Sparse Latent Semantic Analysis

Xi Chen\textsuperscript{1}, Yanjun Qi\textsuperscript{2}, Bing Bai\textsuperscript{2}, Qihang Lin\textsuperscript{1}, Jaime G. Carbonell\textsuperscript{1}

1. Machine Learning Department, Carnegie Mellon University
2. Machine Learning Department, NEC Lab America

The work is done during the internship at NEC Lab America
Vector Space Model:

Document: \( \mathbf{x} = [w_1, \ldots, w_M] \in \mathbb{R}^M \)  \( M \): vocabulary size

\( w_i \): normalized weight (tf-idf) of the \( i \)-th word

\( N \) Documents: \( \mathbf{X} = [\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^N] \in \mathbb{R}^{N \times M} \): Document-Word matrix

Latent Semantic Analysis:

\( D \) latent topics (dimensionality of the latent space)

LSA applies SVD to construct a rank-\( D \) approximation:

\[
\mathbf{X} \approx \mathbf{U}_{N \times D} \mathbf{S}_{D \times D} (\mathbf{V}_{M \times D})^T, \quad \mathbf{U}^T \mathbf{U} = \mathbf{I}, \mathbf{V}^T \mathbf{V} = \mathbf{I}
\]

Projection Matrix: \( \mathbf{A} = \mathbf{S}^{-1} \mathbf{V}^T \in \mathbb{R}^{D \times M} \)

Dimension reduction for a new document \( q \): \( q \in \mathbb{R}^M \Rightarrow \hat{q} = \mathbf{A} q \in \mathbb{R}^D \)
Optimization Formulation for LSA

- **Latent Semantic Analysis**

\[ X \approx U_{N \times D} S_{D \times D} (V_{M \times D})^T, \quad U^T U = I, V^T V = I \]

- **Relaxed Optimization Formulation:**

\[
\min_{U,A} \quad \frac{1}{2} \| X - U A \|_F^2 \\
\text{subject to:} \quad U^T U = I
\]

[K. Yu et al. 05]

- **Sparse Latent Semantic Analysis:**

  Add *sparsity* constraint on the project matrix \( A \):

\[
\min_{U,A} \quad \frac{1}{2} \| X - U A \|_F^2 + \lambda \| A \|_1 \quad \Rightarrow \quad \| A \|_1 = \sum_{d=1}^D \sum_{j=1}^M |a_{d,j}| \quad \ell_1\text{-regularization}
\]

\[ U^T U = I \]
Sparse LSA

\[ \min_{U,A} \quad \frac{1}{2} \| X - UA \|_F^2 + \lambda \| A \|_1 \]
subject to: \[ U^T U = I \]

\[ \begin{array}{c|c|c|c}
N & M & \approx & D \\
\end{array} \quad \begin{array}{c|c|c|c}
X & U & A & D - Latent Concepts \\
\end{array} \]

New Document \( q \): \[ \hat{q} = Aq \in \mathbb{R}^D \]

Comparison to Sparse Coding

\[ \min_{U,A} \quad \frac{1}{2} \| X - UA \|_F^2 + \lambda \| U \|_1 \]
subject to: \[ \| A_j \|_2^2 \leq c, \quad j = 1, \ldots, M \]

New Document \( q \): \[ \hat{q} = \arg\min_{\hat{q}} \frac{1}{2} \| q - A^T \hat{q} \| + \lambda \| \hat{q} \|_1. \]

Simple Projection, Computational Efficient

Lasso Problem
More Computation Time 😞
Advantage of Sparse LSA

- **Better Interpretability:**
  Sparse LSA selects most relevant words for each topic \(a_{dj} \neq 0\).
  Compact representation of topic-word relationship.

- **Efficient Projection:**
  Sparse \(A \implies\) Efficient Projection for new documents:
  \[
  \hat{q} = Aq \in \mathbb{R}^D
  \]

- **Cheap Storage:**
  Cheap storage for sparse \(A\).

