Efficient algorithms for estimating multi-view mixture models

Daniel Hsu

Microsoft Research, New England
Outline

Multi-view mixture models

Multi-view method-of-moments

Some applications and open questions

Concluding remarks
Part 1. Multi-view mixture models

Multi-view mixture models
Unsupervised learning and mixture models
Multi-view mixture models
Complexity barriers

Multi-view method-of-moments

Some applications and open questions

Concluding remarks
Unsupervised learning

- Many modern applications of machine learning:
  - high-dimensional data from many diverse sources,
  - but mostly unlabeled.
Unsupervised learning

- Many modern applications of machine learning:
  - high-dimensional data from many diverse sources,
  - but mostly unlabeled.

- Unsupervised learning: extract useful info from this data.
  - Disentangle sub-populations in data source.
  - Discover useful representations for downstream stages of learning pipeline (e.g., supervised learning).
Mixture models

Simple latent variable model: mixture model

\[ h \in [k] := \{1, 2, \ldots, k\} \text{ (hidden);} \]
\[ \vec{x} \in \mathbb{R}^d \text{ (observed);} \]
\[ \Pr[h = j] = w_j; \quad \vec{x} \mid h \sim \mathbb{P}_h; \]

so \( \vec{x} \) has a mixture distribution

\[ \mathbb{P}(\vec{x}) = w_1 \mathbb{P}_1(\vec{x}) + w_2 \mathbb{P}_2(\vec{x}) + \cdots + w_k \mathbb{P}_k(\vec{x}). \]
Mixture models

Simple latent variable model: mixture model

\[ h \in [k] := \{1, 2, \ldots, k\} \text{ (hidden)}; \]
\[ \vec{x} \in \mathbb{R}^d \text{ (observed)}; \]
\[ \Pr[ h = j ] = w_j; \quad \vec{x} \mid h \sim \mathbb{P}_h; \]

so \( \vec{x} \) has a mixture distribution

\[ \mathbb{P}(\vec{x}) = w_1 \mathbb{P}_1(\vec{x}) + w_2 \mathbb{P}_2(\vec{x}) + \cdots + w_k \mathbb{P}_k(\vec{x}). \]

**Typical use**: learn about constituent sub-populations (e.g., clusters) in data source.
Multi-view mixture models

Can we take advantage of diverse sources of information?
Multi-view mixture models

Can we take advantage of diverse sources of information?

$h \in [k], \quad \vec{x}_i \in \mathbb{R}^{d_i}, \quad \ell = \# \text{ views} (e.g., \text{audio, video, text}).$

$k = \# \text{ components}, \quad \ell = \# \text{ views}.$

View 1: $\vec{x}_1 \in \mathbb{R}^{d_1}$
View 2: $\vec{x}_2 \in \mathbb{R}^{d_2}$
View 3: $\vec{x}_3 \in \mathbb{R}^{d_3}$
Can we take advantage of diverse sources of information?

Can we take advantage of diverse sources of information?

$\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_\ell 

h \in [k], 

\vec{x}_1 \in \mathbb{R}^{d_1}, \vec{x}_2 \in \mathbb{R}^{d_2}, \ldots, \vec{x}_\ell \in \mathbb{R}^{d_\ell}.

k = \# \text{ components}, \ \ell = \# \text{ views (e.g., audio, video, text)}.$
Multi-view mixture models

**Multi-view assumption:**
Views are conditionally independent given the component.

View 1: $\tilde{x}_1 \in \mathbb{R}^{d_1}$  
View 2: $\tilde{x}_2 \in \mathbb{R}^{d_2}$  
View 3: $\tilde{x}_3 \in \mathbb{R}^{d_3}$

Larger $k$ (# components): more sub-populations to disentangle.  
Larger $\ell$ (# views): more non-redundant sources of information.
Semi-parametric estimation task

“Parameters” of component distributions:

Mixing weights $w_j := \Pr[h = j], \quad j \in [k];$

Conditional means $\mu_{v, j} := \mathbb{E}[\tilde{x}_v | h = j] \in \mathbb{R}^{d_v}, \quad j \in [k], \; v \in [\ell].$

**Goal:** Estimate mixing weights and conditional means from independent copies of $(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_\ell).$
Semi-parametric estimation task

“Parameters” of component distributions:

Mixing weights $w_j := \Pr[h = j], \quad j \in [k];$
Conditional means $\bar{\mu}_{v,j} := \mathbb{E}[\bar{x}_v | h = j] \in \mathbb{R}^{dv}, \quad j \in [k], v \in [\ell].$

**Goal:** Estimate mixing weights and conditional means from independent copies of $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_\ell).$

**Questions:**

1. How do we estimate $\{w_j\}$ and $\{\bar{\mu}_{v,j}\}$ without observing $h$?
2. How many views $\ell$ are sufficient to learn with poly$(k)$ computational / sample complexity?
Some barriers to efficient estimation

**Challenge**: many difficult parametric estimation tasks reduce to this estimation problem.
Some barriers to efficient estimation

**Challenge**: many difficult parametric estimation tasks reduce to this estimation problem.

**Cryptographic barrier**: discrete HMM parameter estimation as hard as *learning parity functions with noise* (Mossel-Roch, ’06).
Some barriers to efficient estimation

**Challenge:** many difficult parametric estimation tasks reduce to this estimation problem.

**Cryptographic barrier:** discrete HMM parameter estimation as hard as learning parity functions with noise (Mossel-Roch, ’06).

