Method-of-moments

Daniel Hsu
Example: modeling the topics of a document corpus

Goal: model the topics of document in a corpus.

Sample of documents \rightarrow Learning algorithm \rightarrow Model parameters \rightarrow \theta
**Topic model** (e.g., Hofmann, '99; Blei-Ng-Jordan, '03)

$k$ topics (distributions over vocab words).
Each document $\leftrightarrow$ mixture of topics.
Words in document $\sim_{iid}$ mixture dist.
Topic model \((e.g., \text{Hofmann}, '99; \text{Blei-Ng-Jordan}, '03)\)

\[ k \text{ topics (distributions over vocab words).} \]

Each document \(\leftrightarrow\) mixture of topics.

Words in document \(\sim_{\text{iid}}\) mixture dist.

\[ \text{E.g.,} \]

\[ \sim_{\text{iid}} 0.6 \cdot \text{sports} + 0.3 \cdot \text{science} + 0.1 \cdot \text{politics} + 0. \cdot \text{business} \]

<table>
<thead>
<tr>
<th>Term</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>aardvark</td>
<td>0</td>
</tr>
<tr>
<td>athlete</td>
<td>3</td>
</tr>
<tr>
<td>zygote</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ \Pr_{\theta}[\text{"play" | sports}] = 0.0002 \]
\[ \Pr_{\theta}[\text{"game" | sports}] = 0.0003 \]
\[ \Pr_{\theta}[\text{"season" | sports}] = 0.0001 \]
Learning topic models

Topic model:

$k$ topics (dists. over $d$ words) $\vec{\mu}_1, \ldots, \vec{\mu}_k$;
Each document $\leftrightarrow$ mixture of topics.
Words in document $\sim_{\text{iid}}$ mixture dist.
Learning topic models

Simple topic model: (each document about *single* topic)

$k$ topics (dists. over $d$ words) $\vec{\mu}_1, \ldots, \vec{\mu}_k$;
Topic $t$ chosen with prob. $w_t$,
words in document $\sim_{iid} \vec{\mu}_t$. 

Input: sample of documents, generated by simple topic model with unknown parameters $\theta^\star := \{(\vec{\mu}_t^\star, w_t^\star)\}$.

Task: find parameters $\theta := \{(\vec{\mu}_t, w_t)\}$ so that $\theta \approx \theta^\star$. 

Learning topic models

**Simple topic model**: (each document about *single* topic)

- $k$ topics (dists. over $d$ words) $\vec{\mu}_1, \ldots, \vec{\mu}_k$;
- Topic $t$ chosen with prob. $w_t$;
- words in document $\sim$ iid $\vec{\mu}_t$.

**Input**: sample of documents, generated by simple topic model with unknown parameters $\theta^* := \{(\vec{\mu}_t^*, w_t^*)\}$. 

![Bar chart with categories sports, science, politics, business showing topic distributions](chart)
Learning topic models

Simple topic model: (each document about single topic)

- $k$ topics (dists. over $d$ words) $\vec{\mu}_1, \ldots, \vec{\mu}_k$;
- Topic $t$ chosen with prob. $w_t$,
- words in document $\sim$ iid $\vec{\mu}_t$.

▶ **Input:** sample of documents, generated by simple topic model with unknown parameters $\theta^* := \{(\vec{\mu}_t^*, w_t^*)\}$.

▶ **Task:** find parameters $\theta := \{(\vec{\mu}_t, w_t)\}$ so that $\theta \approx \theta^*$.
Some approaches to estimation

Maximum-likelihood (e.g., Fisher, 1912).\[\theta_{MLE} = \arg \max_\theta \Pr(\text{data}).\]

Current practice (> 40 years): local search for local maxima — can be quite far from \(\theta_{MLE}\).

Method-of-moments (Pearson, 1894).

Find parameters \(\theta\) that (approximately) satisfy system of equations based on the data.

Many ways to instantiate & implement.
Some approaches to estimation

Maximum-likelihood (e.g., Fisher, 1912).

\[ \theta_{MLE} := \arg \max_{\theta} \Pr_{\theta}[\text{data}] . \]
Some approaches to estimation

**Maximum-likelihood** (e.g., Fisher, 1912).

\[ \theta_{\text{MLE}} := \arg \max_\theta \Pr_\theta[\text{data}] \]

Current practice (> 40 years): **local search** for local maxima — can be quite far from \( \theta_{\text{MLE}} \).
Some approaches to estimation

**Maximum-likelihood**  (e.g., Fisher, 1912).

\[ \theta_{\text{MLE}} := \arg \max_{\theta} \Pr_{\theta}[\text{data}] . \]

Current practice (> 40 years): **local search** for local maxima — can be quite far from \( \theta_{\text{MLE}} \).

**Method-of-moments**  (Pearson, 1894).

Find parameters \( \theta \) that (approximately) *satisfy system of equations* based on the data.
Some approaches to estimation

**Maximum-likelihood** (e.g., Fisher, 1912).

\[ \theta_{\text{MLE}} := \arg \max_{\theta} \Pr_{\theta}[\text{data}] \].

Current practice (> 40 years): local search for local maxima — can be quite far from \( \theta_{\text{MLE}} \).

**Method-of-moments** (Pearson, 1894).

Find parameters \( \theta \) that (approximately) satisfy system of equations based on the data.

