Welcome
2017 Faculty Summit Attendees

Faculty Summit 2017
microsoftfacultysummit.com

Microsoft Research
Microsoft.com/research

Facebook
@microsoftresearch

Twitter
@MSFTResearch
#FacSumm
#EdgeofAI

Microsoft Research
Faculty Summit 2017
Discovering hidden structure in the "sparse" regime

Sham M. Kakade
Statistics Computer Science & Engineering
University of Washington
Discovering hidden structure in the “sparse” regime

Sham M. Kakade

Statistics
Computer Science & Engineering

University of Washington
Unsupervised Learning

- Topic models

Can we efficiently estimate the hidden structure, without cluster labels?
We see a matrix of counts.

example: \((i, j)\) entry is \# times word \(i\) preceded word \(j\).
We see a matrix of counts.

example: \((i, j)\) entry is \# times word \(i\) preceded word \(j\).

We want to permute the rows/columns to make the structure clear.
How do we learn latent factor models?

- **Practical heuristics:**
  - $k$-Means / EM / Gibbs sampling
  - Related: (Word) embedding methods  
    (e.g. Word2Vec, Glove)

- **Theory:** spectral (and tensor) methods provide provably learnable algorithms
  - Topic models, hidden Markov models, mixtures of Gaussian  
    [Anandkumar, Ge, Hsu, K. Telgarsky (2012)]
  - Must use three way correlations/ “count tensors”
  - provably efficient in that they have polynomial sample size and polynomial runtime.

**Question:** How does one do this well?  
“provably” -> “automated, practical, and scalable”
An automated algorithm I’d recommend....

Word embedding methods from count matrices:
[Stratos, Kim, Collins, & Hsu; 2014], [Stratos, Collins, & Hsu; 2015]

- For each word \( w \), predict the following words.
  (prediction done with an SVD)
- Let \( \phi(w) \) be this (vector) prediction.

\[
\text{\( w \longrightarrow \phi(w) \)}
\]

- Cluster \( \phi(\text{word 1}) \), \( \phi(\text{word 2}) \), \( \phi(\text{word 3}) \), ... .
- Representation competitive with Word2Vec and Glove.

Related: Brown clustering [Liang ’05]
But our data are sparse...

Word distributions are heavy tailed:

- (also) health care data, comp. bio /genomics, natural language processing
- A sample size that is super-linear (in the vocabulary size) is not practical.
  - Spectral methods fail in the sparse regime.
  - practical methods often transform the count data.
  - RNNs/LSTMs may fail in this regime: [Chen, Grangier & Auli; 2015]
- Provable learning here may be important for practice:
  - The sparse regime is the “small sample” regime.
  - This question is fundamentally related to ideas in sparse random graph theory and community detection.
The joint distribution of two words in a document:

\[ \text{Bi}_{i,j} = \frac{1}{2} p_i p_j + \frac{1}{2} q_i q_j \]

Have \( M \) words in the vocabulary.

**Bigram structure:** an \( M \times M \) matrix:

\[ \text{Bi} = \frac{1}{2} pp^\top + \frac{1}{2} qq^\top \]

Sub-problem (matrix recovery problem) for:
- topic models, hidden Markov models, LDA, community detection
This talk: Optimal estimation of (two) topic models.

- sample size is **LINEAR** in the vocabulary size.
  - applicable to heavy tailed distributions.
- **Key Idea:** a natural subproblem that governs the difficulty of the problem.
- Many connections to sparse random graph theory and community detection.
The spectral idea (and its failure)

- note:
  \[ B_i = \frac{1}{4} (p + q)(p + q)^\top + \frac{1}{4} (p - q)(p - q)^\top \]

- the marginal probability of a word is: \( \rho_w = \frac{1}{2} \rho_w + \frac{1}{2} q_w \), so:
  \[ B_i - \rho \rho^\top = \frac{1}{4} (p - q)(p - q)^\top \]

- do an eigendecomposition

This idea fails due to the “heavy” elements.

- same issue for spectral methods in community detection.
- need \( O(M \log M) \) samples.
Main Result: Estimation in the Sparse Regime

- Suppose we have $M$ words in our dictionary
- Suppose $\Delta := \|p - q\|_1$ is a constant.

**Theorem**

(Huang, K., Kong, Valiant) There is an algorithm which, with high probability and when provided with $N$ sampled bigrams, returns estimates $\hat{p}$ and $\hat{q}$ where:

$$
\|\hat{B}_i - B_i\|_1 \leq c \sqrt{\frac{M}{N}}, \quad \|\hat{p} - p\|_1 \leq c \sqrt{\frac{M}{N}}, \quad \|\hat{q} - q\|_1 \leq c \sqrt{\frac{M}{N}},
$$

where $c$ is a constant.

This is linear sample size. (general $k$-case forthcoming)
What governs the difficulty of learning?

- suppose the $L_1$ distance

$$\Delta := \|p - q\|_1$$

is a constant.