**Statistical barrier:** Gaussian mixtures in $\mathbb{R}^1$ can require $\exp(\Omega(k))$ samples to estimate parameters, even if components are well-separated (Moitra-Valiant, ’10).
Some barriers to efficient estimation

**Challenge**: many difficult parametric estimation tasks reduce to this estimation problem.

**Cryptographic barrier**: discrete HMM parameter estimation as hard as learning parity functions with noise (Mossel-Roch, ’06).

**Statistical barrier**: Gaussian mixtures in $\mathbb{R}^1$ can require $\exp(\Omega(k))$ samples to estimate parameters, even if components are well-separated (Moitra-Valiant, ’10).

**In practice**: resort to local search (e.g., EM), often subject to slow convergence and inaccurate local optima.
Making progress: Gaussian mixture model

**Gaussian mixture model:** problem becomes easier if assume some **large minimum separation** between component means (Dasgupta, ’99):

\[
\text{sep} := \min_{i \neq j} \frac{\|\mu_i - \mu_j\|}{\max\{\sigma_i, \sigma_j\}}.
\]

▶ \(\text{sep} = \Omega(d_c):\) interpoint distance-based methods / EM (Dasgupta, ’99; Dasgupta-Schulman, ’00; Arora-Kannan, ’00)

▶ \(\text{sep} = \Omega(k_c):\) first use PCA to \(k\) dimensions (Vempala-Wang, ’02; Kannan-Salmasian-Vempala, ’05; Achlioptas-McSherry, ’05)

Also works for mixtures of log-concave distributions.

▶ No minimum separation requirement: method-of-moments but \(\exp(\Omega(k))\) running time / sample size (Kalai-Moitra-Valiant, ’10; Belkin-Sinha, ’10; Moitra-Valiant, ’10)
Making progress: Gaussian mixture model

Gaussian mixture model: problem becomes easier if assume some large minimum separation between component means (Dasgupta, ’99):

\[ \text{sep} := \min_{i \neq j} \frac{\| \vec{\mu}_i - \vec{\mu}_j \|}{\max\{\sigma_i, \sigma_j\}}. \]

- \( \text{sep} = \Omega(d^c) \): interpoint distance-based methods / EM (Dasgupta, ’99; Dasgupta-Schulman, ’00; Arora-Kannan, ’00)
  - \( \text{sep} = \Omega(k^c) \): first use PCA to \( k \) dimensions (Vempala-Wang, ’02; Kannan-Salmasian-Vempala, ’05; Achlioptas-McSherry, ’05)
  - Also works for mixtures of log-concave distributions.
Making progress: Gaussian mixture model

**Gaussian mixture model**: problem becomes easier if assume some **large minimum separation** between component means (Dasgupta, ’99):

\[
\text{sep} := \min_{i \neq j} \frac{\|\bar{\mu}_i - \bar{\mu}_j\|}{\max\{\sigma_i, \sigma_j\}}.
\]

- **sep = \Omega(d^c)**: interpoint distance-based methods / EM (Dasgupta, ’99; Dasgupta-Schulman, ’00; Arora-Kannan, ’00)
  - **sep = \Omega(k^c)**: first use PCA to \(k\) dimensions (Vempala-Wang, ’02; Kannan-Salmasian-Vempala, ’05; Achlioptas-McSherry, ’05)
  - Also works for mixtures of log-concave distributions.
- **No minimum separation requirement**: method-of-moments but \(\exp(\Omega(k))\) running time / sample size (Kalai-Moitra-Valiant, ’10; Belkin-Sinha, ’10; Moitra-Valiant, ’10)
Making progress: discrete hidden Markov models

Hardness reductions create HMMs with degenerate output and next-state distributions.

\[
\Pr[\bar{x}_t = \cdot | h_t = 1] 
\]

\[
0.6 \Pr[\bar{x}_t = \cdot | h_t = 2] + 0.4 \Pr[\bar{x}_t = \cdot | h_t = 3]
\]
Making progress: discrete hidden Markov models

Hardness reductions create HMMs with degenerate output and next-state distributions.

\[
\Pr[\bar{x}_t = \cdot | h_t = 1] 
\approx 0.6 + 0.4 
\]

These instances are avoided by assuming parameter matrices are full-rank (Mossel-Roch, ’06; Hsu-Kakade-Zhang, ’09)
What we do

**This work:** given $\geq 3$ views, mild non-degeneracy conditions imply efficient algorithms for estimation.
What we do

**This work:** given $\geq 3$ views, mild non-degeneracy conditions imply efficient algorithms for estimation.

- **Non-degeneracy condition** for multi-view mixture model: Conditional means $\{\vec{\mu}_{v,1}, \vec{\mu}_{v,2}, \ldots, \vec{\mu}_{v,k}\}$ are linearly independent for each view $v \in [\ell]$, and $\vec{w} > \vec{0}$.

  Requires high-dimensional observations ($d_v \geq k$)!
This work: given $\geq 3$ views, mild non-degeneracy conditions imply efficient algorithms for estimation.

