A Method of Moments for Mixture Models and Hidden Markov Models

Anima Anandkumar[@]   Daniel Hsu[#]   Sham M. Kakade[#]

[@]University of California, Irvine

[#]Microsoft Research, New England

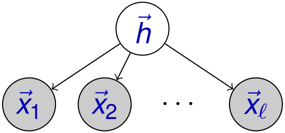
Outline

1. Latent class models and parameter estimation
2. Multi-view method of moments
3. Some applications
4. Concluding remarks

1. Latent class models and parameter estimation

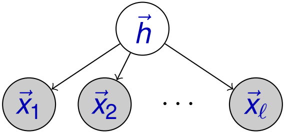
Latent class models / multi-view mixture models

Random vectors $\vec{h} \in \{\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_k\} \in \mathbb{R}^k$, $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_\ell \in \mathbb{R}^d$.



Latent class models / multi-view mixture models

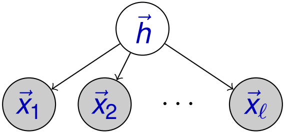
Random vectors $\vec{h} \in \{\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_k\} \in \mathbb{R}^k$, $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_\ell \in \mathbb{R}^d$.



- **Bags-of-words clustering model**: k = number of topics, d = vocabulary size, $\vec{h}$ = topic of document, $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_\ell \in \{\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_d\}$ words in the document.

Latent class models / multi-view mixture models

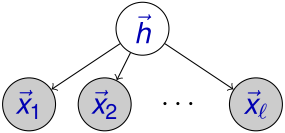
Random vectors $\vec{h} \in \{\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_k\} \in \mathbb{R}^k$, $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_\ell \in \mathbb{R}^d$.



- **Bags-of-words clustering model**: k = number of topics, d = vocabulary size, $\vec{h}$ = topic of document, $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_\ell \in \{\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_d\}$ words in the document.
- **Multi-view clustering**: k = number of clusters, ℓ = number of views (*e.g.*, audio, video, text); views assumed to be conditionally independent given the cluster.

Latent class models / multi-view mixture models

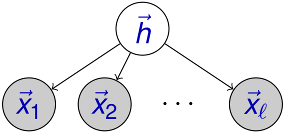
Random vectors $\vec{h} \in \{\vec{e}_1, \vec{e}_2, \dots, \vec{e}_k\} \in \mathbb{R}^k$, $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_\ell \in \mathbb{R}^d$.



- **Bags-of-words clustering model**: k = number of topics, d = vocabulary size, $\vec{h}$ = topic of document, $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_\ell \in \{\vec{e}_1, \vec{e}_2, \dots, \vec{e}_d\}$ words in the document.
- **Multi-view clustering**: k = number of clusters, ℓ = number of views (*e.g.*, audio, video, text); views assumed to be conditionally independent given the cluster.
- **Hidden Markov model**: ($\ell = 3$) past, present, and future observations are conditionally independent given present hidden state.

Latent class models / multi-view mixture models

Random vectors $\vec{h} \in \{\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_k\} \in \mathbb{R}^k$, $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_\ell \in \mathbb{R}^d$.



- **Bags-of-words clustering model**: k = number of topics, d = vocabulary size, $\vec{h}$ = topic of document, $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_\ell \in \{\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_d\}$ words in the document.
- **Multi-view clustering**: k = number of clusters, ℓ = number of views (*e.g.*, audio, video, text); views assumed to be conditionally independent given the cluster.
- **Hidden Markov model**: ($\ell = 3$) past, present, and future observations are conditionally independent given present hidden state.
- etc.

Parameter estimation task

Model parameters: mixing weights and conditional means

$$w_j := \Pr[\vec{h} = \vec{e}_j], \quad j \in [k];$$
$$\vec{\mu}_{v,j} := \mathbb{E}[\vec{x}_v | \vec{h} = \vec{e}_j] \in \mathbb{R}^d, \quad v \in [\ell], j \in [k].$$

Goal: given i.i.d. copies of $(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_\ell)$, estimate matrix of conditional means $M_v := [\vec{\mu}_{v,1} | \vec{\mu}_{v,2} | \cdots | \vec{\mu}_{v,k}]$ for each view $v \in [\ell]$, and mixing weights $\vec{w} := (w_1, w_2, \ldots, w_k)$.