Many ways to instantiate & implement.
Some approaches to estimation

Maximum-likelihood  (e.g., Fisher, 1912).
\[ \theta_{\text{MLE}} := \arg \max_{\theta} \Pr_{\theta} [\text{data}] \]

Current practice (> 40 years): local search for local maxima — can be quite far from \(\theta_{\text{MLE}}\).

Method-of-moments  (Pearson, 1894).
Find parameters \(\theta\) that (approximately) satisfy system of equations based on the data.

Many ways to instantiate & implement.
Moments: normal distribution

**Normal distribution:** $x \sim \mathcal{N}(\mu, \nu)$

**First- and second-order moments:**

\[
E_{(\mu, \nu)}[x] = \mu, \quad E_{(\mu, \nu)}[x^2] = \mu^2 + \nu.
\]
Moments: normal distribution

Normal distribution: \( x \sim \mathcal{N}(\mu, \nu) \)

First- and second-order moments:

\[
\mathbb{E}_{(\mu, \nu)}[x] = \mu, \quad \mathbb{E}_{(\mu, \nu)}[x^2] = \mu^2 + \nu.
\]

Method-of-moments estimators of \( \mu^* \) and \( \nu^* \):
find \( \hat{\mu} \) and \( \hat{\nu} \) s.t.

\[
\hat{\mathbb{E}}_S[x] \approx \hat{\mu}, \quad \hat{\mathbb{E}}_S[x^2] \approx \hat{\mu}^2 + \hat{\nu}.
\]
Moments: normal distribution

Normal distribution: $x \sim \mathcal{N}(\mu, \nu)$

First- and second-order moments:

$$
E_{(\mu, \nu)}[x] = \mu, \quad E_{(\mu, \nu)}[x^2] = \mu^2 + \nu.
$$

Method-of-moments estimators of $\mu^*$ and $\nu^*$:

find $\hat{\mu}$ and $\hat{\nu}$ s.t.

$$
\hat{E}_S[x] \approx \hat{\mu}, \quad \hat{E}_S[x^2] \approx \hat{\mu}^2 + \hat{\nu}.
$$

A reasonable solution:

$$
\hat{\mu} := \hat{E}_S[x], \quad \hat{\nu} := \hat{E}_S[x^2] - \hat{\mu}^2
$$

since $\hat{E}_S[x] \to E_{(\mu^*, \nu^*)}[x]$ and $\hat{E}_S[x^2] \to E_{(\mu^*, \nu^*)}[x^2]$ by LLN.
Moments: simple topic model

For any $n$-tuple $(i_1, i_2, \ldots, i_n) \in \text{Vocabulary}^n$:

*(Population) moments* under some parameter $\theta$:

$$\Pr_\theta \left[ \text{document contains words } i_1, i_2, \ldots, i_n \right].$$

e.g., $\Pr_\theta [\text{“machine” & “learning” co-occur}].$
Moments: simple topic model

For any \( n \)-tuple \((i_1, i_2, \ldots, i_n) \in \text{Vocabulary}^n\):

\textit{(Population) moments} under some parameter \( \theta \):

\[
\text{Pr}_\theta \left[ \text{document contains words } i_1, i_2, \ldots, i_n \right].
\]

\textit{e.g.}, \( \text{Pr}_\theta [\text{“machine” & “learning” co-occur}] \).

\textit{Empirical moments} from sample \( S \) of documents:

\[
\hat{\text{Pr}}_S \left[ \text{document contains words } i_1, i_2, \ldots, i_n \right]
\]

\textit{i.e.}, empirical frequency of co-occurrences \textit{in sample} \( S \).
Method-of-moments

Method-of-moments strategy:
Given data sample $S$, find $\theta$ to satisfy system of equations
\[ \text{moments}_\theta = \hat{\text{moments}}_S. \]

(Recall: we expect $\hat{\text{moments}}_S \approx \text{moments}_{\theta^*}$ by LLN.)

Q1. Which moments should we use?
Q2. How do we (approx.) solve these moment equations?
Q1. Which moments should we use?
Q1. Which moments should we use?

<table>
<thead>
<tr>
<th>moment order</th>
<th>reliable estimates?</th>
<th>unique solution?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st, 2nd</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1\textsuperscript{st} and 2\textsuperscript{nd}-order moments \((e.g.,\) prob. of word pairs)\).

\begin{itemize}
  \item [Arora-Ge-Moitra, ’12]
  \item [Kleinberg-Sandler, ’04]
  \item [Vempala-Wang, ’02]
  \item [McSherry, ’01]
\end{itemize}

\[\Omega(k)\textsuperscript{th}\]
Q1. Which moments should we use?

<table>
<thead>
<tr>
<th>moment order</th>
<th>reliable estimates?</th>
<th>unique solution?</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1^{\text{st}}$, $2^{\text{nd}}$</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

$1^{\text{st}}$- and $2^{\text{nd}}$-order moments (e.g., prob. of word pairs).

- Fairly easy to get reliable estimates.
  
  \[
  \hat{\Pr}_{S}[\text{“machine”, “learning”}] \approx \Pr_{\theta^*}[\text{“machine”, “learning”}] 
  \]

[Arora-Ge-Moitra, ’12]
[Kleinberg-Sandler, ’04]
[Vempala-Wang, ’02]
[McSherry, ’01]

$\Omega(k)^{\text{th}}$ order of moments
Q1. Which moments should we use?