- **Non-degeneracy condition** for multi-view mixture model: Conditional means $\{\vec{\mu}_v, 1, \vec{\mu}_v, 2, \ldots, \vec{\mu}_v, k\}$ are linearly independent for each view $v \in [\ell]$, and $\vec{w} > \vec{0}$.
  
  Requires high-dimensional observations ($d_v \geq k$)!

- **New efficient learning guarantees** for parametric models (e.g., mixtures of Gaussians, general HMMs)
This work: given $\geq 3$ views, mild non-degeneracy conditions imply efficient algorithms for estimation.

- **Non-degeneracy condition** for multi-view mixture model: Conditional means $\{\vec{\mu}_v, 1, \vec{\mu}_v, 2, \ldots, \vec{\mu}_v, k\}$ are linearly independent for each view $v \in [\ell]$, and $\vec{w} > \vec{0}$.

  Requires high-dimensional observations ($d_v \geq k$)!

- **New efficient learning guarantees** for parametric models (e.g., mixtures of Gaussians, general HMMs)

- **General tensor decomposition framework** applicable to a wide variety of estimation problems.

Multi-view mixture models

Multi-view method-of-moments
  Overview
  Structure of moments
  Uniqueness of decomposition
  Computing the decomposition
  Asymmetric views

Some applications and open questions

Concluding remarks
The plan

- First, assume views are \textit{(conditionally) exchangeable}, and derive basic algorithm.
The plan

- First, assume views are \textit{(conditionally) exchangeable}, and derive basic algorithm.

- Then, provide \textit{reduction} from general multi-view setting to exchangeable case.
Simpler case: exchangeable views

*(Conditionally) exchangeable views*: assume the views have the same conditional means, *i.e.*, 

\[
\mathbb{E}[ \mathbf{x}_v | h = j ] \equiv \bar{\mu}_j, \quad j \in [k], v \in [\ell].
\]
Simpler case: exchangeable views

(Conditionally) exchangeable views: assume the views have the same conditional means, i.e.,

$$\mathbb{E}[\tilde{x}_v|h = j] \equiv \mu_j, \quad j \in [k], v \in [\ell].$$

Motivating setting: bag-of-words model, \(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_\ell \equiv \ell\) exchangeable words in a document.

One-hot encoding:
\(\tilde{x}_v = \tilde{e}_i \iff v\)-th word in document is \(i\)-th word in vocab

(where \(\tilde{e}_i \in \{0, 1\}^d\) has 1 in \(i\)-th position, 0 elsewhere).

\((\mu_j)_i = \mathbb{E}[(\tilde{x}_v)_i|h = j] = \Pr[\tilde{x}_v = \tilde{e}_i|h = j], \quad i \in [d], j \in [k].\)
Key ideas

1. **Method-of-moments**: conditional means are revealed by appropriate low-rank decompositions of moment matrices and tensors.

2. **Third-order tensor decomposition** is uniquely determined by directions of (locally) maximum skew.

3. The required **local optimization** can be efficiently performed in poly time.
Algebraic structure in moments

Recall: \( \mathbb{E}[ \tilde{x}_v | h = j ] = \tilde{\mu}_j \).
Algebraic structure in moments

Recall: 
\[ \mathbb{E}[\vec{x}_v | h = j] = \vec{\mu}_j. \]

By conditional independence and exchangeability of \( \vec{x}_1, \vec{x}_2, \ldots, \vec{x}_\ell \) given \( h \),

Pairs \( := \mathbb{E}[\vec{x}_1 \otimes \vec{x}_2] \)

\[ = \mathbb{E}\left[\mathbb{E}[\vec{x}_1 | h] \otimes \mathbb{E}[\vec{x}_2 | h]\right] = \mathbb{E}[\vec{\mu}_h \otimes \vec{\mu}_h] \]

\[ = \sum_{i=1}^{k} w_i \vec{\mu}_i \otimes \vec{\mu}_i \in \mathbb{R}^{d \times d}. \]
Algebraic structure in moments

Recall: \( E[\tilde{x}_v | h = j] = \tilde{\mu}_j \).

By conditional independence and exchangeability of \( \tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_\ell \) given \( h \),

Pairs := \( E[\tilde{x}_1 \otimes \tilde{x}_2] \)
\[ = \mathbb{E}[E[\tilde{x}_1 | h] \otimes E[\tilde{x}_2 | h]] = \mathbb{E}[\tilde{\mu}_h \otimes \tilde{\mu}_h] \]
\[ = \sum_{i=1}^{k} w_i \tilde{\mu}_i \otimes \tilde{\mu}_i \in \mathbb{R}^{d \times d}. \]

Triples := \( E[\tilde{x}_1 \otimes \tilde{x}_2 \otimes \tilde{x}_3] \)
\[ = \sum_{i=1}^{k} w_i \tilde{\mu}_i \otimes \tilde{\mu}_i \otimes \tilde{\mu}_i \in \mathbb{R}^{d \times d \times d}, \text{ etc.} \]

(If only we could extract these “low-rank” decompositions . . . )
2nd moment: subspace spanned by conditional means

Non-degeneracy assumption (\{\vec{\mu}_i\}_{i=1} linearly independent) \Rightarrow Pairs = \sum_{i=1}^k w_i \vec{\mu}_i \otimes \vec{\mu}_i symmetric psd and rank k \Rightarrow Pairs equips k-dim subspace span \{\vec{\mu}_1, \vec{\mu}_2, ..., \vec{\mu}_k\} with inner product Pairs(\vec{x}, \vec{y}) := \vec{x}^\top Pairs \vec{y}.