Parameter estimation task

Model parameters: mixing weights and conditional means

$$w_j := \Pr[\vec{h} = \vec{e}_j], \quad j \in [k];$$
$$\vec{\mu}_{v,j} := \mathbb{E}[\vec{x}_v | \vec{h} = \vec{e}_j] \in \mathbb{R}^d, \quad v \in [\ell], j \in [k].$$

Goal: given i.i.d. copies of $(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_\ell)$, estimate matrix of conditional means $M_v := [\vec{\mu}_{v,1} | \vec{\mu}_{v,2} | \cdots | \vec{\mu}_{v,k}]$ for each view $v \in [\ell]$, and mixing weights $\vec{w} := (w_1, w_2, \ldots, w_k)$.

Unsupervised learning, as $\vec{h}$ is not observed.

Parameter estimation task

Model parameters: mixing weights and conditional means

$$w_j := \Pr[\vec{h} = \vec{e}_j], \quad j \in [k];$$
$$\vec{\mu}_{v,j} := \mathbb{E}[\vec{x}_v | \vec{h} = \vec{e}_j] \in \mathbb{R}^d, \quad v \in [\ell], j \in [k].$$

Goal: given i.i.d. copies of $(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_\ell)$, estimate matrix of conditional means $M_v := [\vec{\mu}_{v,1} | \vec{\mu}_{v,2} | \cdots | \vec{\mu}_{v,k}]$ for each view $v \in [\ell]$, and mixing weights $\vec{w} := (w_1, w_2, \ldots, w_k)$.

Unsupervised learning, as $\vec{h}$ is not observed.

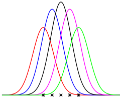
This talk: very general and computationally efficient method-of-moments estimator for $\vec{w}$ and M_v.

Some barriers to efficient estimation

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

Some barriers to efficient estimation

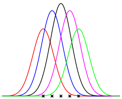
Cryptographic barrier: HMM parameter estimation as hard as learning parity functions with noise (Mossel-Roch, '06).



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

Some barriers to efficient estimation

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



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

Practitioners typically resort to local search heuristics (EM); plagued by 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):

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

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)

Making progress: Gaussian mixture model

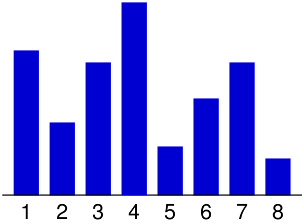
Gaussian mixture model: problem becomes easier if assume some large minimum separation between component means (Dasgupta, '99):

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

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

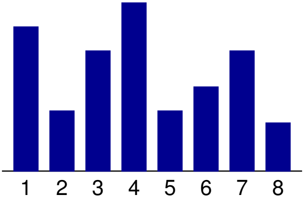
Making progress: hidden Markov models

Hardness reductions create HMMs where different states may have near-identical output and next-state distributions.



$\Pr[\vec{x}_t = \cdot | \vec{h}_t = \vec{e}_1]$

$\approx$

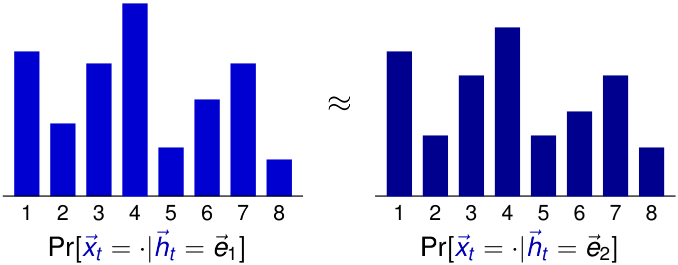


$\Pr[\vec{x}_t = \cdot | \vec{h}_t = \vec{e}_2]$

Can avoid these instances if we assume transition and output parameter matrices are full-rank.

Making progress: hidden Markov models

Hardness reductions create HMMs where different states may have near-identical output and next-state distributions.