- **Sparse Projected Documents:**
  \[
  \text{sparse } \hat{q} = Aq
  \]

- **Document-Topic Relationship:**
  \[
  \hat{q}_d = 0 \iff \hat{q} \text{ not belong to } d\text{-th topic}
  \]

- **Advantage as compared to PCA:**
  Do not need to centralize \(X \implies\) destroy the sparsity of \(X\).
  Do not need the covariance matrix \(X^TX \in \mathbb{R}^{M \times M}\) may not fit in the memory for large vocabulary size.
Optimization Method

Alternating Approach

Fix $U$ and optimize with respect to $A$:

$$
\min_{A} \frac{1}{2} \|X - UA\|_F^2 + \lambda \|A\|_1
$$

subject to: $U^TU = I$

Fix $A$ and optimize with respect to $U$:

$$
\min_{U} \frac{1}{2} \|X - UA\|_F^2 \iff \text{tr}(U^TXA^T)
$$

subject to: $U^TU = I$

Closed-form Solution:

Let $V = XA^T$ (projected documents onto the latent space)

Perform SVD on $V$: $V = P\Delta Q \iff U^* = PQ$

Note: SVD on $V \in \mathbb{R}^{M \times D}$ is much cheaper than that on $X \in \mathbb{R}^{N \times M}$
Algorithm 1 Optimization Algorithm for Sparse LSA

Input: $X$, dimensionality of the latent space $D$, regularization parameter $\lambda$

Initialization: $U^0 = \begin{pmatrix} I_D \\ 0 \end{pmatrix}$

Iterate until convergence of $U$ and $A$:

1. Compute $A$ by solving $M$ independent lasso problems via coordinate descent
2. Project $X$ onto the latent space: $V = XA^T$.
3. Compute the SVD of $V$: $V = P\Delta Q$ and let $U = PQ$.

Output: Sparse projection matrix $A$. 
Extension I : Nonnegative Sparse LSA

Constraint: \( A \geq 0 \):

\[
\min_{U, A} \quad \frac{1}{2} \| X - UA \|_F^2 + \lambda \| A \|_1 \\
\text{subject to:} \quad U^T U = I, \quad A \geq 0.
\]

Simulate the probability of the word \( w_j \) given the topic \( t_d \):

Normalize each row: \( \tilde{a}_{dj} = \frac{a_{dj}}{\sum_{j=1}^{M} a_{dj}} \sim \mathbb{P}(w_j | t_d) \)

Optimization with respect to \( A \):

\[
\min_{A_{j \geq 0}} f(A_{j}) = \frac{1}{2} \| X_j - UA_j \| + \lambda \sum_{d=1}^{D} a_{dj}. \quad j = 1, \ldots, M
\]

Optimize via the coordinate descent approach:
Iterating over \( d \): fix \( a_{\hat{d}j} \) for \( \hat{d} \neq d \) and optimize over \( a_{dj} \)

\[
a_{dj}^* = \begin{cases} \frac{b_d - \lambda}{c_d} & b_d > \lambda \\ 0 & b_d \leq \lambda \end{cases}
\]

\[
c_d = \sum_{i=1}^{N} u_{id}^2, \quad b_d = \sum_{i=1}^{N} u_{id}(x_{ij} - \sum_{k \neq d} u_{ik} a_{kj}).
\]
Application: latent gene-function identification: determine relevant pathways (groups of genes) to a latent gene function (topic)

Group Structured Sparse LSA:
The set of groups of input features \( G = \{g_1, \ldots, g_{|G|}\} \) (available as a priori)

\[
\min_{U,A} \quad \frac{1}{2} \|X - UA\|_F^2 + \lambda \sum_{d=1}^{D} \sum_{g \in G} w_g \|A_{dg}\|_2
\]

subject to: \( U^T U = I \).