However, \{\vec{\mu}_i\}_{i} not generally determined by just Pairs (e.g., \{\vec{\mu}_i\}_{i} are not necessarily orthogonal). Must look at higher-order moments?
2nd moment: subspace spanned by conditional means

**Non-degeneracy assumption** \( \{ \bar{\mu}_i \} \) linearly independent

\[
\begin{align*}
\text{Pairs} = \sum_{i=1}^{k} w_i \bar{\mu}_i \otimes \bar{\mu}_i \\
\text{symmetric psd and rank } k = \Rightarrow \text{Pairs equips } k \text{-dim subspace span } \{ \bar{\mu}_1, \bar{\mu}_2, \ldots, \bar{\mu}_k \} \\
\text{with inner product } \text{Pairs}(\bar{x}, \bar{y}) := \bar{x}^\top \text{Pairs} \bar{y}.
\end{align*}
\]

However, \( \{ \bar{\mu}_i \} \) not generally determined by just \( \text{Pairs}(e.g., \{ \bar{\mu}_i \} \text{are not necessarily orthogonal}) \). Must look at higher-order moments?
2nd moment: subspace spanned by conditional means

**Non-degeneracy assumption** ($\{\bar{\mu}_i\}$ linearly independent)

$\implies$ Pairs $= \sum_{i=1}^{k} w_i \bar{\mu}_i \otimes \bar{\mu}_i$ symmetric psd and rank $k$
Non-degeneracy assumption \( \{\vec{\mu}_i\} \) linearly independent

\[ \implies \text{Pairs} = \sum_{i=1}^{k} w_i \vec{\mu}_i \otimes \vec{\mu}_i \quad \text{symmetric psd and rank } k \]

\[ \implies \text{Pairs} \text{ equips } k\text{-dim subspace span}\{\vec{\mu}_1, \vec{\mu}_2, \ldots, \vec{\mu}_k\} \text{ with inner product} \]

\[ \text{Pairs}(\vec{x}, \vec{y}) := \vec{x}^\top \text{Pairs} \vec{y}. \]
2nd moment: subspace spanned by conditional means

**Non-degeneracy assumption** \( \{\vec{\mu}_i\} \) linearly independent

\[
\implies \text{Pairs} = \sum_{i=1}^{k} w_i \vec{\mu}_i \otimes \vec{\mu}_i \quad \text{symmetric psd and rank } k
\]

\[
\implies \text{Pairs} \text{ equips } k\text{-dim subspace span}\{\vec{\mu}_1, \vec{\mu}_2, \ldots, \vec{\mu}_k\} \text{ with inner product}
\]

\[
\text{Pairs}(\vec{x}, \vec{y}) := \vec{x}^\top \text{Pairs} \vec{y}.
\]

**However**, \( \{\vec{\mu}_i\} \) not generally determined by just **Pairs**
(e.g., \( \{\vec{\mu}_i\} \) are not necessarily orthogonal).
2nd moment: subspace spanned by conditional means

**Non-degeneracy assumption** ($\{\vec{\mu}_i\}$ linearly independent)

$\Longrightarrow$ Pairs $= \sum_{i=1}^{k} w_i \vec{\mu}_i \otimes \vec{\mu}_i$ symmetric psd and rank $k$

$\Longrightarrow$ Pairs equips $k$-dim subspace span$\{\vec{\mu}_1, \vec{\mu}_2, \ldots, \vec{\mu}_k\}$ with inner product

$$
\text{Pairs}(\vec{x}, \vec{y}) := \vec{x}^\top \text{Pairs} \vec{y}.
$$

However, $\{\vec{\mu}_i\}$ not generally determined by just Pairs (e.g., $\{\vec{\mu}_i\}$ are not necessarily orthogonal).

Must look at higher-order moments?
Claim: **Up to third-moment (i.e., 3 views) suffices.**

View **Triples**: $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ as trilinear form.
Claim: **Up to third-moment (i.e., 3 views) suffices.**

View **Triples**: $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ as trilinear form.

<table>
<thead>
<tr>
<th>Theorem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Each isolated local maximizer $\vec{\eta}^*$ of</td>
</tr>
<tr>
<td>$\max_{\vec{\eta} \in \mathbb{R}^d} \text{Triples}(\vec{\eta}, \vec{\eta}, \vec{\eta})$ s.t. $\text{Pairs}(\vec{\eta}, \vec{\eta}) = 1$</td>
</tr>
<tr>
<td>satisfies, for some $i \in [k]$,</td>
</tr>
<tr>
<td>$\text{Pairs}(\vec{\eta}^<em>) = \sqrt{w_i} \bar{\mu}_i$, $\text{Triples}(\vec{\eta}^</em>, \vec{\eta}^<em>, \vec{\eta}^</em>) = \frac{1}{\sqrt{w_i}}$.</td>
</tr>
</tbody>
</table>
3rd moment: (cross) skew maximizers

Claim: **Up to third-moment (i.e., 3 views) suffices.**
View **Triples**: $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as trilinear form.

