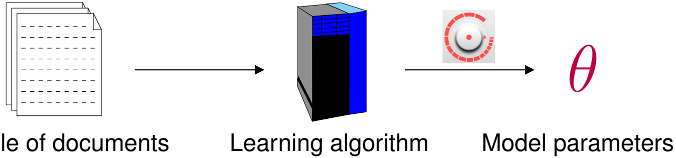
Method-of-moments

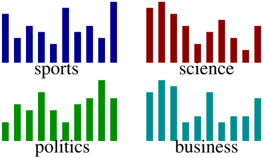
Daniel Hsu

Example: modeling the topics of a document corpus

Goal: model the topics of document in a corpus.



Topic model (*e.g.*, Hofmann, '99; Blei-Ng-Jordan, '03)

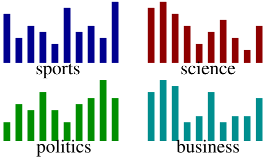


k topics (distributions over vocab words).

Each document $\leftrightarrow$ mixture of topics.

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

Topic model (*e.g.*, Hofmann, ’99; Blei-Ng-Jordan, ’03)



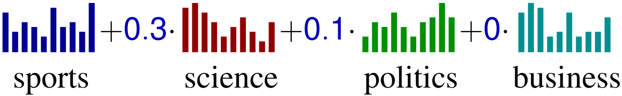
k topics (distributions over vocab words).

Each document $\leftrightarrow$ mixture of topics.

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

E.g.,

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



aardvark	0
athlete	3
	⋮
zygote	1

$\Pr_\theta[\text{“play”} \mid \text{sports}] = 0.0002$

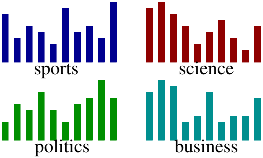
$\Pr_\theta[\text{“game”} \mid \text{sports}] = 0.0003$

$\Pr_\theta[\text{“season”} \mid \text{sports}] = 0.0001$

⋮

Learning topic models

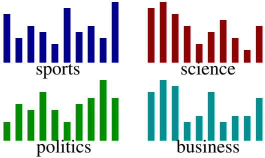
Topic model:



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

Learning topic models

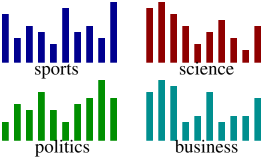
Simple topic model: (each document about *single* topic)



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

Learning topic models

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

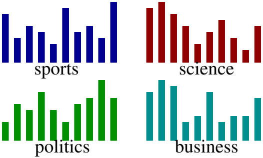


k topics (dists. over d words) $\vec{\mu}_1, \dots, \vec{\mu}_k$;
Topic t chosen with prob. w_t,
words in document $\sim_{\text{iid}} \vec{\mu}_t$.

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

Learning topic models

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



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

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

Some approaches to estimation

Some approaches to estimation

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

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

Some approaches to estimation

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

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

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

Some approaches to estimation

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

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

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

Method-of-moments (Pearson, 1894).

Find parameters θ that (approximately) *satisfy system of equations* based on the data.

Some approaches to estimation

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

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

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

Method-of-moments (Pearson, 1894).

Find parameters θ that (approximately) *satisfy system of equations* based on the data.

Many ways to instantiate & implement.

Some approaches to estimation

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

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

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

Method-of-moments (Pearson, 1894).

Find parameters θ that (approximately) *satisfy system of equations* based on the data.

Many ways to instantiate & implement.

Moments: normal distribution

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

First- and second-order moments:

$$\mathbb{E}_{(\mu,v)}[x] = \mu, \qquad \mathbb{E}_{(\mu,v)}[x^2] = \mu^2 + v.$$

Moments: normal distribution

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

First- and second-order moments:

$$\mathbb{E}_{(\mu,v)}[x] = \mu, \qquad \mathbb{E}_{(\mu,v)}[x^2] = \mu^2 + v.$$

Method-of-moments estimators of $\mu^\star$ and $v^\star$:
find $\hat{\mu}$ and $\hat{v}$ s.t.