<table>
<thead>
<tr>
<th>moment order</th>
<th>reliable estimates?</th>
<th>unique solution?</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1^{\text{st}}$, $2^{\text{nd}}$</td>
<td>✓</td>
<td>X</td>
</tr>
</tbody>
</table>

$1^{\text{st}}$- and $2^{\text{nd}}$-order moments (e.g., prob. of word pairs).

- Fairly easy to get reliable estimates.

$$\hat{\Pr}_S[\text{“machine”, “learning”}] \approx \Pr_{\theta^*}[\text{“machine”, “learning”}]$$

- Can have multiple solutions to moment equations.

$$\text{moments}_{\theta_1} = \underbrace{\text{moments}} = \text{moments}_{\theta_2}, \quad \theta_1 \neq \theta_2$$

[Arora-Ge-Moitra, ’12]
[Kleinberg-Sandler, ’04]
[Vempala-Wang, ’02]
[McSherry, ’01]

$\Omega(k)^{\text{th}}$ order of moments
Q1. Which moments should we use?

<table>
<thead>
<tr>
<th>moment order</th>
<th>reliable estimates?</th>
<th>unique solution?</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1^{st}$, $2^{nd}$</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>$\Omega(k)^{th}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$\Omega(k)^{th}$-order moments (prob. of word $k$-tuples)

[Arora-Ge-Moitra, ’12]
[Kleinberg-Sandler, ’04]
[Vempala-Wang, ’02]
[McSherry, ’01]

[Gravin et al, ’12]
[Moitra-Valiant, ’10]
[Lindsay, ’89]
[Prony, 1795]
Q1. Which moments should we use?

<table>
<thead>
<tr>
<th>moment order</th>
<th>reliable estimates?</th>
<th>unique solution?</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1^{st}$, $2^{nd}$</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>$\Omega(k)^{th}$</td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

$\Omega(k)^{th}$-order moments (prob. of word $k$-tuples)

- Uniquely pins down the solution.

[Arora-Ge-Moitra, ’12]
[Kleinberg-Sandler, ’04]
[Vempala-Wang, ’02]
[McSherry, ’01]

[Gravin et al, ’12]
[Moitra-Valiant, ’10]
[Lindsay, ’89]
[Prony, 1795]
# Q1. Which moments should we use?

<table>
<thead>
<tr>
<th>moment order</th>
<th>reliable estimates?</th>
<th>unique solution?</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1^{st}$, $2^{nd}$</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>$\Omega(k)^{th}$</td>
<td>✗</td>
<td>✓</td>
</tr>
</tbody>
</table>

$\Omega(k)^{th}$-order moments (prob. of word $k$-tuples)

- Uniquely pins down the solution.
- Empirical estimates very unreliable.

[Arora-Ge-Moitra, ’12]  
[Kleinberg-Sandler, ’04]  
[Vempala-Wang, ’02]  
[McSherry, ’01]  

[Gravin et al, ’12]  
[Moitra-Valiant, ’10]  
[Lindsay, ’89]  
[Prony, 1795]
Q1. Which moments should we use?

<table>
<thead>
<tr>
<th>moment order</th>
<th>reliable estimates?</th>
<th>unique solution?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1\textsuperscript{st}, 2\textsuperscript{nd}</td>
<td>✔</td>
<td>✗</td>
</tr>
<tr>
<td>$\Omega(k)^{th}$</td>
<td>✗</td>
<td>✔</td>
</tr>
</tbody>
</table>

Can we get best-of-both-worlds?

[Arora-Ge-Moitra, '12]
[Kleinberg-Sandler, '04]
[Vempala-Wang, '02]
[McSherry, '01]

[Gravin et al, '12]
[Moitra-Valiant, '10]
[Lindsay, '89]
[Prony, 1795]

1\textsuperscript{st} 2\textsuperscript{nd} order of moments $\Omega(k)^{th}$
Q1. Which moments should we use?

<table>
<thead>
<tr>
<th>moment order</th>
<th>reliable estimates?</th>
<th>unique solution?</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1^{st}$, $2^{nd}$</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>$\Omega(k)^{th}$</td>
<td>✗</td>
<td>✓</td>
</tr>
</tbody>
</table>

Can we get best-of-both-worlds? **Yes!**

**In high-dimensions, low-order multivariate moments suffice.**

(1$^{st}$-, 2$^{nd}$-, and 3$^{rd}$-order moments)

---

[Arora-Ge-Moitra, ’12]  
[Kleinberg-Sandler, ’04]  
[Vempala-Wang, ’02]  
[McSherry, ’01]  

[Gravin et al, ’12]  
[Moitra-Valiant, ’10]  
[Lindsay, ’89]  
[Prony, 1795]  

this work
Low-order multivariate moments suffice

**Key observation:** in high dimensions \((d \gg k)\), low-order moments have simple ("low-rank") algebraic structure.
Low-order multivariate moments suffice

\[ P_\theta := \Pr_\theta[\text{words } i, j] \quad \text{(Empirical: } \hat{P}) \]

Given a document about topic \( t \),

\[ \Pr_\theta[\text{words } i, j \mid \text{topic } t] = (\bar{\mu}_t)_i \cdot (\bar{\mu}_t)_j. \]
Low-order multivariate moments suffice

Given a document about topic $t$, 

$$\Pr_\theta[\text{words } i, j \mid \text{topic } t ] = (\vec{\mu}_t \otimes \vec{\mu}_t)_{i,j}. $$
Low-order multivariate moments suffice

\[ P_\theta := \Pr_\theta[\text{words } i, j] \]

(Empirical: \( \hat{P} \))