**Theorem**
*Each isolated local maximizer $\vec{\eta}^*$ of*

$$\max_{\vec{\eta} \in \mathbb{R}^d} \text{Triples}(\vec{\eta}, \vec{\eta}, \vec{\eta}) \text{ s.t. } \text{Pairs}(\vec{\eta}, \vec{\eta}) = 1$$

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

$$\text{Pairs} \, \vec{\eta}^* = \sqrt{w_i} \, \mu_i, \quad \text{Triples}(\vec{\eta}^*, \vec{\eta}^*, \vec{\eta}^*) = \frac{1}{\sqrt{w_i}}.$$

Also: these maximizers can be found **efficiently** and **robustly**.
Variational analysis

\[
\max_{\tilde{\eta} \in \mathbb{R}^d} \text{Tripples}(\tilde{\eta}, \tilde{\eta}, \tilde{\eta}) \quad \text{s.t.} \quad \text{Pairs}(\tilde{\eta}, \tilde{\eta}) = 1
\]
Variational analysis

\[
\max_{\tilde{\eta} \in \mathbb{R}^d} \text{Triples}(\tilde{\eta}, \tilde{\eta}, \tilde{\eta}) \quad \text{s.t.} \quad \text{Pairs}(\tilde{\eta}, \tilde{\eta}) = 1
\]

(Substitute \( \text{Pairs} = \sum_{i=1}^{k} w_i \tilde{\mu}_i \otimes \tilde{\mu}_i \) and \( \text{Triples} = \sum_{i=1}^{k} w_i \tilde{\mu}_i \otimes \tilde{\mu}_i \otimes \tilde{\mu}_i \))
Variational analysis

\[
\max_{\vec{\eta} \in \mathbb{R}^d} \sum_{i=1}^{k} w_i (\vec{\eta}^\top \vec{\mu}_i)^3 \quad \text{s.t.} \quad \sum_{i=1}^{k} w_i (\vec{\eta}^\top \vec{\mu}_i)^2 = 1
\]

(Substitute Pairs = \(\sum_{i=1}^{k} w_i \vec{\mu}_i \otimes \vec{\mu}_i\) and Triples = \(\sum_{i=1}^{k} w_i \vec{\mu}_i \otimes \vec{\mu}_i \otimes \vec{\mu}_i\).)
Variational analysis

\[
\max_{\vec{\eta} \in \mathbb{R}^d} \sum_{i=1}^k w_i (\vec{\eta}^\top \vec{\mu}_i)^3 \quad \text{s.t.} \quad \sum_{i=1}^k w_i (\vec{\eta}^\top \vec{\mu}_i)^2 = 1
\]

(Let \(\theta_i := \sqrt{w_i (\vec{\eta}^\top \vec{\mu}_i)}\) for \(i \in [k]\).)
Variational analysis

\[
\max_{\tilde{\eta} \in \mathbb{R}^d} \sum_{i=1}^{k} \frac{1}{\sqrt{w_i}} (\sqrt{w_i} \tilde{\eta}^\top \tilde{\mu}_i)^3 \quad \text{s.t.} \quad \sum_{i=1}^{k} (\sqrt{w_i} \tilde{\eta}^\top \tilde{\mu}_i)^2 = 1
\]

(Let \( \theta_i := \sqrt{w_i} (\tilde{\eta}^\top \tilde{\mu}_i) \) for \( i \in [k] \).)
Variational analysis

\[
\max_{\tilde{\theta} \in \mathbb{R}^k} \sum_{i=1}^{k} \frac{1}{\sqrt{w_i}} \theta_i^3 \quad \text{s.t.} \quad \sum_{i=1}^{k} \theta_i^2 = 1
\]

(Let \( \theta_i := \sqrt{w_i} (\eta_i^\top \bar{\mu}_i) \) for \( i \in [k] \).)
Variational analysis

\[
\max_{\vec{\theta} \in \mathbb{R}^k} \sum_{i=1}^{k} \frac{1}{\sqrt{w_i}} \theta_i^3 \quad \text{s.t.} \quad \sum_{i=1}^{k} \theta_i^2 = 1
\]

(Let \( \theta_i := \sqrt{w_i} (\vec{\eta}^* \vec{\mu}_i) \) for \( i \in [k] \).)

**Isolated local maximizers** \( \vec{\theta}^* \) (found via gradient ascent) are

\( \vec{e}_1 = (1, 0, 0, \ldots) \), \( \vec{e}_2 = (0, 1, 0, \ldots) \), etc.

which means that each \( \vec{\eta}^* \) satisfies, for some \( i \in [k] \),

\[
\sqrt{w_j} (\vec{\eta}^* \vec{\mu}_j) = \begin{cases} 
1 & j = i \\
0 & j \neq i.
\end{cases}
\]
Variational analysis

\[
\max_{\vec{\theta} \in \mathbb{R}^k} \sum_{i=1}^{k} \frac{1}{\sqrt{w_i}} \theta_i^3 \quad \text{s.t.} \quad \sum_{i=1}^{k} \theta_i^2 = 1
\]

(Let \( \theta_i := \sqrt{w_i} (\vec{\eta}^\top \vec{\mu}_i) \) for \( i \in [k] \).)