$$\widehat{\mathbb{E}}_S[x] \approx \hat{\mu}, \qquad \widehat{\mathbb{E}}_S[x^2] \approx \hat{\mu}^2 + \hat{v}.$$

Moments: normal distribution

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

First- and second-order moments:

$$\mathbb{E}_{(\mu,v)}[x] = \mu, \qquad \mathbb{E}_{(\mu,v)}[x^2] = \mu^2 + v.$$

Method-of-moments estimators of $\mu^\star$ and $v^\star$:
find $\hat{\mu}$ and $\hat{v}$ s.t.

$$\widehat{\mathbb{E}}_S[x] \approx \hat{\mu}, \qquad \widehat{\mathbb{E}}_S[x^2] \approx \hat{\mu}^2 + \hat{v}.$$

A reasonable solution:

$$\hat{\mu} := \widehat{\mathbb{E}}_S[x], \qquad \hat{v} := \widehat{\mathbb{E}}_S[x^2] - \hat{\mu}^2$$

since $\widehat{\mathbb{E}}_S[x] \to \mathbb{E}_{(\mu^\star,v^\star)}[x]$ and $\widehat{\mathbb{E}}_S[x^2] \to \mathbb{E}_{(\mu^\star,v^\star)}[x^2]$ by LLN.

Moments: simple topic model

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

(Population) moments under some parameter θ:

$$\Pr_\theta\Big[\text{document contains words } i_1, i_2, \dots, i_n\Big].$$

e.g., $\Pr_\theta$[“machine” & “learning” co-occur].

Moments: simple topic model

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

(Population) moments under some parameter θ:

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

e.g., $\Pr_\theta$[“machine” & “learning” co-occur].

Empirical moments from sample S of documents:

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

i.e., empirical frequency of co-occurrences *in sample S*.

Method-of-moments

Method-of-moments strategy:

Given data sample S, find θ to satisfy system of equations

$$\text{moments}_\theta = \widehat{\text{moments}}_S.$$

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

Q1. Which moments should we use?

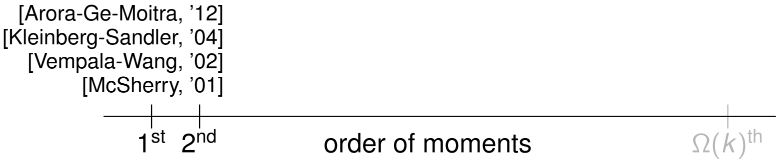
Q2. How do we (approx.) solve these moment equations?

Q1. Which moments should we use?

Q1. Which moments should we use?

moment order	reliable estimates?	unique solution?
1^{st}, 2^{nd}		

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



Q1. Which moments should we use?

moment order	reliable estimates?	unique solution?
1^{st}, 2^{nd}	✓	

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

- Fairly easy to get reliable estimates.

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

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

1^{st} 2^{nd} order of moments $\Omega(k)^{\text{th}}$

Q1. Which moments should we use?

moment order	reliable estimates?	unique solution?
1st, 2nd	✓	✗

1st- and 2nd-order moments (*e.g.*, prob. of word pairs).

- Fairly easy to get reliable estimates.

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

- Can have multiple solutions to moment equations.

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

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

1st 2nd order of moments $\Omega(k)^{\text{th}}$

Q1. Which moments should we use?

moment order	reliable estimates?	unique solution?
1st, 2nd	✓	✗
$\Omega(k)^{\text{th}}$		

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

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

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

1st 2nd order of moments $\Omega(k)^{\text{th}}$

Q1. Which moments should we use?

moment order	reliable estimates?	unique solution?
1^{st}, 2^{nd}	✓	✗
$\Omega(k)^{\text{th}}$		✓

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

- Uniquely pins down the solution.

