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

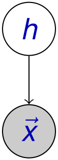
- **Many modern applications of machine learning**:
  - high-dimensional data from many diverse sources,
  - but mostly unlabeled.

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



$h \in [k] := \{1, 2, \ldots, k\}$ (hidden);

$\vec{x} \in \mathbb{R}^d$ (observed);

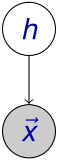
$\Pr[\ h = j\ ] = w_j$; $\quad \vec{x} \big| h \sim \mathbb{P}_h$;

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



$h \in [k] := \{1, 2, \ldots, k\}$ (hidden);

$\vec{x} \in \mathbb{R}^d$ (observed);

$\Pr[\ h = j\ ] = w_j$; $\quad \vec{x} \big| h \sim \mathbb{P}_h$;

so $\vec{x}$ has a mixture distribution

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

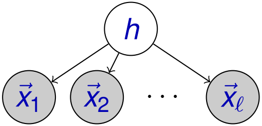
Typical use: learn about constituent sub-populations (*e.g.*, clusters) in data source.

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?



$h \in [k]$,

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

k = # components, ℓ = # views (*e.g.*, audio, video, text).

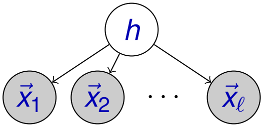
View 1: $\vec{x}_1 \in \mathbb{R}^{d_1}$

View 2: $\vec{x}_2 \in \mathbb{R}^{d_2}$

View 3: $\vec{x}_3 \in \mathbb{R}^{d_3}$

Multi-view mixture models

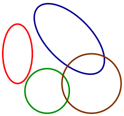
Can we take advantage of diverse sources of information?



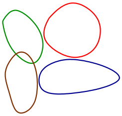
$h \in [k]$,

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

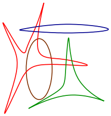
$k = \#$ components, $\ell = \#$ views (*e.g.*, audio, video, text).



View 1: $\vec{x}_1 \in \mathbb{R}^{d_1}$



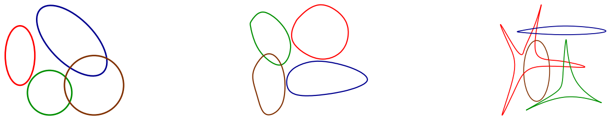
View 2: $\vec{x}_2 \in \mathbb{R}^{d_2}$



View 3: $\vec{x}_3 \in \mathbb{R}^{d_3}$

Multi-view mixture models

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



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

Larger k (# components): more sub-populations to disentangle.

Larger ℓ (# views): more non-redundant sources of information.

Semi-parametric estimation task

"Parameters" of component distributions:

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

$$\text{Conditional means } \vec{\mu}_{v,j} := \mathbb{E}[\vec{x}_v | h = j] \in \mathbb{R}^{d_v}, \quad j \in [k], v \in [\ell].$$

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

Semi-parametric estimation task

"Parameters" of component distributions:

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

$$\text{Conditional means } \vec{\mu}_{v,j} := \mathbb{E}[\vec{x}_v | h = j] \in \mathbb{R}^{d_v}, \quad j \in [k], v \in [\ell].$$

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

Questions:

1. How do we estimate $\{w_j\}$ and $\{\vec{\mu}_{v,j}\}$ without observing h?
2. How many views ℓ 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

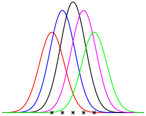
Challenge: many difficult parametric estimation tasks reduce to this estimation problem.

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).

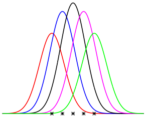


Statistical barrier: Gaussian mixtures in $\mathbb{R}^1$ can require $\exp(\Omega(k))$ samples to estimate parameters, even if components are well-separated (Moitra-Valiant, '10).

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):

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

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

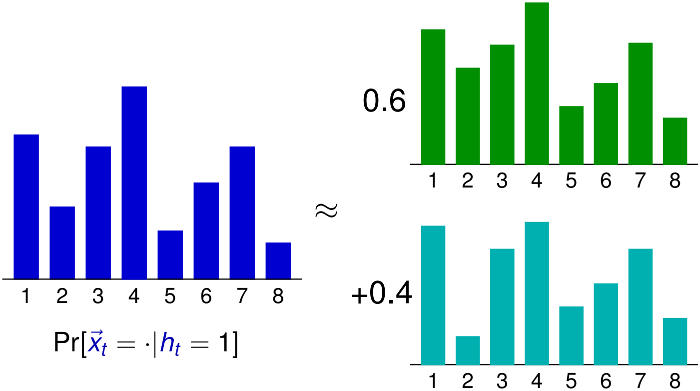
Gaussian mixture model: problem becomes easier if assume some large minimum separation between component means (Dasgupta, '99):

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

- $\text{sep} = \Omega(d^c)$: 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.
- 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.