Can avoid these instances if we assume transition and output parameter matrices are full-rank.

- $d = k$: eigenvalue decompositions (Chang, '96; Mossel-Roch, '06)
- $d \geq k$: subspace ID + observable operator model (Hsu-Kakade-Zhang, '09)

What we do

This work: Concept of “full rank” parameter matrices is generic and very powerful; adapt Chang’s method for more general mixture models.

What we do

This work: Concept of “full rank” parameter matrices is generic and very powerful; adapt Chang’s method for more general mixture models.

- **Non-degeneracy condition** for latent class model: M_v has full column rank ($\forall v \in [\ell]$), and $\vec{w} > 0$.

What we do

This work: Concept of "full rank" parameter matrices is generic and very powerful; adapt Chang's method for more general mixture models.

- **Non-degeneracy condition** for latent class model: M_v has full column rank ($\forall v \in [\ell]$), and $\vec{w} > 0$.
- New efficient learning results for:
  - Certain Gaussian mixture models, with no minimum separation requirement and poly(k) sample / computational complexity
  - HMMs with discrete or continuous output distributions (*e.g.*, Gaussian mixture outputs)

2. Multi-view method of moments

Simplified model and low-order statistics

Simplification: $M_v \equiv M$ (same conditional means for all views);

Simplified model and low-order statistics

Simplification: $M_v \equiv M$ (same conditional means for all views);

If $\vec{x}_v \in \{\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_d\}$ (discrete outputs), then

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

Simplified model and low-order statistics

Simplification: $M_v \equiv M$ (same conditional means for all views);

If $\vec{x}_v \in \{\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_d\}$ (discrete outputs), then

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

So pair-wise and triple-wise statistics are:

$$\text{Pairs}_{i,j} := \Pr[\vec{x}_1 = \vec{e}_i \ \wedge \ \vec{x}_2 = \vec{e}_j], \quad i, j \in [d]$$
$$\text{Triples}_{i,j,\kappa} := \Pr[\vec{x}_1 = \vec{e}_i \ \wedge \ \vec{x}_2 = \vec{e}_j \ \wedge \ \vec{x}_3 = \vec{e}_\kappa], \quad i, j, \kappa \in [d].$$

Simplified model and low-order statistics

Simplification: $M_v \equiv M$ (same conditional means for all views);

If $\vec{x}_v \in \{\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_d\}$ (discrete outputs), then

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

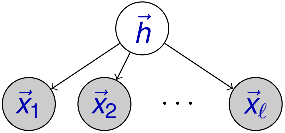
So pair-wise and triple-wise statistics are:

$$\text{Pairs}_{i,j} := \Pr[\vec{x}_1 = \vec{e}_i \ \wedge \ \vec{x}_2 = \vec{e}_j], \quad i,j \in [d]$$
$$\text{Triples}_{i,j,\kappa} := \Pr[\vec{x}_1 = \vec{e}_i \ \wedge \ \vec{x}_2 = \vec{e}_j \ \wedge \ \vec{x}_3 = \vec{e}_\kappa], \quad i,j,\kappa \in [d].$$

Notation: for $\vec{\eta} = (\eta_1, \eta_2, \ldots, \eta_d) \in \mathbb{R}^d$,

$$\text{Triples}_{i,j}(\vec{\eta}) := \sum_{\kappa=1}^{d} \eta_\kappa \Pr[\vec{x}_1 = \vec{e}_i \ \wedge \ \vec{x}_2 = \vec{e}_j \ \wedge \ \vec{x}_3 = \vec{e}_\kappa], \quad i,j \in [d].$$

Algebraic structure in moments



By conditional independence of $\vec{x}_1, \vec{x}_2, \vec{x}_3$ given $\vec{h}$,

$$\begin{aligned} \text{Pairs} &= M \operatorname{diag}(\vec{w}) M^\top \\ \text{Triples}(\vec{\eta}) &= M \operatorname{diag}(M^\top \vec{\eta}) \operatorname{diag}(\vec{w}) M^\top . \end{aligned}$$

(Low-rank matrix factorizations,
but M not necessarily orthonormal.)