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

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

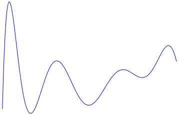
1^{st} 2^{nd} order of moments $\Omega(k)^{\text{th}}$

Q1. Which moments should we use?

moment order	reliable estimates?	unique solution?
1^{st}, 2^{nd}	✓	✗
$\Omega(k)^{th}$	✗	✓

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

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



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

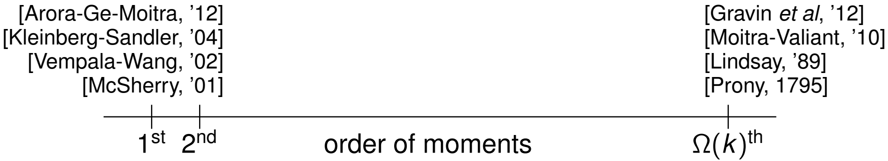
[Gravin *et al*, '12]
[Moitra-Valiant, '10]
[Lindsay, '89]
[Prony, 1795]

1^{st} 2^{nd} order of moments $\Omega(k)^{th}$

Q1. Which moments should we use?

moment order	reliable estimates?	unique solution?
1^{st}, 2^{nd}	✓	✗
$\Omega(k)^{th}$	✗	✓

Can we get best-of-both-worlds?

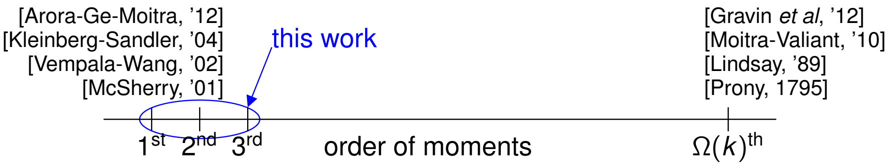


Q1. Which moments should we use?

moment order	reliable estimates?	unique solution?
1^{st}, 2^{nd}	✓	✗
$\Omega(k)^{\text{th}}$	✗	✓

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

In high-dimensions,
low-order multivariate moments suffice.
(1^{st}-, 2^{nd}-, and 3^{rd}-order moments)

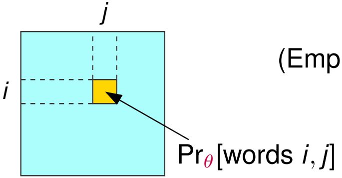


Low-order multivariate moments suffice

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

Low-order multivariate moments suffice

$P_\theta :=$



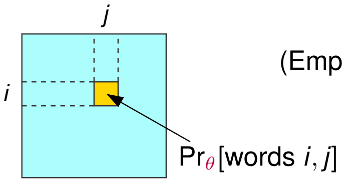
(Empirical: $\widehat{P}$)

Given a document about topic t,

$$\Pr_\theta[\text{ words } i, j \mid \text{topic } t\ ] = (\vec{\mu}_t)_i \cdot (\vec{\mu}_t)_j.$$

Low-order multivariate moments suffice

$P_\theta :=$



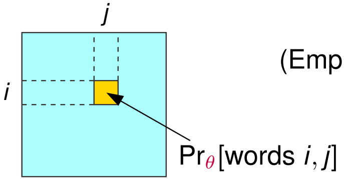
(Empirical: $\widehat{P}$)

Given a document about topic t,

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

Low-order multivariate moments suffice

$P_\theta :=$



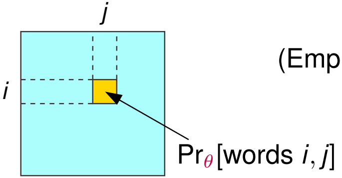
(Empirical: $\widehat{P}$)

Averaging over topics,

$$\Pr_\theta[\text{ words } i, j\,] = \sum_t w_t \cdot (\vec{\mu}_t \otimes \vec{\mu}_t)_{i,j}.$$

Low-order multivariate moments suffice

$P_\theta :=$

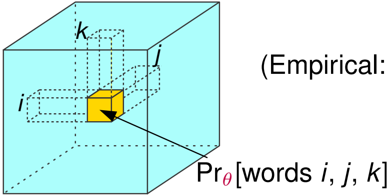


(Empirical: $\widehat{P}$)

