- Pick length \( N \)
- For \( i = 1 \) to \( N \)
  - Independent \( \{ \text{pick topic } t_i \sim w \} \)
  - Bag-of-words \( \{ \text{pick word } z_i \sim A_{:, t_i} \} \)

- Matrix representation

- Different choices for \( D_{\text{topic}} \)
  - Pure documents:
    - \( w = e_i \) with probability \( \alpha_i \)
  - Latent Dirichlet Allocation (Blei, Ng, Jordan 03)
    - \( \text{Pr}[w] \propto \prod_{i=1}^K \alpha_i \alpha_i^{w_i} \)
    - \( \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K) \)
  - Property: \( w \) is sparse (when \( \alpha_i \ll 1 \))
    - \( \alpha = \sum_i \alpha_i \) is roughly \# large entries
  - Correlated Topic Model (Blei Lafferty)
- Using NMF for topic models
  
  - Separability
    - Recall: \( i = 1, \ldots, k \), there is a row \( r_i \) such that \( a_{ri} > 0 \), \( \forall j \neq i \), \( a_{rj} = 0 \).
    - For every topic, there is a word ('anchor-word') that only appears in this topic. (also: has reasonable prob.)

  \[
  M^T = W A^T
  \]

  For every topic, there is a pure document.

  (Q: why do we use "anchor-words" instead of "pure doc")

- Sampling noise
  
  Example: \( \left( \frac{1}{10000}, \ldots, \frac{1}{10000} \right) \), 100 samples \( \Rightarrow \) vector with \( \leq 100 \) nonzero entries

  \( l_1 \) distance \( \approx 2 \)!

  too large to be called "perturbation"

- Idea: reducing noise with more documents

- Word-word correlation matrix

  \( q_{ij} = \Pr[z_i = i, z_j = j] \)

  Claim: \( Q = A A^T \), \( R = E[w w^T] \), \( w \sim \text{Topic} \)

  Proof: \( z_i, z_j \) independent conditioned on \( w \)

  \[
  Q = E[z_i z_j^T] = E[E[z_i z_j^T | w] \]  
  \[
  = E[E[z_i | w] E[z_j^T | w]] 
  \]

  \( w = a_1 \wedge \ldots \wedge a_k \wedge \Delta_1^T, \ldots \wedge \rho \Delta_k^T \)
\[
\begin{align*}
\mathbf{E}[\mathbf{w}^\top A \mathbf{w} A^\top] &= \mathbf{A} \mathbf{R} \mathbf{A}^\top \\
\end{align*}
\]

# words in vocabulary fixed, # doc \(\to\) infinity

\(\Rightarrow\) can estimate \(Q\)

(recall: why not "pure doc"? hard to reduce noise)

- **High level Alg.**
  
  1. estimate \(Q\) matrix
  2. apply separable NMF to get \(Q = \tilde{A} \tilde{W}^*\)
  3. fix the scaling to find \(A\)

\[
\bar{Q}_{ij} = \frac{Q_{ij}}{\|Q_{ii}\|} = \frac{\Pr[Z_2 = j \mid Z_1 = i]}{\sum_{l=1}^{K} \Pr[Z_2 = j \mid t_1 = l] \Pr[t_1 = l \mid Z_1 = i]}
\]

\[
\begin{bmatrix}
\bar{Q} \\
\tilde{A} \\
\tilde{W}^* \\
\end{bmatrix}
\]


\[
\begin{align*}
\bar{Q}_{i,j} &= \frac{Q_{i,j}}{\|Q_{ii}\|} = \Pr[Z_2 = j \mid Z_1 = i] \\
\Pr[Z_2 = j \mid Z_1 = i] &= \sum_{l=1}^{K} \Pr[Z_2 = j \mid t_1 = l] \Pr[t_1 = l \mid Z_1 = i] \\
\end{align*}
\]

\(\bar{R}_l\): anchor word for topic \(l\), then

\[
\Pr[Z_2 = j \mid Z_1 = \bar{R}_l] = \Pr[Z_2 = j \mid t_1 = l]
\]

NMF:

\[
\bar{A}_{i,l} = \Pr[t_1 = l \mid Z_1 = i]
\]

want:

\[
A_{i,l} = \Pr[Z_2 = i \mid t_1 = l]
\]