Developing a method of moments

For simplicity, assume $d = k$ (all matrices are square).

Developing a method of moments

For simplicity, assume $d = k$ (all matrices are square). Recall:

$$\text{Pairs} = M \ \text{diag}(\vec{w}) \ M^\top$$
$$\text{Triples}(\vec{\eta}) = M \ \text{diag}(M^\top \vec{\eta}) \ \text{diag}(\vec{w}) \ M^\top$$

Developing a method of moments

For simplicity, assume $d = k$ (all matrices are square). Recall:

$$\text{Pairs} = M \ \text{diag}(\vec{w}) \ M^\top$$
$$\text{Triples}(\vec{\eta}) = M \ \text{diag}(M^\top \vec{\eta}) \ \text{diag}(\vec{w}) \ M^\top$$

and therefore

$$\text{Triples}(\vec{\eta}) \ \text{Pairs}^{-1} = M \, \text{diag}(M^\top \vec{\eta}) M^{-1},$$

Developing a method of moments

For simplicity, assume $d = k$ (all matrices are square). Recall:

$$\text{Pairs} = M \, \text{diag}(\vec{w}) \, M^\top$$
$$\text{Triples}(\vec{\eta}) = M \, \text{diag}(M^\top \vec{\eta}) \, \text{diag}(\vec{w}) \, M^\top$$

and therefore

$$\text{Triples}(\vec{\eta}) \, \text{Pairs}^{-1} = M \, \text{diag}(M^\top \vec{\eta}) M^{-1},$$

a diagonalizable matrix of the form $V \Lambda V^{-1}$,
where $V = M$ (eigenvectors) and $\Lambda = \text{diag}(M^\top \vec{\eta})$ (eigenvalues).

Developing a method of moments

For simplicity, assume $d = k$ (all matrices are square). Recall:

$$\text{Pairs} = M \ \text{diag}(\vec{w}) \ M^\top$$
$$\text{Triples}(\vec{\eta}) = M \ \text{diag}(M^\top \vec{\eta}) \ \text{diag}(\vec{w}) \ M^\top$$

and therefore

$$\text{Triples}(\vec{\eta}) \ \text{Pairs}^{-1} = M \, \text{diag}(M^\top \vec{\eta}) M^{-1},$$

a diagonalizable matrix of the form $V \Lambda V^{-1}$,
where $V = M$ (eigenvectors) and $\Lambda = \text{diag}(M^\top \vec{\eta})$ (eigenvalues).

(If $d > k$, use SVD to reduce dimension.)

Plug-in estimator

1. Obtain empirical estimates $\widehat{\text{Pairs}}$ and $\widehat{\text{Triples}}$ of Pairs and Triples.
2. Compute matrix of k orthonormal left singular vectors $\widehat{U}$ using rank-k SVD of $\widehat{\text{Pairs}}$.
3. Randomly pick unit vector $\vec{\theta} \in \mathbb{R}^k$.
4. Compute right eigenvectors $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_k$ of

$$\left(\widehat{U}^\top \widehat{\text{Triples}}(\widehat{U}\vec{\theta})\widehat{U}\right)\left(\widehat{U}^\top \widehat{\text{Pairs}}\,\widehat{U}\right)^{-1}$$

   and return

$$\widehat{M} := [\widehat{U}\vec{v}_1 | \widehat{U}\vec{v}_2 | \cdots | \widehat{U}\vec{v}_k]$$

   as conditional mean parameter estimates (up to scaling).

In general, proper scaling can be determined from *eigenvalues*.