Optimization with respect to A:

Optimize via the coordinate descent approach:

\[
A_{dg}^* = \begin{cases} 
\frac{B_{dg}(\|B_{dg}\|_2 - \lambda w_g)}{C_d\|B_{dg}\|_2} & \|B_{dg}\|_2 > \lambda w_g \\
0 & \|B_{dg}\|_2 \leq \lambda w_g 
\end{cases}
\]

\[
C_d = \sum_{i=1}^{N} u_{id}^2, \quad (B_{dg})_{j \in g} = \sum_{i=1}^{N} u_{id}(x_{ij} - \sum_{k \neq d} u_{ik} a_{kj}).
\]
## Experimental Setup

### Methods Compared

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional LSA</td>
<td></td>
</tr>
<tr>
<td>Sparse Coding (Code from Lee et. al. 07)</td>
<td></td>
</tr>
<tr>
<td>Latent Dirichlet allocation (LDA) (Code from Blei et. al. 03)</td>
<td></td>
</tr>
<tr>
<td>Sparse LSA</td>
<td></td>
</tr>
<tr>
<td>Nonnegative Sparse LSA (NN Sparse LSA)</td>
<td></td>
</tr>
</tbody>
</table>

### Text Classification Data

<table>
<thead>
<tr>
<th>Data Set</th>
<th>N (No. of Documents)</th>
<th>M (Vocabulary Size)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 news group (20NG)</td>
<td>1,425</td>
<td>17,390</td>
</tr>
<tr>
<td>(alt.atheism vs talk.religion.misc)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RCV1 (20 classes)</td>
<td>15,564</td>
<td>7,413</td>
</tr>
</tbody>
</table>

### Topic-Word Relationship Data

<table>
<thead>
<tr>
<th>Data Set</th>
<th>N (No. of Documents)</th>
<th>M (Vocabulary Size)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIPS Proceedings from 98 to 99</td>
<td>1,714</td>
<td>13,649</td>
</tr>
</tbody>
</table>
Text Classification Performance

### 20NG

<table>
<thead>
<tr>
<th>Dimension</th>
<th>10</th>
<th>50</th>
<th>100</th>
<th>500</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparse LSA</td>
<td>1.48</td>
<td>0.80</td>
<td>0.74</td>
<td>0.32</td>
<td>0.18</td>
</tr>
<tr>
<td>NN Sparse LSA</td>
<td>1.44</td>
<td>0.72</td>
<td>0.55</td>
<td>0.31</td>
<td>0.17</td>
</tr>
<tr>
<td>Other Methods</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

**20NG:** Density of $A$ (%) ($\lambda=0.05$)

### RCV1

<table>
<thead>
<tr>
<th>Dimension</th>
<th>10</th>
<th>50</th>
<th>100</th>
<th>500</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparse LSA</td>
<td>13.52</td>
<td>7.46</td>
<td>7.40</td>
<td>2.71</td>
<td>1.13</td>
</tr>
<tr>
<td>NN Sparse LSA</td>
<td>11.65</td>
<td>4.97</td>
<td>0.40</td>
<td>1.91</td>
<td>0.79</td>
</tr>
<tr>
<td>Other Methods</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

**RCV1:** Density of $A$ (%) ($\lambda=0.05$)

**Conclusion:** For large $D$, the classification performance of Sparse LSA is almost the same as LSA but with a much more sparse projection matrix $A$. 

---

[Image of bar charts and tables showing classification accuracy for different methods and dimensions.]
# Efficiency and Storage