**Isolated local maximizers** \( \vec{\theta}^* \) (found via gradient ascent) are

\[
\vec{e}_1 = (1, 0, 0, \ldots), \quad \vec{e}_2 = (0, 1, 0, \ldots), \quad \text{etc.}
\]

which means that each \( \vec{\eta}^* \) satisfies, for some \( i \in [k] \),

\[
\sqrt{w_j} (\vec{\eta}^* \top \vec{\mu}_j) = \begin{cases} 1 & j = i \\ 0 & j \neq i. \end{cases}
\]

Therefore

\[
\text{Pairs } \vec{\eta}^* = \sum_{j=1}^{k} w_j \vec{\mu}_j (\vec{\eta}^* \top \vec{\mu}_j) = \sqrt{w_i} \vec{\mu}_i.
\]
Extracting all isolated local maximizers

1. Start with $T := \text{Triples}$. 

Goto step 2.

A variant of this runs in polynomial time (w.h.p.), and is robust to perturbations to $\text{Pairs}$ and $\text{Triples}$. 
1. Start with $T := \text{Triples}$.

2. Find isolated local maximizer of

$$T(\vec{\eta}, \vec{\eta}, \vec{\eta}) \text{ s.t. } \text{Pairs}(\vec{\eta}, \vec{\eta}) = 1$$

via gradient ascent from random $\vec{\eta} \in \text{range} \text{(Pairs)}$.

Say maximum is $\lambda^*$ and maximizer is $\vec{\eta}^*$.
Extracting all isolated local maximizers

1. Start with $T := \text{Triples}$.

2. Find isolated local maximizer of

$$T(\vec{\eta}, \vec{\eta}, \vec{\eta}) \text{ s.t. } \text{Pairs}(\vec{\eta}, \vec{\eta}) = 1$$

via gradient ascent from random $\vec{\eta} \in \text{range(Pairs)}$.

Say maximum is $\lambda^*$ and maximizer is $\vec{\eta}^*$.

3. Deflation: replace $T$ with $T - \lambda^* \vec{\eta}^* \otimes \vec{\eta}^* \otimes \vec{\eta}^*$.

Goto step 2.
Extracting all isolated local maximizers

1. Start with $T := \text{Triples}$.
2. Find isolated local maximizer of
   
   $$T(\vec{\eta}, \vec{\eta}, \vec{\eta}) \text{ s.t. } \text{Pairs}(\vec{\eta}, \vec{\eta}) = 1$$

   via gradient ascent from random $\vec{\eta} \in \text{range}(\text{Pairs})$.
   Say maximum is $\lambda^*$ and maximizer is $\vec{\eta}^*$.
3. Deflation: replace $T$ with $T - \lambda^* \vec{\eta}^* \otimes \vec{\eta}^* \otimes \vec{\eta}^*$.  
   Goto step 2.

A variant of this runs in polynomial time (w.h.p.), and is robust to perturbations to Pairs and Triples.
General case: asymmetric views

Each view $\nu$ has different set of conditional means
\[ \{ \vec{\mu}_{\nu,1}, \vec{\mu}_{\nu,2}, \ldots, \vec{\mu}_{\nu,k} \} \subset \mathbb{R}^{d_{\nu}}. \]
General case: asymmetric views

Each view \( v \) has different set of conditional means 
\[ \{ \bar{\mu}_{v,1}, \bar{\mu}_{v,2}, \ldots, \bar{\mu}_{v,k} \} \subset \mathbb{R}^{d_v}. \]

Reduction: transform \( \tilde{x}_1 \) and \( \tilde{x}_2 \) to “look like” \( \tilde{x}_3 \) via linear transformations.
Asymmetric cross moments

Define asymmetric cross moment:

$$\text{Pairs}_{u,v} := \mathbb{E}[\vec{x}_u \otimes \vec{x}_v].$$
Asymmetric cross moments

Define asymmetric cross moment:

\[ \text{Pairs}_{u,v} := \mathbb{E}[\tilde{x}_u \otimes \tilde{x}_v]. \]

Transforming view \( v \) to view 3:

\[ C_{v \rightarrow 3} := \mathbb{E}[\tilde{x}_3 \otimes \tilde{x}_u] \mathbb{E}[\tilde{x}_v \otimes \tilde{x}_u]^\dagger \in \mathbb{R}^{d_3 \times d_v} \]

where \(^\dagger\) denotes Moore-Penrose pseudoinverse.
Asymmetric cross moments

Define asymmetric cross moment:

\[
Pairs_{u,v} := E[\vec{x}_u \otimes \vec{x}_v].
\]

Transforming view \(v\) to view 3:

\[
C_{v\rightarrow 3} := E[\vec{x}_3 \otimes \vec{x}_u] E[\vec{x}_v \otimes \vec{x}_u]^{\dagger} \in \mathbb{R}^{d_3 \times d_v}
\]

where \(\dagger\) denotes Moore-Penrose pseudoinverse.

Simple exercise to show

\[
E[C_{v\rightarrow 3} \vec{x}_v | h = j] = \bar{\mu}_{3,j}
\]

so \(C_{v\rightarrow 3} \vec{x}_v\) behaves like \(\vec{x}_3\) (as far as our algorithm can tell).
Part 3. Some applications and open questions

Multi-view mixture models

Multi-view method-of-moments

Some applications and open questions
- Mixtures of Gaussians
- Hidden Markov models and other models
- Topic models
- Open questions

Concluding remarks
Mixtures of axis-aligned Gaussians

Mixture of axis-aligned Gaussian in $\mathbb{R}^n$, with component means $\vec{\mu}_1, \vec{\mu}_2, \ldots, \vec{\mu}_k \in \mathbb{R}^n$; no minimum separation requirement.