Averaging over topics,

\[ \Pr_\theta[\text{words } i, j] = \sum_t w_t \cdot (\vec{\mu}_t \otimes \vec{\mu}_t)_{i,j}. \]
Low-order multivariate moments suffice

\[ P_\theta := \text{Pr}_\theta[\text{words } i, j] \]

(Empirical: \( \hat{P} \))

In matrix notation \( P_\theta \),

\[ P_\theta = \sum_t w_t \, \vec{\mu}_t \otimes \vec{\mu}_t. \]
Low-order multivariate moments suffice

\[ T_\theta := \text{(Empirical: } \hat{T} \text{)} \]

Similarly,

\[ \Pr_\theta[ \text{words } i, j, k ] = \sum_t w_t \cdot (\vec{\mu}_t \otimes \vec{\mu}_t \otimes \vec{\mu}_t)_{i,j,k}. \]
Low-order multivariate moments suffice

\[ T_\theta := \begin{array}{c}
\text{(Empirical: } \hat{T}) \\
\Pr_{\theta}[\text{words } i, j, k] \\
\end{array} \]

In tensor notation \( T_\theta \),

\[ T_\theta = \sum_t w_t \vec{\mu}_t \otimes \vec{\mu}_t \otimes \vec{\mu}_t. \]
Low-order multivariate moments suffice

\[ P_\theta := \sum_{i,j} \text{Pr}_\theta[\text{words } i, j] \]

\[ T_\theta := \sum_{i,j,k} \text{Pr}_\theta[\text{words } i, j, k] \]

\[ P_\theta = \sum_{t=1}^{k} w_t \vec{\mu}_t \otimes \bar{\vec{\mu}}_t \quad \text{and} \quad T_\theta = \sum_{t=1}^{k} w_t \vec{\mu}_t \otimes \bar{\vec{\mu}}_t \otimes \bar{\bar{\vec{\mu}}}_t \]

Claim: \( P_\theta \) and \( T_\theta \) uniquely determine the parameters \( \theta \).
Low-order multivariate moments suffice

\[ P_\theta := \sum_{t=1}^{k} w_t \vec{\mu}_t \otimes \vec{\mu}_t \]  
and  
\[ T_\theta := \sum_{t=1}^{k} w_t \vec{\mu}_t \otimes \vec{\mu}_t \otimes \vec{\mu}_t \]

Low-rank matrix and tensor
Low-order multivariate moments suffice

\[ P_\theta := \text{Pr}_\theta[\text{words } i, j] \]
\[ T_\theta := \text{Pr}_\theta[\text{words } i, j, k] \]

Moment equations: \( P_\theta = \hat{P}, \ T_\theta = \hat{T} \)
(i.e., find low-rank decompositions of empirical moments).
Low-order multivariate moments suffice

\[ P_\theta := \text{Pr}_\theta[\text{words } i, j] \]

\[ T_\theta := \text{Pr}_\theta[\text{words } i, j, k] \]

\[ P_\theta = \sum_{t=1}^{k} w_t \vec{\mu}_t \otimes \vec{\mu}_t \quad \text{and} \quad T_\theta = \sum_{t=1}^{k} w_t \vec{\mu}_t \otimes \vec{\mu}_t \otimes \vec{\mu}_t \]

**Moment equations:** \( P_\theta = \hat{P}, \ T_\theta = \hat{T} \)

(i.e., find low-rank decompositions of empirical moments).

**Claim:** \( P_\theta \) and \( T_\theta \) uniquely determine the parameters \( \theta \).
Reduction to orthogonal case via whitening

\[ P_\theta = \sum_t w_t \vec{\mu}_t \otimes \vec{\mu}_t \] defines “whitened” coord. system.
Reduction to orthogonal case via whitening

\[ P_\theta = \sum_t w_t \mu_t \otimes \tilde{\mu}_t \] defines “whitened” coord. system.

Technical reduction:

Apply *change-of-basis* transformation \( P_\theta^{-1/2} \) to \( T_\theta \):

\[ T_\theta = \sum_{t=1}^k w_t \mu_t \otimes \mu_t \otimes \mu_t \quad \rightarrow \quad B_\theta = \sum_{t=1}^k \lambda_t \tilde{v}_t \otimes \tilde{v}_t \otimes \tilde{v}_t \]

where \( \lambda_t = 1/\sqrt{w_t}, \) \( \tilde{v}_t = P_\theta^{-1/2} (\sqrt{w_t} \mu_t). \)
Reduction to orthogonal case via whitening

$P_\theta = \sum_t w_t \mu_t \otimes \mu_t$ defines “whitened” coord. system.

**Technical reduction:**
Apply *change-of-basis* transformation $P_{\theta}^{-1/2}$ to $T_\theta$:

$$T_\theta = \sum_{t=1}^{k} w_t \mu_t \otimes \mu_t \otimes \mu_t \quad \longrightarrow \quad B_\theta = \sum_{t=1}^{k} \lambda_t \tilde{v}_t \otimes \tilde{v}_t \otimes \tilde{v}_t$$

where $\lambda_t = 1/\sqrt{w_t}$, $\tilde{v}_t = P_{\theta}^{-1/2} (\sqrt{w_t} \mu_t)$.

**Upshot:** $\{\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_k\}$ are orthonormal.
Reduction to orthogonal case via whitening

\[ P_\theta = \sum_t w_t \tilde{\mu}_t \otimes \tilde{\mu}_t \] defines “whitened” coord. system.