In matrix notation P_θ,

$$P_\theta = \sum_t w_t \; \vec{\mu}_t \otimes \vec{\mu}_t .$$

Low-order multivariate moments suffice

$T_\theta :=$ 

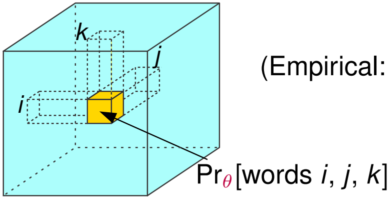
(Empirical: $\widehat{T}$)

Similarly,

$$\Pr_\theta[\text{ words } i, j, k \text{ }] = \sum_t w_t \cdot (\vec{\mu}_t \otimes \vec{\mu}_t \otimes \vec{\mu}_t)_{i,j,k}.$$

Low-order multivariate moments suffice

$T_\theta :=$

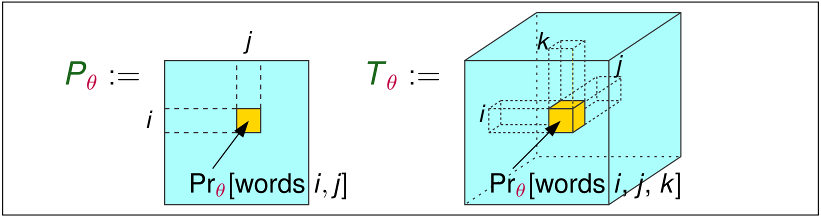


(Empirical: $\widehat{T}$)

In tensor notation T_θ,

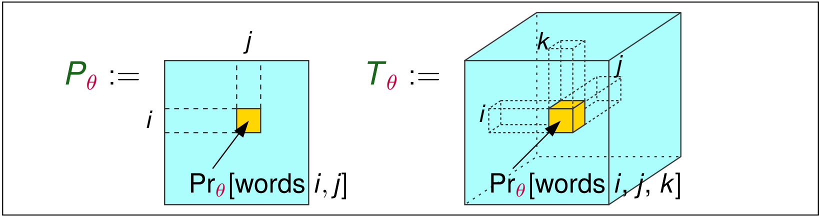
$$T_\theta = \sum_t w_t \; \vec{\mu}_t \otimes \vec{\mu}_t \otimes \vec{\mu}_t .$$

Low-order multivariate moments suffice



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

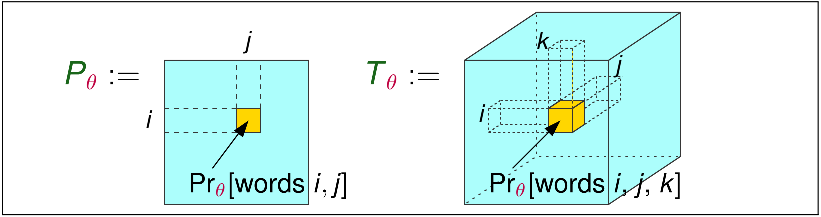
Low-order multivariate moments suffice



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

Low-rank matrix and tensor

Low-order multivariate moments suffice

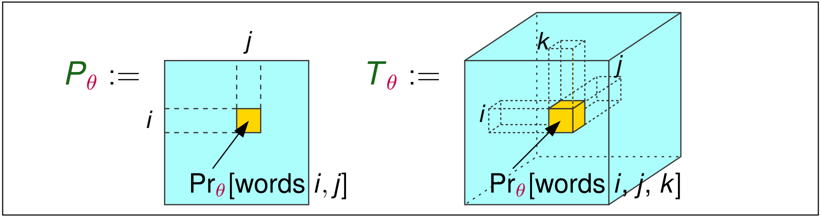


$$P_\theta = \sum_{t=1}^{k} w_t \, \vec{\mu}_t \otimes \vec{\mu}_t \quad \text{and} \quad T_\theta = \sum_{t=1}^{k} w_t \, \vec{\mu}_t \otimes \vec{\mu}_t \otimes \vec{\mu}_t$$

Moment equations: $P_\theta = \widehat{P}$, $T_\theta = \widehat{T}$

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

Low-order multivariate moments suffice



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

Moment equations: $P_\theta = \widehat{P}$, $T_\theta = \widehat{T}$

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

Claim: P_θ and T_θ uniquely determine the parameters θ.