- this means there exists an “anchor set” $\mathcal{A}$ and $\mathcal{B}$ s.t.

$$p(\mathcal{A}) - q(\mathcal{A}) > \text{constant} \times \Delta$$

$$q(\mathcal{B}) - p(\mathcal{B}) > \text{constant} \times \Delta$$

Suppose we knew some “anchor set”.

S. M. Kakade (UW)  Spectral Learning
Suppose we knew such an anchor set.

Let’s consider a superword where we merge $\mathcal{A}$ into a single word and we merge $\mathcal{B}$ into a single word.

for any word $w$, let’s look at the bigram probabilities $B_{i_{w,\mathcal{A}}}$ and $B_{i_{w,\mathcal{B}}}$:

$$B_{i_{w,\mathcal{A}}} = \frac{1}{2} p_w p_{\mathcal{A}} + \frac{1}{2} q_w q_{\mathcal{A}}$$

$$B_{i_{w,\mathcal{B}}} = \frac{1}{2} p_w p_{\mathcal{B}} + \frac{1}{2} q_w q_{\mathcal{B}}$$

Algorithm which knows the anchor set: First, use spectral to estimate $p_{\mathcal{A}}$ and $q_{\mathcal{A}}$. Second, estimate and $p_w$ and $q_w$ by solving (separate) linear equations.

Clearly linear sample size for this algorithm.
How do we find the anchor sets?

Phase I: Finding the anchor sets
The mix of “heavy” and “light” words are problematic.

- Step 1: Create empirical buckets with words of roughly the same frequencies.
- Step 2: Carefully reconstruct the matrix, restricted to each sub-matrix.
- Step 3: Find the anchor sets in each sub-matrix (if they exist) and merge.

Techniques: borrow from random graph theory/comm. detection literature.
Le, Levina, & Vershynin (2014); Le & Vershynin; (2015)

Phase II: Use the anchor sets to fill in the matrix.
Anchor set discovery, Step 1: Empirical bucketing

- construct estimates $\hat{\rho}_w$ of the empirical frequency of word $w$.
- Collect all the “heavy” words into one bin:
  \[ I_{\text{heavy}} = \{ w : \text{s.t. } \hat{\rho}_w > \log M / M \} \]
- Create “light” bins.
  \[ I_k = \left\{ w : \text{s.t. } \frac{2^k}{M} \leq \hat{\rho}_w < \frac{2^{k+1}}{M} \right\} \]
  for $k = 0, \ldots \log \log M$.
- There is spill over.
The “heavy” bucket
- We have good estimates here.
- Do spectral and reconstruct the sub-matrix.

The “light” bins:
- Empirical estimates are not good. Use regularized algorithms:
  Le, Levina, & Vershynin (2014); Le & Vershynin; (2015)
- algo: draw a fresh sample; throw away “heavy”; run spectral

technical lemma: “many” sub-matrices, which contain a constant fraction of the anchor set, will be estimated accurately
Anchor set discover, Step 3: Find and merge anchor sub-sets

- if a sub-matrix was estimated correctly, then either
  - if it is rank 2, extract two anchor sub-sets.
  - if it is rank 1, either there is no anchor set OR all the words belong in one anchor sub-set.
- Merge the anchor sub-sets from each sub-matrix (as they co-occur together).
- Guarantee: we will find anchor sets with constant separation.
Known:

- Need to use three way correlation.
  - count matrices $\rightarrow$ count tensors.
- again, sample size may be too large for current algorithms.
- Lower Bound Theorem: (Huang, K., Kong, Valiant) Testing if the model is an HMM is just as hard as learning (in terms of sample size).

Open:

- figure out informationally optimal $k > 2$ case (must use tensors)
- low rank reconstruction/approximation questions
A plug:
The MusicNet Dataset
MusicNet (large labeled dataset for music)

New (polyphonic) dataset with start/end times of each note labeled.