“Whitened” third-order moment tensor \( B_\theta \) has orthogonal decomposition

\[ B_\theta = \sum_{t=1}^{k} \lambda_t \tilde{v}_t \otimes \tilde{v}_t \otimes \tilde{v}_t. \]

(And \( \{ (\lambda_t, \tilde{v}_t) \} \) are related to parameters \( \{ (w_t, \tilde{\mu}_t) \} \).)

**Upshot:** \( \{ \tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_k \} \) are orthonormal.

**Claim:** Orthogonal decomposition of \( B_\theta \) is unique.
The spectral theorem and eigendecompositions
The spectral theorem and eigendecompositions

Any symmetric matrix

\[ A = \sum_{i=1}^{k} \lambda_i \vec{v}_i \otimes \vec{v}_i \]

Decomposition is unique only if all eigenvalues \( \lambda_i \) are distinct.
The spectral theorem and eigendecompositions

Any symmetric matrix

\[ A = \sum_{i=1}^{k} \lambda_i \vec{v}_i \otimes \vec{v}_i \]

Decomposition is unique only if all eigenvalues \( \lambda_i \) are distinct.
The spectral theorem and eigendecompositions

Any symmetric matrix

\[ A = \sum_{i=1}^{k} \lambda_i \, \vec{v}_i \otimes \vec{v}_i \]

Decomposition is unique only if all eigenvalues \( \lambda_i \) are distinct.
The spectral theorem and eigendecompositions

**Any symmetric matrix**

\[ A = \sum_{i=1}^{k} \lambda_i \vec{v}_i \otimes \vec{v}_i \]

Decomposition is unique only if all eigenvalues \( \lambda_i \) are distinct.

**Special 3rd-order tensor**

\[ B = \sum_{i=1}^{k} \lambda_i \vec{v}_i \otimes \vec{v}_i \otimes \vec{v}_i \]

If decomposition exists, then it’s always unique (even if \( \lambda_i \) all same).
The spectral theorem and eigendecompositions

Any symmetric matrix

\[ A = \sum_{i=1}^{k} \lambda_i \vec{v}_i \otimes \vec{v}_i \]

Decomposition is unique only if all eigenvalues \( \lambda_i \) are distinct.

Special 3rd-order tensor

\[ B = \sum_{i=1}^{k} \lambda_i \vec{v}_i \otimes \vec{v}_i \otimes \vec{v}_i \]

If decomposition exists, then it’s always unique (even if \( \lambda_i \) all same).
The spectral theorem and eigendecompositions

Any symmetric matrix

\[ A = \sum_{i=1}^{k} \lambda_i \vec{v}_i \otimes \vec{v}_i \]

Decomposition is unique only if all eigenvalues \( \lambda_i \) are distinct.

Special 3\(^{rd}\)-order tensor

\[ B = \sum_{i=1}^{k} \lambda_i \vec{v}_i \otimes \vec{v}_i \otimes \vec{v}_i \]

If decomposition exists, then it’s always unique (even if \( \lambda_i \) all same).
The spectral theorem and eigendecompositions

Any symmetric matrix

\[ A = \sum_{i=1}^{k} \lambda_i \vec{v}_i \otimes \vec{v}_i \]

Decomposition is unique only if all eigenvalues \( \lambda_i \) are distinct.

Special 3rd-order tensor

\[ B = \sum_{i=1}^{k} \lambda_i \vec{v}_i \otimes \vec{v}_i \otimes \vec{v}_i \]

If decomposition exists, then it’s always unique (even if \( \lambda_i \) all same).

Uniqueness of orthogonal decomposition (\(+\)low-rank structure) implies that \( P_{\theta} \) and \( T_{\theta} \) uniquely determine \( \theta \).
The spectral theorem and eigendecompositions

Any symmetric matrix

\[ A = \sum_{i=1}^{k} \lambda_i \vec{v}_i \otimes \vec{v}_i \]

Decomposition is unique only if all eigenvalues \( \lambda_i \) are distinct.

Special 3rd-order tensor

\[ B = \sum_{i=1}^{k} \lambda_i \vec{v}_i \otimes \vec{v}_i \otimes \vec{v}_i \]

If decomposition exists, then it’s always unique (even if \( \lambda_i \) all same).

Uniqueness of orthogonal decomposition (+low-rank structure) implies that \( P_\theta \) and \( T_\theta \) uniquely determine \( \theta \).
Q2. How to solve the moment equations?

Solve moment equations via optimization problem

\[
\min_{\theta} \| T\theta - \hat{T} \|_2 \quad \text{s.t.} \quad P\theta = \hat{P}.
\]

Not convex in parameters \( \theta = \{ (\vec{\mu}_i, w_i) \} \).

What we do: find one topic \( (\vec{\mu}_i, w_i) \) at a time, using local optimization on rank-1 approximation objective:

Can approximate all local optima, each corresponding to a topic.

\( \rightarrow \) Near-optimal solution to \((\dagger)\).
Q2. How to solve the moment equations?

Solve moment equations via optimization problem

$$\min_{\theta} \| T_\theta - \hat{T} \|^2 \quad \text{s.t.} \quad P_\theta = \hat{P}. \quad \quad \text{(†)}$$
Q2. How to solve the moment equations?

Solve moment equations via optimization problem

$$\min_\theta \| T_\theta - \hat{T} \|^2 \quad \text{s.t.} \quad P_\theta = \hat{P}. \quad \text{(†)}$$

Not convex in parameters $\theta = \{ (\vec{\mu}_i, w_i) \}$. 

