Recap: Using NMF for topic modeling
- estimate $Q: Q_{i,j} = \Pr[z_i = i, z_j = j]
- NMF: $\bar{Q} = \bar{A} \bar{W}^*$ ($\bar{W}^* = \bar{R} \bar{A}^T$)
- Fix normalization in $\bar{A}$.

Problem 1. How to estimate $Q$?

Lemma: Let $u$ be the word count vector for a length $N$ document, then $E[u u^T - \text{diag}(u)] = N(N-1)Q$

Proof: $u = \sum_{i=1}^{N} z_i \iff i$th word

we know $E[z_i | w] = Aw \iff$ mixture of topics

$E[z_i z_j^T] = \bar{A} \bar{R}^T = Q$ (See last lecture)

but $E[z_i z_j^T] = \bar{E}[z_i z_j^T] \quad$ (Symmetry)

So $E[u u^T] = E\left[\sum_{i=1}^{N} \left(\sum_{j=1}^{N} z_i z_j^T\right)\right]$

$= \sum_{i \neq j} E[z_i z_j^T] + \sum_{i=1}^{N} E[z_i z_i^T]$

$= (N-1)Q + NQ = N(N+1)Q$
\[ \mathbb{E} [\mathbf{u} \mathbf{u}^T - \text{diag}(\mathbf{u})] = N(N-1) \mathbf{Q} \]

Why is this better? Use more samples, get better estimate of $\mathbf{Q}$.

**Problem 2:** Algorithm for NMF is slow.

Fast algorithm: repeatedly look for the furthest point.

Given: $V_1, V_2, \ldots, V_n$

Find: vertices $u_1, u_2, \ldots, u_k$

$u_1$ = vector with largest norm

$u_2$ = vector farthest away from $u_1$

for $i = 3$ to $n$

$u_i$ = vector farthest to $\text{aff}(u_1, u_2, \ldots, u_{i-1})$

\[ \text{aff}(u_1, u_2, \ldots, u_k) = \left\{ u \mid u = \sum_{i=1}^{k} c_i u_i, \sum_{i=1}^{k} c_i = 1 \right\} \]

Recall \[ \text{conv}(u_1, u_2, \ldots, u_k) = \text{cone}(u_1, \ldots, u_k) \cap \text{aff}(u_1, \ldots, u_k) \]

**Inference Problem**

Given $A$, document $\mathbf{u}$, how to find $\mathbf{w}$?

- Observations:
  - Cannot hope to get exact $\mathbf{w}$
  - In practice, try to approximate posterior dist.
  
  \[ \text{Prior} \quad P(\mathbf{w}) \propto \prod_{i=1}^{k} w_i \]

  \[ \mathbf{w} \mid A, \mathbf{u} \]

  - Can also try to find a matrix $\mathbf{B}$ such that
  \[ \mathbf{B} \mathbf{A} = \mathbf{I} \]
and use $\hat{\omega} = \frac{B\nu}{|\nu|}$

$$E[\hat{\omega}] = \frac{B E[\nu]}{N} = \frac{B \cdot N \cdot A\omega}{N} = B A \omega = \omega$$

Problem: how to control variance?

see http://128.84.21.199/abs/1605.08491