Robust Spectral Inference for Joint Stochastic Matrix Factorization

Kun Dong
Cornell University

October 20, 2016
Introduction

Topic Modeling

• Idea: Represent documents as combination of topics.

• Advantages:
  • Low-dimensional representation of documents
  • Uncover hidden structure from large collections

• Applications:
  • Summarizing documents with the topics
  • Clustering documents by similarity in topics
Co-occurrence Matrix

- The relationships between words can be more revealing than the words themselves.

\[ C \approx BAB^T \]

- \( C \in \mathbb{R}^{n \times n} \) - Word-Word Matrix. \( C_{ij} = p(X_1 = i, X_2 = j) \)
- \( A \in \mathbb{R}^{k \times k} \) - Topic-Topic Matrix. \( A_{k\ell} = p(Z_1 = k, Z_2 = \ell) \)
- \( B \in \mathbb{R}^{n \times k} \) - Word-Topic Matrix. \( B_{ik} = p(X = i | Z = k) \)
What We Observe

\[ B \quad A \quad B^T \quad \hat{C} \]
• **Separability:** The word-topic matrix $B$ is $p$-separable if for each topic $k$ there is some word $i$ such that $A_{i,k} \geq p$ and $A_{i,\ell} = 0$ for $\ell \neq k$

• Every topic $k$ has an anchor word $i$ exclusive to it.

• Documents containing anchor word $i$ must contain topic $k$. 

![Image of a word-topic matrix](image-url)
Anchor Word Algorithm

- Under this assumption, Arora et al. (2013) showed Anchor Word Algorithm computes this decomposition in polynomial time.

- Use QR with row-pivoting after random projection on $C$. Choose the points that are farthest away from each other.

- However, it fails to produce doubly nonnegative topic-topic matrix.

- It tends to choose rare words as anchors and generate less meaningful topics.
- For $m$-th document with $n_m$ words, we view it as $n_m(n_m - 1)$ pairs.
- Generate a distribution $A$ over pairs of topics with parameter $\alpha$.
- Sample two topics $(z_1, z_2) \sim A$.
- Sample actual word-pair $(x_1, x_2) \sim (B_{z_1}, B_{z_2})$. 
• Let $f(\alpha)$ be a distribution of topic-distributions.
• Documents are $M$ i.i.d. samples $\{W_1, \cdots, W_m\} \sim f(\alpha)$.
• Let the posterior topic-topic matrix $A^*_M = \frac{1}{M} \sum_{m=1}^{M} W_m W_m^T$ and the expectation $A^* = \mathbb{E}[W_m W_m^T]$. $A^*_M \to A^*$ as $M \to \infty$.
• Let posterior word-word matrix $C^*_m = BW_m W_m^T B^T$ and $C^* = \frac{1}{M} \sum_{m=1}^{M} C^*_m$.
• Let $C$ be the noisy observation for all samples.

$$C \to \mathbb{E}[C] = C^* = BA^*_M B^T \to BA^* B$$

• $A^*_M, A^* \in \mathcal{DNN}_K$ and $C^* \in \mathcal{DNN}_N$. 
Generating Co-occurrence C

- Let $H_m$ be the vector of word counts for $m$-th document and $W_m$ be the latent topic distribution.

- Let $p_m = BW_m$, and we assume $H_m \sim Multi(n_m, p_m)$.

- $\mathbb{E}[H_m] = n_m p_m = n_m BW_m$ and $\text{Cov}(H_m) = n_m (\text{diag}(p_m) - p_m p_m^T)$.

- Let co-occurrence $C_m = H_m H_m^T - \text{diag}(H_m) \over n_m(n_m-1)$.