Accuracy guarantee

Theorem (discrete outputs)

*Assume **non-degeneracy condition** holds.*
If $\widehat{\text{Pairs}}$ and $\widehat{\text{Triples}}$ are empirical frequencies obtained from random sample of size

$$\frac{\text{poly}\left(k, \sigma_{\min}(M)^{-1}, w_{\min}^{-1}\right)}{\epsilon^2},$$

then with high probability, there exists a permutation matrix Π such that the $\widehat{M}$ returned by plug-in estimator satisfies

$$\|\widehat{M}\Pi - M\| \leq \epsilon.$$

Accuracy guarantee

Theorem (discrete outputs)

*Assume **non-degeneracy condition** holds.*
If $\widehat{\text{Pairs}}$ and $\widehat{\text{Triples}}$ are empirical frequencies obtained from random sample of size

$$\frac{\text{poly}\left(k, \sigma_{\min}(M)^{-1}, w_{\min}^{-1}\right)}{\epsilon^2},$$

then with high probability, there exists a permutation matrix Π such that the $\widehat{M}$ returned by plug-in estimator satisfies

$$\|\widehat{M}\Pi - M\| \leq \epsilon.$$

Role of non-degeneracy: $\sigma_{\min}(M)^{-1}$ and $w_{\min}^{-1}$ in sample complexity bound.

Additional details (see paper)

Additional details (see paper)

- Can also obtain estimate for mixing weights $\vec{w}$.

Additional details (see paper)

- Can also obtain estimate for mixing weights $\vec{w}$.
- General setting: different conditional mean matrices for different views; some non-discrete observed variables.

Additional details (see paper)

- Can also obtain estimate for mixing weights $\vec{w}$.
- General setting: different conditional mean matrices for different views; some non-discrete observed variables.
  - Similar sample complexity bound for models with continuous but subgaussian (or log-concave, etc.) $\vec{x}_v$'s.

Additional details (see paper)

- Can also obtain estimate for mixing weights $\vec{w}$.
- General setting: different conditional mean matrices for different views; some non-discrete observed variables.
  - Similar sample complexity bound for models with continuous but subgaussian (or log-concave, etc.) $\vec{x}_v$'s.
  - Delicate alignment issue: how to make sure columns of $\widehat{M}_1$ are in same order as columns of $\widehat{M}_2$?

Additional details (see paper)

- Can also obtain estimate for mixing weights $\vec{w}$.
- General setting: different conditional mean matrices for different views; some non-discrete observed variables.
  - Similar sample complexity bound for models with continuous but subgaussian (or log-concave, etc.) $\vec{x}_v$'s.
  - Delicate alignment issue: how to make sure columns of $\widehat{M}_1$ are in same order as columns of $\widehat{M}_2$?
  - Solution: reuse eigenvectors whenever possible and align based on eigenvalues.

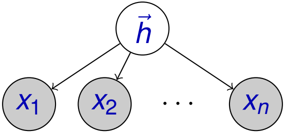
Additional details (see paper)

- Can also obtain estimate for mixing weights $\vec{w}$.
- General setting: different conditional mean matrices for different views; some non-discrete observed variables.
  - Similar sample complexity bound for models with continuous but subgaussian (or log-concave, etc.) $\vec{x}_v$'s.
  - Delicate alignment issue: how to make sure columns of $\widehat{M}_1$ are in same order as columns of $\widehat{M}_2$?
  - Solution: reuse eigenvectors whenever possible and align based on eigenvalues.
- Many variants possible (*e.g.*, symmetrization to only deal with orthogonal eigenvectors) — easy to design once you see the structure.

3. Some applications

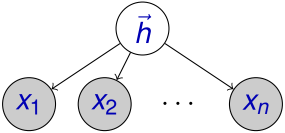
Mixtures of axis-aligned Gaussians

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.

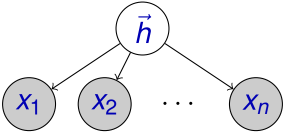


Assumptions:

- non-degeneracy: component means span k dimensional subspace.
- 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.

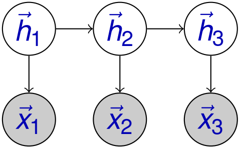


Assumptions:

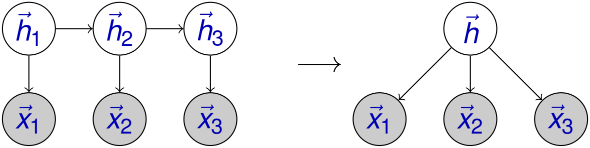
- non-degeneracy: component means span k dimensional subspace.
- incoherence condition: component means not perfectly aligned with coordinate axes — similar to *spreading condition* of (Chaudhuri-Rao, '08).