<table>
<thead>
<tr>
<th>20NG</th>
<th>Proj. Time (ms)</th>
<th>Storage (MB)</th>
<th>Density of Proj. Doc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparse LSA</td>
<td>0.25 (4.05E-2)</td>
<td>0.6314</td>
<td>35.81 (15.39)</td>
</tr>
<tr>
<td>NN Sparse LSA</td>
<td>0.22 (2.78E-2)</td>
<td>0.6041</td>
<td>35.44 (15.17)</td>
</tr>
<tr>
<td>LSA</td>
<td>31.6 (1.10)</td>
<td>132.68</td>
<td>100 (0)</td>
</tr>
<tr>
<td>Sparse Coding</td>
<td>1711.1 (323.9)</td>
<td>132.68</td>
<td>86.94 (3.63)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RCV1</th>
<th>Proj. Time (ms)</th>
<th>Storage (MB)</th>
<th>Density of Proj. Doc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparse LSA</td>
<td>0.59 (7.36E-2)</td>
<td>1.3374</td>
<td>55.38 (11.77)</td>
</tr>
<tr>
<td>NN Sparse LSA</td>
<td>0.46 (6.66E-2)</td>
<td>0.9537</td>
<td>46.47 (11.90)</td>
</tr>
<tr>
<td>LSA</td>
<td>13.2 (0.78)</td>
<td>113.17</td>
<td>100 (0)</td>
</tr>
<tr>
<td>Sparse Coding</td>
<td>370.5 (23.3)</td>
<td>113.17</td>
<td>83.88 (2.11)</td>
</tr>
</tbody>
</table>

**Conclusion:** Sparse LSA or NN Sparse LSA

- Efficient projection with less time
- Less storage for the projection matrix $A$
- Sparse projected documents: more efficient for subsequent retrieval tasks, e.g. ranking, text categorization, etc

$D = 1,000, \lambda = 0.05$  
Table entry: mean (std)
# Topic–Word Relationship

## Nonnegative Sparse LSA

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>network</td>
<td>learning</td>
<td>network</td>
<td>model</td>
</tr>
<tr>
<td>neural</td>
<td>reinforcement</td>
<td>learning</td>
<td>data</td>
</tr>
<tr>
<td>networks</td>
<td>algorithm</td>
<td>data</td>
<td>models</td>
</tr>
<tr>
<td>system</td>
<td>function</td>
<td>neural</td>
<td>parameters</td>
</tr>
<tr>
<td>neurons</td>
<td>rule</td>
<td>training</td>
<td>mixture</td>
</tr>
<tr>
<td>neuron</td>
<td>control</td>
<td>set</td>
<td>likelihood</td>
</tr>
<tr>
<td>input</td>
<td>learn</td>
<td>function</td>
<td>distribution</td>
</tr>
<tr>
<td>output</td>
<td>weight</td>
<td>model</td>
<td>gaussian</td>
</tr>
<tr>
<td>time</td>
<td>action</td>
<td>input</td>
<td>em</td>
</tr>
<tr>
<td>systems</td>
<td>policy</td>
<td>networks</td>
<td>variables</td>
</tr>
</tbody>
</table>

## LDA

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>learning</td>
<td>figure</td>
<td>algorithm</td>
<td>single</td>
</tr>
<tr>
<td>data</td>
<td>model</td>
<td>method</td>
<td>general</td>
</tr>
<tr>
<td>model</td>
<td>output</td>
<td>networks</td>
<td>sets</td>
</tr>
<tr>
<td>training</td>
<td>neurons</td>
<td>process</td>
<td>time</td>
</tr>
<tr>
<td>information</td>
<td>vector</td>
<td>learning</td>
<td>maximum</td>
</tr>
<tr>
<td>number</td>
<td>networks</td>
<td>input</td>
<td>paper</td>
</tr>
<tr>
<td>algorithm</td>
<td>state</td>
<td>based</td>
<td>rates</td>
</tr>
<tr>
<td>performance</td>
<td>layer</td>
<td>function</td>
<td>features</td>
</tr>
<tr>
<td>linear</td>
<td>system</td>
<td>error</td>
<td>estimated</td>
</tr>
<tr>
<td>input</td>
<td>order</td>
<td>parameter</td>
<td>neural</td>
</tr>
</tbody>
</table>

## Conclusion

The topics learned by **NN Sparse LSA** are discriminative while the topics learned by **LDA** are all closely related to **neural network**.
Thank You!