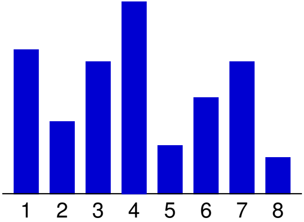


$\Pr[\vec{x}_t = \cdot | h_t = 1]$

$0.6 \Pr[\vec{x}_t = \cdot | h_t = 2] + 0.4 \Pr[\vec{x}_t = \cdot | h_t = 3]$

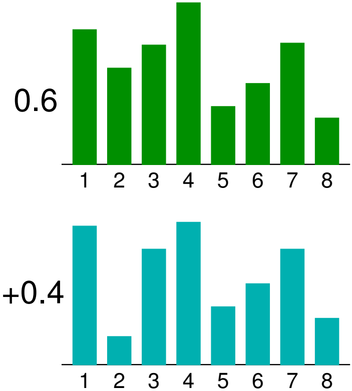
Making progress: discrete hidden Markov models

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



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

$\approx$



$0.6 \Pr[\vec{x}_t = \cdot | h_t = 2] + 0.4 \Pr[\vec{x}_t = \cdot | h_t = 3]$

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

What we do

This work: given ≥ 3 views, mild non-degeneracy conditions imply efficient algorithms for estimation.

What we do

This work: given ≥ 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 ($d_v \geq k$)!

What we do

This work: given ≥ 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)

What we do

This work: given ≥ 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.

Part 2. Multi-view method-of-moments

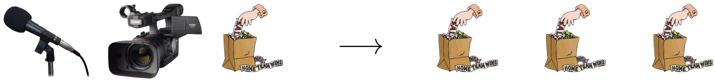
The plan

- First, assume views are **(conditionally) exchangeable**, and derive basic algorithm.

The plan

- First, assume views are **(conditionally) exchangeable**, and derive basic algorithm.

- Then, provide **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.*,

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

Simpler case: exchangeable views

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

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

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

One-hot encoding:

$$\vec{x}_v = \vec{e}_i \quad \Leftrightarrow \quad v\text{-th word in document is } i\text{-th word in vocab}$$

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

$$(\vec{\mu}_j)_i \ = \ \mathbb{E}[(\vec{x}_v)_i | h = j] \ = \ \Pr[\vec{x}_v = \vec{e}_i | h = j], \quad i \in [d], j \in [k].$$

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

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,

$$\begin{aligned}
\text{Pairs} &:= \mathbb{E}[\vec{x}_1 \otimes \vec{x}_2] \\
&= \mathbb{E}\big[\mathbb{E}[\vec{x}_1 | h] \otimes \mathbb{E}[\vec{x}_2 | h]\big] = \mathbb{E}[\vec{\mu}_h \otimes \vec{\mu}_h] \\
&= \sum_{i=1}^{k} w_i\, \vec{\mu}_i \otimes \vec{\mu}_i \;\in \mathbb{R}^{d \times d}.
\end{aligned}$$

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,

$$\begin{aligned}
\text{Pairs} &:= \mathbb{E}[\vec{x}_1 \otimes \vec{x}_2] \\
&= \mathbb{E}\big[\mathbb{E}[\vec{x}_1 | h] \otimes \mathbb{E}[\vec{x}_2 | h]\big] = \mathbb{E}[\vec{\mu}_h \otimes \vec{\mu}_h] \\
&= \sum_{i=1}^{k} w_i\ \vec{\mu}_i \otimes \vec{\mu}_i\ \in \mathbb{R}^{d \times d}.
\end{aligned}$$

$$\begin{aligned}
\text{Triples} &:= \mathbb{E}[\vec{x}_1 \otimes \vec{x}_2 \otimes \vec{x}_3] \\
&= \sum_{i=1}^{k} w_i\ \vec{\mu}_i \otimes \vec{\mu}_i \otimes \vec{\mu}_i\ \in \mathbb{R}^{d \times d \times d}, \quad \textit{etc.}
\end{aligned}$$

(If only we could extract these “low-rank” decompositions . . .)