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

Hidden Markov models



Hidden Markov models



Bag-of-words clustering model

$M_{i,j} = \Pr[\text{see word } i \text{ in article} | \text{article topic is } j]$.

- 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 ordered by $\widehat{M}_{i,j}$ value.

Bag-of-words clustering model

$M_{i,j} = \Pr[\text{see word } i \text{ in article} | \text{article topic is } j]$.

- 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 ordered by $\widehat{M}_{i,j}$ value.

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.

4. Concluding remarks

Concluding remarks

Take-home messages:

Concluding remarks

Take-home messages:

- Some provably hard parameter estimation problems become easy after ruling out "degenerate" cases.

Concluding remarks

Take-home messages:

- Some provably hard parameter estimation problems become easy after ruling out “degenerate” cases.
- Algebraic structure of moments can be exploited using simple eigendecomposition techniques.

Concluding remarks

Take-home messages:

- Some provably hard parameter estimation problems become easy after ruling out "degenerate" cases.
- Algebraic structure of moments can be exploited using simple eigendecomposition techniques.

Some follow-up works (see arXiv reports):

Concluding remarks

Take-home messages:

- Some provably hard parameter estimation problems become easy after ruling out “degenerate” cases.
- Algebraic structure of moments can be exploited using simple eigendecomposition techniques.

Some follow-up works (see arXiv reports):

- Mixtures of (single-view) spherical Gaussians — non-degeneracy, without incoherence condition.

Concluding remarks

Take-home messages:

- Some provably hard parameter estimation problems become easy after ruling out “degenerate” cases.
- Algebraic structure of moments can be exploited using simple eigendecomposition techniques.

Some follow-up works (see arXiv reports):

- Mixtures of (single-view) spherical Gaussians — non-degeneracy, without incoherence condition.
- Latent Dirichlet Allocation (joint with **Dean Foster** and **Yi-Kai Liu**).

Concluding remarks

Take-home messages:

- Some provably hard parameter estimation problems become easy after ruling out "degenerate" cases.
- Algebraic structure of moments can be exploited using simple eigendecomposition techniques.

Some follow-up works (see arXiv reports):

- Mixtures of (single-view) spherical Gaussians — non-degeneracy, without incoherence condition.
- Latent Dirichlet Allocation (joint with **Dean Foster** and **Yi-Kai Liu**).
- Dynamic parsing models (joint with **Percy Liang**) — need a new trick to handle *unobserved* random tree structure (*e.g.*, PCFGs, dependency parsing trees).

Concluding remarks

Take-home messages:

- Some provably hard parameter estimation problems become easy after ruling out “degenerate” cases.
- Algebraic structure of moments can be exploited using simple eigendecomposition techniques.

Some follow-up works (see arXiv reports):

- Mixtures of (single-view) spherical Gaussians — non-degeneracy, without incoherence condition.
- Latent Dirichlet Allocation (joint with **Dean Foster** and **Yi-Kai Liu**).
- Dynamic parsing models (joint with **Percy Liang**) — need a new trick to handle *unobserved* random tree structure (*e.g.*, PCFGs, dependency parsing trees).

The end. Thanks!

5. Blank slide ————————————————

6. Connections to other moment methods

Connections to other moment methods

Basic recipe:

- Express moments of observable variables as system of polynomials in the desired parameters.
- Solve system of polynomials for desired parameters.

Connections to other moment methods

Basic recipe:

- Express moments of observable variables as system of polynomials in the desired parameters.
- Solve system of polynomials for desired parameters.

Pros:

- Very general technique; does not even require explicit specification of likelihood.
- Example: learn vertices of convex polytope from random samples (Gravin-Lassere-Pasechnik-Robins, ’12) — very powerful generalization of Prony’s method.

Connections to other moment methods

Basic recipe:

- Express moments of observable variables as system of polynomials in the desired parameters.
- Solve system of polynomials for desired parameters.