What we do: find one topic $(\vec{\mu}_i, w_i)$ at a time, using local optimization on rank-1 approximation objective:

Can approximate all local optima, each corresponding to a topic.

Near-optimal solution to $(†)$. 
Q2. How to solve the moment equations?

Solve moment equations via optimization problem

\[
\min_{\theta} \| T_{\theta} - \hat{T} \|^2 \quad \text{s.t.} \quad P_{\theta} = \hat{P}. \tag{†}
\]

Not convex in parameters \( \theta = \{(\hat{\mu}_i, w_i)\} \).

What we do: find one topic \((\hat{\mu}_i, w_i)\) at a time, using local optimization on rank-1 approximation objective:

\[
\min_{\lambda, \vec{v}} \| \lambda \vec{v} \otimes \vec{v} \otimes \vec{v} - \hat{B} \|^2 \tag{‡}
\]

(after change-of-coord. system via \( \hat{P}: \hat{T} \rightarrow \hat{B} \)).
Q2. How to solve the moment equations?

Solve moment equations via optimization problem

\[
\min_{\theta} \| T_{\theta} - \hat{T} \|^2 \quad \text{s.t.} \quad P_{\theta} = \hat{P}. \quad (\dagger)
\]

Not convex in parameters \( \theta = \{ (\mu_i, w_i) \} \).

What we do: find one topic \((\mu_i, w_i)\) at a time, using local optimization on rank-1 approximation objective:

\[
\max_{\| \bar{u} \| \leq 1} \sum_{i,j,k} \hat{B}_{i,j,k} u_i u_j u_k \quad (\ddagger)
\]
Q2. How to solve the moment equations?

Solve moment equations via optimization problem

$$\min_{\theta} \| T_{\theta} - \hat{T} \|^2 \quad \text{s.t.} \quad P_{\theta} = \hat{P}. \quad (\dagger)$$

Not convex in parameters $\theta = \{(\vec{\mu}_i, w_i)\}$.

What we do: find one topic $(\vec{\mu}_i, w_i)$ at a time, using local optimization on rank-1 approximation objective:
Q2. How to solve the moment equations?

Solve moment equations via optimization problem

\[
\min_{\theta} \| T_\theta - \hat{T} \|^2 \quad \text{s.t.} \quad P_\theta = \hat{P}.
\]

(†)

Not convex in parameters \( \theta = \{(\vec{\mu}_i, w_i)\} \).

What we do: find one topic \((\vec{\mu}_i, w_i)\) at a time, using local optimization on rank-1 approximation objective:

\[
\vec{u}_1^* \vec{u}_1^* \ldots \vec{u}_k^* (\vec{\mu}_1^*, w_1^*) (\vec{\mu}_2^*, w_2^*) (\vec{\mu}_k^*, w_k^*)
\]

Can approximate all local optima, each corresp. to a topic.

\[\rightarrow\] Near-optimal solution to (†).
Variational argument

Interpret $P_\theta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and $T_\theta : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as bi-linear and tri-linear forms.
Variational argument

Interpret $P_\theta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and $T_\theta : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as bi-linear and tri-linear forms.

**Lemma**

Assuming $\{\vec{\mu}_i\}$ linearly independent and $w_i > 0$, each of the $k$ distinct, isolated local maximizers $\vec{u}^*$ of

$$\max_{\vec{u} \in \mathbb{R}^d} T_\theta(\vec{u}, \vec{u}, \vec{u}) \quad \text{s.t.} \quad P_\theta(\vec{u}, \vec{u}) \leq 1 \quad (\dagger)$$

satisfies, for some $i \in [k]$,

$$P_\theta \vec{u}^* = \sqrt{w_i} \vec{\mu}_i, \quad T_\theta(\vec{u}^*, \vec{u}^*, \vec{u}^*) = \frac{1}{\sqrt{w_i}}.$$
Variational argument

Interpret $P_\theta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and $T_\theta : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as bi-linear and tri-linear forms.

Lemma
Assuming $\{\vec{\mu}_i\}$ linearly independent and $w_i > 0$, each of the $k$ distinct, isolated local maximizers $\vec{u}^*$ of

$$\max_{\vec{u} \in \mathbb{R}^d} T_\theta(\vec{u}, \vec{u}, \vec{u}) \quad \text{s.t.} \quad P_\theta(\vec{u}, \vec{u}) \leq 1 \quad (\dagger)$$

satisfies, for some $i \in [k]$,

$$P_\theta \vec{u}^* = \sqrt{w_i} \vec{\mu}_i, \quad T_\theta(\vec{u}^*, \vec{u}^*, \vec{u}^*) = \frac{1}{\sqrt{w_i}}.$$ 

∴ $\{(\vec{\mu}_i, w_i) : i \in [k]\}$ uniquely determined by $P_\theta$ and $T_\theta$. 
Implementation of topic model estimator

**Potential deal-breakers:** Explicitly form $\hat{T}$, count word-triples $\rightarrow \Omega(d^3)$ space, $\Omega(length^3)$ time / doc.
Implementation of topic model estimator

Potential deal-breakers: Explicitly form $\hat{T}$, count word-triples $\longrightarrow \Omega(d^3)$ space, $\Omega(\text{length}^3)$ time / doc.