<table>
<thead>
<tr>
<th>Minutes</th>
<th>Labels</th>
<th>Recordings</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2,357</td>
<td>1,316,868</td>
<td>383</td>
<td>4.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ensemble</th>
<th>Minutes</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solo Piano</td>
<td>1,058</td>
<td>485,459</td>
</tr>
<tr>
<td>String Quartet</td>
<td>496</td>
<td>299,363</td>
</tr>
<tr>
<td>Accompanied Violin</td>
<td>166</td>
<td>120,594</td>
</tr>
<tr>
<td>Piano Trio</td>
<td>97</td>
<td>86,123</td>
</tr>
<tr>
<td>Piano Quartet</td>
<td>73</td>
<td>60,594</td>
</tr>
<tr>
<td>String Sextet</td>
<td>56</td>
<td>42,920</td>
</tr>
<tr>
<td>Accompanied Cello</td>
<td>62</td>
<td>38,076</td>
</tr>
<tr>
<td>Piano Quintet</td>
<td>25</td>
<td>27,566</td>
</tr>
<tr>
<td>Wind Quintet</td>
<td>42</td>
<td>24,832</td>
</tr>
<tr>
<td>Horn Piano Trio</td>
<td>29</td>
<td>18,834</td>
</tr>
<tr>
<td>Wind Sextet</td>
<td>30</td>
<td>16,955</td>
</tr>
<tr>
<td>Clarinet Quintet</td>
<td>32</td>
<td>16,523</td>
</tr>
<tr>
<td>Wind Octet</td>
<td>22</td>
<td>14,659</td>
</tr>
<tr>
<td>Clarinet-Cello-Piano Trio</td>
<td>24</td>
<td>13,451</td>
</tr>
<tr>
<td>Solo Violin</td>
<td>36</td>
<td>11,973</td>
</tr>
<tr>
<td>Solo Cello</td>
<td>49</td>
<td>10,876</td>
</tr>
<tr>
<td>Violin and Harpsichord</td>
<td>15</td>
<td>7,470</td>
</tr>
<tr>
<td>Viola Quintet</td>
<td>15</td>
<td>4,160</td>
</tr>
<tr>
<td>Solo Flute</td>
<td>7</td>
<td>2,214</td>
</tr>
<tr>
<td>Accompanied Clarinet</td>
<td>20</td>
<td>2,193</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Composer</th>
<th>Minutes</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beethoven</td>
<td>1,274</td>
<td>725,092</td>
</tr>
<tr>
<td>Schubert</td>
<td>286</td>
<td>162,394</td>
</tr>
<tr>
<td>Brahms</td>
<td>217</td>
<td>148,292</td>
</tr>
<tr>
<td>Mozart</td>
<td>228</td>
<td>137,797</td>
</tr>
<tr>
<td>Dvorak</td>
<td>63</td>
<td>34,827</td>
</tr>
<tr>
<td>Bach</td>
<td>190</td>
<td>33,266</td>
</tr>
<tr>
<td>Cambini</td>
<td>42</td>
<td>24,832</td>
</tr>
<tr>
<td>Faure</td>
<td>32</td>
<td>22,558</td>
</tr>
<tr>
<td>Ravel</td>
<td>27</td>
<td>21,382</td>
</tr>
<tr>
<td>Haydn</td>
<td>14</td>
<td>6,428</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Minutes</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Piano</td>
<td>1,558</td>
<td>732,027</td>
</tr>
<tr>
<td>Violin</td>
<td>1,045</td>
<td>269,865</td>
</tr>
<tr>
<td>Cello</td>
<td>933</td>
<td>116,829</td>
</tr>
<tr>
<td>Viola</td>
<td>699</td>
<td>116,238</td>
</tr>
<tr>
<td>Clarinet</td>
<td>173</td>
<td>27,901</td>
</tr>
<tr>
<td>Bassoon</td>
<td>96</td>
<td>16,629</td>
</tr>
<tr>
<td>Horn</td>
<td>125</td>
<td>12,426</td>
</tr>
<tr>
<td>Oboe</td>
<td>65</td>
<td>8,710</td>
</tr>
<tr>
<td>Flute</td>
<td>50</td>
<td>8,317</td>
</tr>
<tr>
<td>Harpsichord</td>
<td>15</td>
<td>4,914</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Piano</th>
<th>Violin</th>
<th>Cello</th>
<th>Viola</th>
<th>Clarinet</th>
<th>Bassoon</th>
<th>Horn</th>
<th>Oboe</th>
<th>Flute</th>
<th>Harpsichord</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,558</td>
<td>732,027</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,045</td>
<td>269,865</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>933</td>
<td>116,829</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>699</td>
<td>116,238</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>173</td>
<td>27,901</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>96</td>
<td>16,629</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>125</td>
<td>12,426</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>8,710</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>8,317</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>4,914</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example Labels in MusicNet

<table>
<thead>
<tr>
<th>Start</th>
<th>End</th>
<th>Instrument</th>
<th>Note</th>
<th>Measure</th>
<th>Beat</th>
<th>Note Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>45.29</td>
<td>45.49</td>
<td>Violin</td>
<td>G5</td>
<td>21</td>
<td>3</td>
<td>Eighth</td>
</tr>
<tr>
<td>48.99</td>
<td>50.13</td>
<td>Cello</td>
<td>A#3</td>
<td>24</td>
<td>2</td>
<td>Dotted Half</td>
</tr>
<tr>
<td>82.91</td>
<td>83.12</td>
<td>Viola</td>
<td>C5</td>
<td>51</td>
<td>2.5</td>
<td>Eighth</td>
</tr>
</tbody>
</table>

All music is freely licensed, under Creative Commons.
MusicNet: Learning Features from “Scratch”

preliminary results

[Waveform and Fourier transform diagrams]
Thanks!

- Obtaining the information theoretical limit for learning 2-factor latent models.
  - new algorithmic insights
  - future: new practical algorithms + new theory
    - forthcoming: generalize to $k > 2$ and extend to richer models

Collaborators:

Q. Huang  W. Kong  G. Valiant
Thank you