**Assumptions:**
- Non-degeneracy: component means span $k$-dim subspace.
- Weak incoherence condition: component means not perfectly aligned with coordinate axes — similar to spreading condition of (Chaudhuri-Rao, '08).

Then, randomly partitioning coordinates into $\ell \geq 3$ views guarantees (w.h.p.) that non-degeneracy holds in all $\ell$ views.
Mixtures of axis-aligned Gaussians

Mixture of axis-aligned Gaussian in $\mathbb{R}^n$, with component means $\vec{\mu}_1, \vec{\mu}_2, \ldots, \vec{\mu}_k \in \mathbb{R}^n$; no minimum separation requirement.

Assumptions:
- **non-degeneracy**: component means span $k$ dim subspace.
- **weak incoherence condition**: component means not perfectly aligned with coordinate axes — similar to spreading condition of (Chaudhuri-Rao, ’08).
Mixtures of axis-aligned Gaussians

Mixture of axis-aligned Gaussian in $\mathbb{R}^n$, with component means $\vec{\mu}_1, \vec{\mu}_2, \ldots, \vec{\mu}_k \in \mathbb{R}^n$; no minimum separation requirement.

Assumptions:

- **non-degeneracy**: component means span $k$ dim subspace.
- **weak incoherence condition**: component means not perfectly aligned with coordinate axes — similar to spreading condition of (Chaudhuri-Rao, '08).

Then, randomly partitioning coordinates into $\ell \geq 3$ views guarantees (w.h.p.) that non-degeneracy holds in all $\ell$ views.
Hidden Markov models and others

- Mixtures of Gaussians (Hsu-Kakade, ITCS'13)
- HMMs (Anandkumar-Hsu-Kakade, COLT'12)
- Latent Dirichlet Allocation (Anandkumar-Foster-Hsu-Kakade-Liu, NIPS'12)
- Latent parse trees (Hsu-Kakade-Liang, NIPS'12)
- Independent Component Analysis (Arora-Ge-Moitra-Sachdeva, NIPS'12; Hsu-Kakade, ITCS'13)
Hidden Markov models and others

- Mixtures of Gaussians (Hsu-Kakade, ITCS'13)
- HMMs (Anandkumar-Hsu-Kakade, COLT'12)
- Latent Dirichlet Allocation (Anandkumar-Foster-Hsu-Kakade-Liu, NIPS'12)
- Latent parse trees (Hsu-Kakade-Liang, NIPS'12)
- Independent Component Analysis (Arora-Ge-Moitra-Sachdeva, NIPS'12; Hsu-Kakade, ITCS'13)
Hidden Markov models and others

Other models:

1. Mixtures of Gaussians (Hsu-Kakade, ITCS’13)
2. HMMs (Anandkumar-Hsu-Kakade, COLT’12)
3. Latent Dirichlet Allocation
   (Anandkumar-Foster-Hsu-Kakade-Liu, NIPS’12)
4. Latent parse trees (Hsu-Kakade-Liang, NIPS’12)
5. Independent Component Analysis
   (Arora-Ge-Moitra-Sachdeva, NIPS’12; Hsu-Kakade, ITCS’13)
Bag-of-words clustering model

$$(\tilde{\mu}_j)_i = \text{Pr}[\text{see word } i \text{ in document} \mid \text{document topic is } j].$$

- Vocabulary size: $d = 102660$ words.
- Chose $k = 50$.
- For each topic $j$, show top 10 words $i$. 
Bag-of-words clustering model

$$(\vec{\mu}_j)_i = \Pr[\text{see word } i \text{ in document } | \text{ document topic is } j].$$

- Vocabulary size: $d = 102660$ words.
- Chose $k = 50$.
- For each topic $j$, show top 10 words $i$.

<table>
<thead>
<tr>
<th>sales</th>
<th>run</th>
<th>school</th>
<th>drug</th>
<th>player</th>
</tr>
</thead>
<tbody>
<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>
Bag-of-words clustering model

| palestinian | tax        | cup        | point       | yard       |
| israel      | cut        | minutes    | game        | game       |
| israeli     | percent    | oil        | team        | play       |
| yasser_arafat | bush      | water      | shot        | season     |
| peace       | plan       | add        | play        | touchdown  |
| israeli     | bill       | tablespoon | laker       | quarterback|
| israelis    | taxes      | food       | season      | coach      |
| leader      | million    | teaspoon   | half        | defense    |
| official    | congress   | pepper     | lead        | quarter    |
| attack      |           | sugar      | games       |           |
### Bag-of-words Clustering Model

<table>
<thead>
<tr>
<th>percent</th>
<th>al_gore</th>
<th>car</th>
<th>book</th>
<th>taliban</th>
</tr>
</thead>
<tbody>
<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>official</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>
## Bag-of-words clustering model