Pros:

- Very general technique; does not even require explicit specification of likelihood.
- Example: learn vertices of convex polytope from random samples (Gravin-Lassere-Pasechnik-Robins, '12) — very powerful generalization of Prony's method.

Cons:

- Typically require high-order moments, which are difficult to estimate.
- Computationally prohibitive to solve general systems of multivariate polynomials.

7. Moments

Simplified model and low-order moments

Simplification: $M_v \equiv M$ (same conditional means for all views);

Simplified model and low-order moments

Simplification: $M_v \equiv M$ (same conditional means for all views);

By conditional independence of $\vec{x}_1, \vec{x}_2, \vec{x}_3$ given $\vec{h}$,

$$\begin{aligned} \text{Pairs} &:= \mathbb{E}[\vec{x}_1 \otimes \vec{x}_2] \\ &= \mathbb{E}[(M\vec{h}) \otimes (M\vec{h})] \\ &= M \operatorname{diag}(\vec{w}) M^\top \end{aligned}$$

Simplified model and low-order moments

Simplification: $M_v \equiv M$ (same conditional means for all views);

By conditional independence of $\vec{x}_1, \vec{x}_2, \vec{x}_3$ given $\vec{h}$,

$$\begin{aligned} \text{Pairs} &:= \mathbb{E}[\vec{x}_1 \otimes \vec{x}_2] \\ &= \mathbb{E}[(M\vec{h}) \otimes (M\vec{h})] \\ &= M \,\text{diag}(\vec{w}) M^\top \end{aligned}$$

$$\begin{aligned} \text{Triples} &:= \mathbb{E}[\vec{x}_1 \otimes \vec{x}_2 \otimes \vec{x}_3] \\ &= \mathbb{E}[(M\vec{h}) \otimes (M\vec{h}) \otimes (M\vec{h})] \\ &= \mathbb{E}[\vec{h} \otimes \vec{h} \otimes \vec{h}](M, M, M) \end{aligned}$$

Simplified model and low-order moments

Simplification: $M_v \equiv M$ (same conditional means for all views);

By conditional independence of $\vec{x}_1, \vec{x}_2, \vec{x}_3$ given $\vec{h}$,

$$
\begin{aligned}
\text{Pairs} &:= \mathbb{E}[\vec{x}_1 \otimes \vec{x}_2] \\
&= \mathbb{E}[(M\vec{h}) \otimes (M\vec{h})] \\
&= M \operatorname{diag}(\vec{w}) M^\top
\end{aligned}
$$

$$
\begin{aligned}
\text{Triples} &:= \mathbb{E}[\vec{x}_1 \otimes \vec{x}_2 \otimes \vec{x}_3] \\
&= \mathbb{E}[(M\vec{h}) \otimes (M\vec{h}) \otimes (M\vec{h})] \\
&= \mathbb{E}[\vec{h} \otimes \vec{h} \otimes \vec{h}](M, M, M)
\end{aligned}
$$

$$
\begin{aligned}
\text{Triples}(\vec{\eta}) &:= \mathbb{E}[\langle \vec{\eta}, \vec{x}_1 \rangle (\vec{x}_2 \otimes \vec{x}_3)] \\
&= \mathbb{E}[\langle M^\top \vec{\eta}, \vec{h} \rangle ((M\vec{h}) \otimes (M\vec{h}))] \\
&= M \operatorname{diag}(M^\top \vec{\eta}) \operatorname{diag}(\vec{w}) M^\top .
\end{aligned}
$$

8. Symmetric plug-in estimator

Symmetric plug-in estimator

1. Obtain empirical estimates $\widehat{\text{Pairs}}$ and $\widehat{\text{Triples}}$ of Pairs and Triples.
2. Compute matrix of k orthonormal left singular vectors $\widehat{U}$ using rank-k SVD of $\widehat{\text{Pairs}}$;

$$W := \widehat{U}\left(\widehat{U}^\top \widehat{\text{Pairs}}\,\widehat{U}\right)^{-1/2}, \qquad B := \widehat{U}\left(\widehat{U}^\top \widehat{\text{Pairs}}\,\widehat{U}\right)^{1/2}.$$

3. Randomly pick unit vector $\vec{\theta} \in \mathbb{R}^k$.
4. Compute right eigenvectors $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_k$ of

$$\widehat{W}^\top \widehat{\text{Triples}}(\widehat{W}\vec{\theta})\widehat{W}$$

and return

$$\widehat{M} := [\widehat{B}\vec{v}_1 | \widehat{B}\vec{v}_2 | \cdots | \widehat{B}\vec{v}_k]$$

as conditional mean parameter estimates (up to scaling).