- $\mathbb{E}[C_m|W_m] = C^*_m$ so $\mathbb{E}[C|W] = C^*$. 
Rectifying Co-occurrence $C$

- In reality $C$ could still mismatch $C^*$ because of model assumption violation and limited data.
- We can rectify $C$ into low-rank, doubly non-negative and joint-stochastic by Alternating Projection (Dykstra’s Algorithm).

\[
\begin{align*}
\Pi_{PSD_{NK}}(C) &= U\Lambda^+_K U^T \\
\Pi_{N\otimes R_N}(C) &= C + \frac{1 - \sum_{i,j} C_{ij}}{N^2} 11^T \\
\Pi_{N\cdot N_N}(C) &= \max\{C, 0\}
\end{align*}
\]
Finding Anchor Words

- Use a column-pivoting QR algorithm to greedily find topics farthest away from each other.
- Exploit sparsity and avoid using random projection.

Figure 1: 2D visualizations show the low-quality convex hull found by Anchor Words [6] (left) and a better convex hull (middle) found by discovering anchor words on a rectified space (right).
Recovering Word-Topic Matrix $B$

- If we row-normalize $C$ to get $\overline{C}$, $\overline{C}_{ij} = p(w_2 = j|w_1 = i)$.
- Under separability assumption,

  $$\overline{C}_{s_k,j} = \sum_{k'} p(z_1 = k'|w_1 = s_k)p(w_2 = j|z_1 = k') = p(w_2 = j|z_1 = k)$$

- The row-space of $\overline{C}$ lies in the convex hull of $\overline{C}_{s_k}$ rows.

  $$\overline{C}_{ij} = \sum_k p(z_1 = k|w_1 = i)p(w_2 = j|z_1 = k) = \sum_k Q_{ik} \overline{C}_{s_k,j}$$

- Find $Q_{ik}$ through NNLS and infer $B_{ik}$ with Bayes’ rule.
### Example of Recovered Topics

Table 3: Each line is a topic from NIPS ($K = 5$). Previous work simply repeats the most frequent words in the corpus five times.

<table>
<thead>
<tr>
<th>Arora et al. 2013 (Baseline)</th>
</tr>
</thead>
<tbody>
<tr>
<td>neuron layer hidden recognition signal cell noise</td>
</tr>
<tr>
<td>neuron layer hidden cell signal representation noise</td>
</tr>
<tr>
<td>neuron layer cell hidden signal noise dynamic</td>
</tr>
<tr>
<td>neuron layer cell hidden control signal noise</td>
</tr>
<tr>
<td>neuron layer hidden cell signal recognition noise</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>This paper (AP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>neuron circuit cell synaptic signal layer activity</td>
</tr>
<tr>
<td>control action dynamic optimal policy controller reinforcement</td>
</tr>
<tr>
<td>recognition layer hidden word speech image net</td>
</tr>
<tr>
<td>cell field visual direction image motion object orientation</td>
</tr>
<tr>
<td>gaussian noise hidden approximation matrix bound examples</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Probabilistic LDA (Gibbs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>neuron cell visual signal response field activity</td>
</tr>
<tr>
<td>control action policy optimal reinforcement dynamic robot</td>
</tr>
<tr>
<td>recognition image object feature word speech features</td>
</tr>
<tr>
<td>hidden net layer dynamic neuron recurrent noise</td>
</tr>
<tr>
<td>gaussian approximation matrix bound component variables</td>
</tr>
</tbody>
</table>
Recovering Topic-Topic Matrix $A$

\[ p(w_1|z_1) \quad p(z_1, z_2|m) \quad p(w_2|z_2) \quad p(w_1, w_2|m) \]

- Diagonal submatrix $D$ in $B$.

\[ \Rightarrow \hat{C} = \begin{bmatrix} D \\ B' \end{bmatrix} A \begin{bmatrix} D^T & B'^T \end{bmatrix} = \begin{bmatrix} DAD^T & DAB'^T \\ B'AD^T & B'AB'^T \end{bmatrix}, \quad A = D^{-1}C_{SS}D^{-1} \]
• This algorithm can handle noisy co-occurrence by rectification.
• It produces quality anchor words and topics, even when sample size is small.
• Preserve the structure of the decomposition under our assumption.

Thank you!