2nd moment: subspace spanned by conditional means

2nd moment: subspace spanned by conditional means

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

2nd moment: subspace spanned by conditional means

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

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

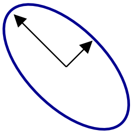
2nd moment: subspace spanned by conditional means

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$\{\vec{\mu}_1, \vec{\mu}_2, \dots, \vec{\mu}_k\}$ with inner product

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



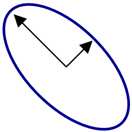
2nd moment: subspace spanned by conditional means

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

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

$\implies$ 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).

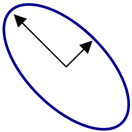
2nd moment: subspace spanned by conditional means

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

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

$\implies$ 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?

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.

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

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

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

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

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*

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

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

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

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

Variational analysis

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

Variational analysis

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

(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

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

(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

$$\boxed{\max_{\vec{\eta} \in \mathbb{R}^d} \sum_{i=1}^{k} w_i \, (\vec{\eta}^\top \vec{\mu}_i)^3 \;\; \text{s.t.} \;\; \sum_{i=1}^{k} w_i \, (\vec{\eta}^\top \vec{\mu}_i)^2 \leq 1}$$

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

Variational analysis

$$\boxed{\max_{\vec{\eta} \in \mathbb{R}^d} \sum_{i=1}^{k} \frac{1}{\sqrt{w_i}} \left(\sqrt{w_i} \vec{\eta}^\top \vec{\mu}_i\right)^3 \quad \text{s.t.} \quad \sum_{i=1}^{k} \left(\sqrt{w_i} \vec{\eta}^\top \vec{\mu}_i\right)^2 \leq 1}$$

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

Variational analysis

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

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

Variational analysis

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

(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

$$\vec{e}_1 = (1, 0, 0, \dots), \quad \vec{e}_2 = (0, 1, 0, \dots), \quad \textit{etc.}$$

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

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

Variational analysis

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

(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

$$\vec{e}_1 = (1, 0, 0, \dots), \quad \vec{e}_2 = (0, 1, 0, \dots), \quad \textit{etc.}$$

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 $T :=$ Triples.

Extracting all isolated local maximizers

1. Start with $T :=$ Triples.
2. Find isolated local maximizer of

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

   via gradient ascent from random $\vec{\eta} \in \text{range}(\text{Pairs})$.
   Say maximum is λ^* and maximizer is $\vec{\eta}^*$.

Extracting all isolated local maximizers

1. Start with $T :=$ Triples.
2. Find isolated local maximizer of

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

   via gradient ascent from random $\vec{\eta} \in$ range(Pairs).
   Say maximum is λ^* 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 $T :=$ Triples.
2. Find isolated local maximizer of

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

   via gradient ascent from random $\vec{\eta} \in \text{range}(\text{Pairs})$.
   Say maximum is λ^* 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.

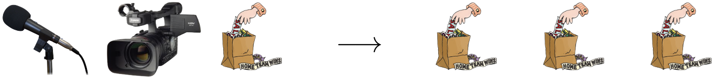
General case: asymmetric views

Each view v has different set of conditional means $\{\vec{\mu}_{v,1}, \vec{\mu}_{v,2}, \ldots, \vec{\mu}_{v,k}\} \subset \mathbb{R}^{d_v}$.

General case: asymmetric views

Each view v has different set of conditional means $\{\vec{\mu}_{v,1}, \vec{\mu}_{v,2}, \ldots, \vec{\mu}_{v,k}\} \subset \mathbb{R}^{d_v}$.

Reduction: transform $\vec{x}_1$ and $\vec{x}_2$ to “look like” $\vec{x}_3$ 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:

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

Transforming view v to view 3:

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

where † denotes Moore-Penrose pseudoinverse.

Asymmetric cross moments

Define asymmetric cross moment:

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

Transforming view v to view 3:

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

where † denotes Moore-Penrose pseudoinverse.