Symmetric plug-in estimator

Recall:

$$W := U\big(U^\top \mathsf{Pairs}\, U\big)^{-1/2}, \qquad B := U\big(U^\top \mathsf{Pairs}\, U\big)^{1/2}.$$

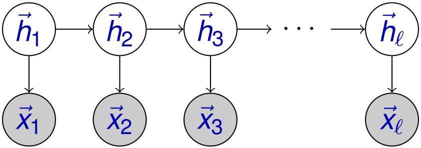
Then

$$\mathsf{Triples}(W, W, W) = \sum_{i=1}^{k} \lambda_i (\vec{v}_i \otimes \vec{v}_i \otimes \vec{v}_i)$$

where $[\vec{v}_1 | \vec{v}_2 | \cdots | \vec{v}_k] = (U^\top \mathsf{Pairs}\, U)^{-1/2} (U^\top M \operatorname{diag}(\vec{w})^{1/2})$ is orthogonal.
Therefore $B\vec{v}_i$ is i-th column of M scaled by $\sqrt{w_i}$.

9. Hidden Markov models

Hidden Markov models



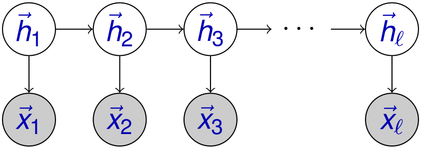
Parameters $(\vec{\pi}, T, O)$:

$$\Pr[\vec{h}_1 = \vec{e}_i] = \pi_i, \quad i \in [k]$$

$$\Pr[\vec{h}_{t+1} = \vec{e}_i | \vec{h}_t = \vec{e}_j] = T_{i,j}, \quad i, j \in [k]$$

$$\mathbb{E}[\vec{x}_t | \vec{h}_t = \vec{e}_j] = O\vec{e}_j, \quad j \in [k].$$

Hidden Markov models



Parameters $(\vec{\pi}, T, O)$:

$$\Pr[\vec{h}_1 = \vec{e}_i] = \pi_i, \quad i \in [k]$$
$$\Pr[\vec{h}_{t+1} = \vec{e}_i | \vec{h}_t = \vec{e}_j] = T_{i,j}, \quad i, j \in [k]$$
$$\mathbb{E}[\vec{x}_t | \vec{h}_t = \vec{e}_j] = O\vec{e}_j, \quad j \in [k].$$

As a latent class model:

$$\vec{w} := T\vec{\pi} \qquad M_1 := O \operatorname{diag}(\vec{\pi}) T^\top \operatorname{diag}(T\vec{\pi})^{-1}$$
$$M_2 := O \qquad M_3 := OT.$$

10. Comparison to HKZ

Comparison to previous spectral methods

- **Previous works** for estimating observable operator model for HMMs and other sequence / fixed-tree models (Hsu-Kakade-Zhang, '09; Langford-Salakhutdinov-Zhang, '09; Siddiqi-Boots-Gordon, '10; Song *et al*, '10; Foster *et al*, '11; Parikh *et al*, '11; Song *et al*, '11; Cohen *et al*, '12; Balle *et al*, '12; etc.)
  - Based on regression idea: best prediction of $\vec{x}_{t+1}$ given history $\vec{x}_{\leq t}$.
  - Observable operator model (Jaeger, '00) provides way to predict further ahead $\vec{x}_{t+1}, \vec{x}_{t+2}, \dots$.
- **This work**: Eigendecomposition method is rather different — looks for skewed directions using third-order moments. (Related to looking for kurtotic directions using fourth-order moments, like ICA.)
  - Can recover actual HMM parameters (transition and emission matrices).