Can exploit algebraic structure to avoid bottlenecks.
Implementation of topic model estimator

**Potential deal-breakers:** Explicitly form $\hat{T}$, count word-triples $\rightarrow \Omega(d^3)$ space, $\Omega(\text{length}^3)$ time / doc.

Can **exploit algebraic structure** to avoid bottlenecks.

Implicit representation of $\hat{T}$:

$$\hat{T} \approx \frac{1}{|S|} \sum_{\vec{h} \in S} \vec{h} \otimes \vec{h} \otimes \vec{h}$$

where $\vec{h} \in \mathbb{N}^d$ is (sparse) histogram vector for a document.
Potential deal-breakers: Explicitly form $\hat{T}$, count word-triples $\rightarrow \Omega(d^3)$ space, $\Omega(\text{length}^3)$ time / doc.

Can exploit algebraic structure to avoid bottlenecks.

Computation of objective gradient at vector $\tilde{u} \in \mathbb{R}^d$:

$$\hat{T}(\tilde{u}) \approx \frac{1}{|S|} \sum_{\tilde{h} \in S} (\tilde{h} \otimes \tilde{h} \otimes \tilde{h})(\tilde{u}) = \frac{1}{|S|} \sum_{\tilde{h} \in S} (\tilde{h}^\top \tilde{u})^2 \tilde{h}$$

(sparse vector operations; time $= O(\text{input size})$).
Illustrative empirical results

- Vocabulary size: 102660 words.
- Set number of topics $k := 50$. 

Predictive performance of straightforward implementation:

\[ \approx 4 – 8 \times \text{speed-up over Gibbs sampling.} \]
Illustrative empirical results

- Vocabulary size: 102660 words.
- Set number of topics $k := 50$.

Predictive performance of straightforward implementation: ≈ 4–8× speed-up over Gibbs sampling.

![Graph showing Log loss vs. Training time (×10^4 sec)]
Illustrative empirical results

**Sample topics**: (showing top 10 words for each topic)

<table>
<thead>
<tr>
<th>Econ.</th>
<th>Baseball</th>
<th>Edu.</th>
<th>Health care</th>
<th>Golf</th>
</tr>
</thead>
<tbody>
<tr>
<td>sales</td>
<td>run</td>
<td>school</td>
<td>drug</td>
<td>player</td>
</tr>
<tr>
<td>economic</td>
<td>inning</td>
<td>student</td>
<td>patient</td>
<td>tiger_wood</td>
</tr>
<tr>
<td>consumer</td>
<td>hit</td>
<td>teacher</td>
<td>million</td>
<td>won</td>
</tr>
<tr>
<td>major</td>
<td>game</td>
<td>program</td>
<td>company</td>
<td>shot</td>
</tr>
<tr>
<td>home</td>
<td>season</td>
<td>official</td>
<td>doctor</td>
<td>play</td>
</tr>
<tr>
<td>indicator</td>
<td>home</td>
<td>public</td>
<td>companies</td>
<td>round</td>
</tr>
<tr>
<td>weekly</td>
<td>right</td>
<td>children</td>
<td>percent</td>
<td>win</td>
</tr>
<tr>
<td>order</td>
<td>games</td>
<td>high</td>
<td>cost</td>
<td>tournament</td>
</tr>
<tr>
<td>claim</td>
<td>dodger</td>
<td>education</td>
<td>program</td>
<td>tour</td>
</tr>
<tr>
<td>scheduled</td>
<td>left</td>
<td>district</td>
<td>health</td>
<td>right</td>
</tr>
</tbody>
</table>
Illustrative empirical results

Sample topics: (showing top 10 words for each topic)

<table>
<thead>
<tr>
<th>Invest.</th>
<th>Election</th>
<th>auto race</th>
<th>Child’s Lit.</th>
<th>Afghan War</th>
</tr>
</thead>
<tbody>
<tr>
<td>percent</td>
<td>al_gore</td>
<td>car</td>
<td>book</td>
<td>taliban</td>
</tr>
<tr>
<td>stock</td>
<td>campaign</td>
<td>race</td>
<td>children</td>
<td>attack</td>
</tr>
<tr>
<td>market</td>
<td>president</td>
<td>driver</td>
<td>ages</td>
<td>afghanistan</td>
</tr>
<tr>
<td>fund</td>
<td>george_bush</td>
<td>team</td>
<td>author</td>
<td>military</td>
</tr>
<tr>
<td>investor</td>
<td>bush</td>
<td>won</td>
<td>read</td>
<td>u_s</td>
</tr>
<tr>
<td>companies</td>
<td>clinton</td>
<td>win</td>
<td>newspaper</td>
<td>united_states</td>
</tr>
<tr>
<td>analyst</td>
<td>vice</td>
<td>racing</td>
<td>web</td>
<td>terrorist</td>
</tr>
<tr>
<td>money</td>
<td>presidential</td>
<td>track</td>
<td>writer</td>
<td>war</td>
</tr>
<tr>
<td>investment</td>
<td>million</td>
<td>season</td>
<td>written</td>
<td>bin</td>
</tr>
<tr>
<td>economy</td>
<td>democratic</td>
<td>lap</td>
<td>sales</td>
<td></td>
</tr>
</tbody>
</table>
### Sample topics: (showing top 10 words for each topic)