Simple exercise to show

$$\mathbb{E}[C_{v \to 3}\vec{x}_v | h = j] = \vec{\mu}_{3,j}$$

so $C_{v \to 3}\vec{x}_v$ 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

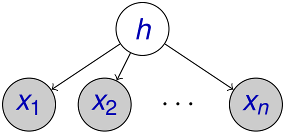
Some applications and open questions

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

Concluding remarks

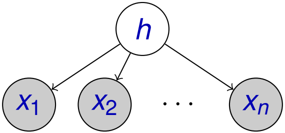
Mixtures of axis-aligned Gaussians

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.



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.

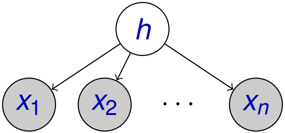


Assumptions:

- 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.

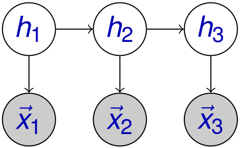


Assumptions:

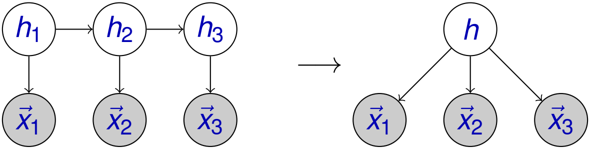
- non-degeneracy: component means span k dim subspace.
- weak incoherence condition: component means not perfectly aligned with coordinate axes — similar to *spreading condition* of (Chaudhuri-Rao, '08).

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

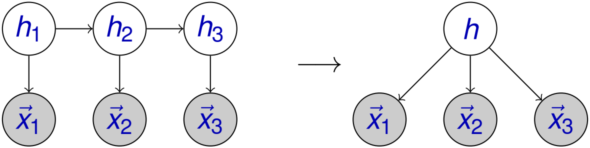
Hidden Markov models and others



Hidden Markov models and others



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

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

- Corpus: New York Times (from UCI), 300000 articles.
- Vocabulary size: $d = 102660$ words.
- Chose $k = 50$.
- For each topic j, show top 10 words i.

Bag-of-words clustering model

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

- Corpus: New York Times (from UCI), 300000 articles.
- Vocabulary size: $d = 102660$ words.
- Chose $k = 50$.
- For each topic j, show top 10 words i.

sales	run	school	drug	player
economic	inning	student	patient	tiger_wood
consumer	hit	teacher	million	won
major	game	program	company	shot
home	season	official	doctor	play
indicator	home	public	companies	round
weekly	right	children	percent	win
order	games	high	cost	tournament
claim	dodger	education	program	tour
scheduled	left	district	health	right

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	billion	add	play	team
israeli	plan	tablespoon	laker	touchdown
israelis	bill	food	season	quarterback
leader	taxes	teaspoon	half	coach
official	million	pepper	lead	defense
attack	congress	sugar	games	quarter

Bag-of-words clustering model

percent	al_gore	car	book	taliban
stock	campaign	race	children	attack
market	president	driver	ages	afghanistan
fund	george_bush	team	author	official
investor	bush	won	read	military
companies	clinton	win	newspaper	u_s
analyst	vice	racing	web	united_states
money	presidential	track	writer	terrorist
investment	million	season	written	war
economy	democratic	lap	sales	bin

Bag-of-words clustering model

com	court	show	film	music
www	case	network	movie	song
site	law	season	director	group
web	lawyer	nbc	play	part
sites	federal	cb	character	new_york
information	government	program	actor	company
online	decision	television	show	million
mail	trial	series	movies	band
internet	microsoft	night	million	show
telegram	right	new_york	part	album

etc.

Some open questions

What if $\mathbf{k} > \mathbf{d_v}$? (relevant to overcomplete dictionary learning)

Some open questions

What if $\mathbf{k > d_v}$? (relevant to overcomplete dictionary learning)

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

Some open questions

What if $\mathbf{k > d_v}$? (relevant to overcomplete dictionary learning)

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

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

Some open questions

What if $\mathbf{k} > \mathbf{d_v}$? (relevant to overcomplete dictionary learning)

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

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

Can we relax the multi-view assumption?

Some open questions

What if $\mathbf{k > d_v}$? (relevant to overcomplete dictionary learning)

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

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

Can we relax the multi-view assumption?

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

Some open questions

What if $\mathbf{k > d_v}$? (relevant to overcomplete dictionary learning)

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

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

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

Concluding remarks

Take-home messages:

Concluding remarks

Take-home messages:

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

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`