<table>
<thead>
<tr>
<th>com</th>
<th>court</th>
<th>show</th>
<th>film</th>
<th>music</th>
</tr>
</thead>
<tbody>
<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>song</td>
</tr>
<tr>
<td>web</td>
<td>lawyer</td>
<td>nbc</td>
<td>play</td>
<td>song</td>
</tr>
<tr>
<td>sites</td>
<td>federal</td>
<td>cb</td>
<td>character</td>
<td>song</td>
</tr>
<tr>
<td>information</td>
<td>government</td>
<td>program</td>
<td>actor</td>
<td>song</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>show</td>
</tr>
<tr>
<td>internet</td>
<td>microsoft</td>
<td>night</td>
<td>million</td>
<td>part</td>
</tr>
<tr>
<td>telegram</td>
<td>right</td>
<td>new_york</td>
<td>part</td>
<td>album</td>
</tr>
</tbody>
</table>

etc.
Some open questions

What if $k > d_v$? (relevant to overcomplete dictionary learning)
Some open questions

What if $k > d_v$? (relevant to overcomplete dictionary learning)

- Apply some non-linear transformations $\vec{x}_v \mapsto f_v(\vec{x}_v)$?
Some open questions

What if $k > d_v$? (relevant to overcomplete dictionary learning)

▶ Apply some non-linear transformations $\tilde{x}_v \mapsto f_v(\tilde{x}_v)$?

▶ Combine views, e.g., via tensor product

\[
\tilde{x}_{1,2} := \tilde{x}_1 \otimes \tilde{x}_2, \quad \tilde{x}_{3,4} := \tilde{x}_3 \otimes \tilde{x}_4, \quad \tilde{x}_{5,6} := \tilde{x}_5 \otimes \tilde{x}_6, \quad \text{etc.} \]
Some open questions

What if \( k > d_v \)? (relevant to overcomplete dictionary learning)

- Apply some non-linear transformations \( \tilde{x}_v \mapsto f_v(\tilde{x}_v) \)?
- Combine views, e.g., via tensor product

\[
\tilde{x}_{1,2} := \tilde{x}_1 \otimes \tilde{x}_2, \quad \tilde{x}_{3,4} := \tilde{x}_3 \otimes \tilde{x}_4, \quad \tilde{x}_{5,6} := \tilde{x}_5 \otimes \tilde{x}_6, \quad \text{etc.}
\]

Can we relax the multi-view assumption?
Some open questions

What if $k > d_v$? (relevant to overcomplete dictionary learning)

- Apply some non-linear transformations $\bar{x}_v \mapsto f_v(\bar{x}_v)$?
- Combine views, e.g., via tensor product

$$\tilde{x}_{1,2} := \bar{x}_1 \otimes \bar{x}_2, \quad \tilde{x}_{3,4} := \bar{x}_3 \otimes \bar{x}_4, \quad \tilde{x}_{5,6} := \bar{x}_5 \otimes \bar{x}_6,$$
  etc.?

Can we relax the multi-view assumption?

- Allow for richer hidden state?
  (e.g., independent component analysis)

"Gaussianization" via random projection?
Some open questions

What if $k > d_v$? (relevant to overcomplete dictionary learning)

- Apply some non-linear transformations $\tilde{x}_v \mapsto f_v(\tilde{x}_v)$?
- Combine views, e.g., via tensor product

\[ \tilde{x}_{1,2} := \tilde{x}_1 \otimes \tilde{x}_2, \quad \tilde{x}_{3,4} := \tilde{x}_3 \otimes \tilde{x}_4, \quad \tilde{x}_{5,6} := \tilde{x}_5 \otimes \tilde{x}_6, \quad \text{etc.} \]

Can we relax the multi-view assumption?

- Allow for richer hidden state?
  (e.g., independent component analysis)
- “Gaussianization” via random projection?
Part 4. Concluding remarks

Multi-view mixture models

Multi-view method-of-moments

Some applications and open questions

Concluding remarks
Concluding remarks

Take-home messages:

▶ Power of multiple views: Can take advantage of diverse / non-redundant sources of information in unsupervised learning.

▶ Overcoming complexity barriers: Some provably hard estimation problems become easy after ruling out "degenerate" cases.

▶ "Blessing of dimensionality" for estimators based on method-of-moments.
Concluding remarks

Take-home messages:

▶ **Power of multiple views**: Can take advantage of diverse / non-redundant sources of information in unsupervised learning.

▶ Overcoming complexity barriers: Some provably hard estimation problems become easy after ruling out "degenerate" cases.

▶ "Blessing of dimensionality" for estimators based on method-of-moments.
Concluding remarks

Take-home messages:

- **Power of multiple views**: Can take advantage of diverse / non-redundant sources of information in unsupervised learning.

- **Overcoming complexity barriers**: Some provably hard estimation problems become easy after ruling out “degenerate” cases.
Concluding remarks

Take-home messages:

- **Power of multiple views**: Can take advantage of diverse / non-redundant sources of information in unsupervised learning.

- **Overcoming complexity barriers**: Some provably hard estimation problems become easy after ruling out “degenerate” cases.

- “Blessing of dimensionality” for estimators based on method-of-moments.
Thanks!

(Co-authors: Anima Anandkumar, Dean Foster, Rong Ge, Sham Kakade, Yi-Kai Liu, Matus Telgarsky)

http://arxiv.org/abs/1210.7559