<table>
<thead>
<tr>
<th>Web</th>
<th>Antitrust</th>
<th>TV</th>
<th>Movies</th>
<th>Music</th>
</tr>
</thead>
<tbody>
<tr>
<td>com</td>
<td>court</td>
<td>show</td>
<td>film</td>
<td>music</td>
</tr>
<tr>
<td>www</td>
<td>case</td>
<td>network</td>
<td>movie</td>
<td>song</td>
</tr>
<tr>
<td>site</td>
<td>law</td>
<td>season</td>
<td>director</td>
<td>group</td>
</tr>
<tr>
<td>web</td>
<td>lawyer</td>
<td>nbc</td>
<td>play</td>
<td>part</td>
</tr>
<tr>
<td>sites</td>
<td>federal</td>
<td>cb</td>
<td>character</td>
<td>new_york</td>
</tr>
<tr>
<td>information</td>
<td>government</td>
<td>program</td>
<td>actor</td>
<td>company</td>
</tr>
<tr>
<td>online</td>
<td>decision</td>
<td>television</td>
<td>show</td>
<td>million</td>
</tr>
<tr>
<td>mail</td>
<td>trial</td>
<td>series</td>
<td>movies</td>
<td>band</td>
</tr>
<tr>
<td>internet</td>
<td>microsoft</td>
<td>night</td>
<td>million</td>
<td>show</td>
</tr>
<tr>
<td>telegram</td>
<td>right</td>
<td>new_york</td>
<td>part</td>
<td>album</td>
</tr>
</tbody>
</table>

*etc.*
Efficient learning algorithms for topic models, based on solving moment equations

\[ \text{moments}_\theta = \text{moments}_S. \]
**Recap**

**Efficient learning algorithms** for topic models, based on solving moment equations

\[ \text{moments}_\theta = \text{moments}_S. \]

**Q1.** Which moments should we use? Suffices to use low-order (up to 3\textsuperscript{rd}-order) moments, and exploit multivariate structure in high-dimensions.
Recap

**Efficient learning algorithms** for topic models, based on solving moment equations

\[ \text{moments}_\theta = \text{moments}_S. \]

**Q1.** Which moments should we use?

Suffices to use low-order (up to 3\textsuperscript{rd}-order) moments, and exploit multivariate structure in high-dimensions.

**Q2.** How do we (approx.) solve these moment equations?

Local optimization based on orthogonal tensor decompositions.
“Eigen-structure” found in low-order moments for many other models of high-dimensional data

\[ \sum_{i=1}^{k} \lambda_i \vec{v}_i \otimes \vec{v}_i \otimes \vec{v}_i \]
Latent Dirichlet Allocation (Blei-Ng-Jordan, ’02) topic model:

- $k$ topics (distributions over $d$ words).
- Each document $\leftrightarrow$ mixture of topics.
- Doc.’s mixing weights $\sim$ Dirichlet($\vec{\alpha}$).
- Words in doc. $\sim_{iid}$ mixture dist.
Latent Dirichlet Allocation (Blei-Ng-Jordan, ’02) topic model:

- $k$ topics (distributions over $d$ words).
- Each document $\leftrightarrow$ mixture of topics.
- Doc.’s mixing weights $\sim$ Dirichlet($\vec{\alpha}$).
- Words in doc. $\sim_{iid}$ mixture dist.

Mixtures of Gaussians (Pearson, 1894)

- $k$ sub-populations in $\mathbb{R}^d$;
- $t$-th sub-pop. modeled as Gaussian $\mathcal{N}(\vec{\mu}_t, \Sigma_t)$
- with mixing weight $w_t$. 
Finding the relevant eigenstructure

In both LDA and mixtures of axis-aligned Gaussians:

\[ f \left( \leq 2^{\text{nd}}\text{-order moments}_{\theta} \right) = \sum w_t \vec{\mu}_t \otimes \vec{\mu}_t \]

\[ g \left( \leq 3^{\text{rd}}\text{-order moments}_{\theta} \right) = \sum w_t \vec{\mu}_t \otimes \vec{\mu}_t \otimes \vec{\mu}_t \]

for suitable \( f \) and \( g \) based on additional model structure.
Hidden Markov Models (HMMs)

Workhorse statistical model for sequence data

/k/    /a/    /t/

\[ h_1 \rightarrow h_2 \rightarrow \ldots \rightarrow h_\ell \]

\[ \vec{x}_1 \rightarrow \vec{x}_2 \rightarrow \vec{x}_\ell \]
Hidden Markov Models (HMMs)

Workhorse statistical model for sequence data

Hidden state variables $h_1 \rightarrow h_2 \rightarrow \cdots$ form a Markov chain.

Observation $\vec{x}_t$ at time $t$ depends only on hidden state $h_t$ at time $t$. 
Learning HMMs

Correlations between past, present, and future

$h_{t-1} \rightarrow h_t \rightarrow h_{t+1}$

$\tilde{x}_{t-1} \rightarrow \tilde{x}_t \rightarrow \tilde{x}_{t+1}$
Learning HMMs

Correlations between past, present, and future

\[ h_{t-1} \rightarrow h_t \rightarrow h_{t+1} \]

\[ \tilde{x}_{t-1} \rightarrow \tilde{x}_t \rightarrow \tilde{x}_{t+1} \]

Suffices to use **low-order (asymmetric) cross moments**

\[ \mathbb{E}_\theta [ \tilde{x}_{t-1} \otimes \tilde{x}_t \otimes \tilde{x}_{t+1} ] \]
Where to read more

Tensor decompositions for learning latent variable models
A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, M. Telgarsky

http://jmlr.org/papers/v15/anandkumar14b.html