Reduction to orthogonal case via whitening

$P_\theta = \sum_t w_t \, \vec{\mu}_t \otimes \vec{\mu}_t$ defines "whitened" coord. system.

Reduction to orthogonal case via whitening

$P_\theta = \sum_t w_t \, \vec{\mu}_t \otimes \vec{\mu}_t$ defines "whitened" coord. system.

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

$$T_\theta = \sum_{t=1}^{k} w_t \, \vec{\mu}_t \otimes \vec{\mu}_t \otimes \vec{\mu}_t \quad \longmapsto \quad B_\theta = \sum_{t=1}^{k} \lambda_t \, \vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t$$

where $\lambda_t = 1/\sqrt{w_t}, \quad \vec{v}_t = P_\theta^{-1/2} \left(\sqrt{w_t} \, \vec{\mu}_t \right)$.

Reduction to orthogonal case via whitening

$P_\theta = \sum_t w_t \, \vec{\mu}_t \otimes \vec{\mu}_t$ defines "whitened" coord. system.

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

$$T_\theta = \sum_{t=1}^{k} w_t \, \vec{\mu}_t \otimes \vec{\mu}_t \otimes \vec{\mu}_t \quad \longmapsto \quad B_\theta = \sum_{t=1}^{k} \lambda_t \, \vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t$$

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

Upshot: $\{\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_k\}$ are **orthonormal.**

Reduction to orthogonal case via whitening

$P_\theta = \sum_t w_t \, \vec{\mu}_t \otimes \vec{\mu}_t$ defines "whitened" coord. system.

"Whitened" third-order moment tensor B_θ has *orthogonal decomposition*

$$B_\theta = \sum_{t=1}^{k} \lambda_t \, \vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t .$$

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

Upshot: $\{\vec{v}_1, \vec{v}_2, \dots, \vec{v}_k\}$ are **orthonormal.**

Claim: *Orthogonal* decomposition of B_θ is unique.

The spectral theorem and eigendecompositions

The spectral theorem and eigendecompositions

Any symmetric matrix

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

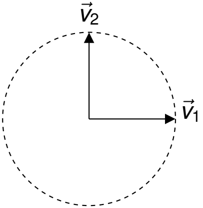
Decomposition is unique only if all eigenvalues λ_i are distinct.

The spectral theorem and eigendecompositions

Any symmetric matrix

$$A = \sum_{i=1}^{k} \lambda_i \; \vec{v}_i \otimes \vec{v}_i$$

Decomposition is unique only if all eigenvalues λ_i are distinct.

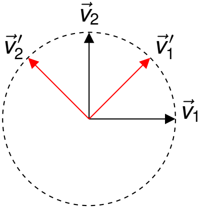


The spectral theorem and eigendecompositions

Any symmetric matrix

$$A = \sum_{i=1}^{k} \lambda_i \; \vec{v}_i \otimes \vec{v}_i$$

Decomposition is unique only if all eigenvalues λ_i are distinct.



The spectral theorem and eigendecompositions

Any symmetric matrix

$$A = \sum_{i=1}^{k} \lambda_i \; \vec{v}_i \otimes \vec{v}_i$$

Decomposition is unique only if all eigenvalues λ_i are distinct.

Special 3rd-order tensor

$$B = \sum_{i=1}^{k} \lambda_i \; \vec{v}_i \otimes \vec{v}_i \otimes \vec{v}_i$$

If decomposition exists, then it's always unique (even if λ_i all same).

The spectral theorem and eigendecompositions

Any symmetric matrix

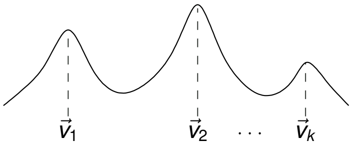
$$A = \sum_{i=1}^{k} \lambda_i \, \vec{v}_i \otimes \vec{v}_i$$

Decomposition is unique only if all eigenvalues λ_i are distinct.

Special 3rd-order tensor

$$B = \sum_{i=1}^{k} \lambda_i \, \vec{v}_i \otimes \vec{v}_i \otimes \vec{v}_i$$

If decomposition exists, then it's always unique (even if λ_i all same).

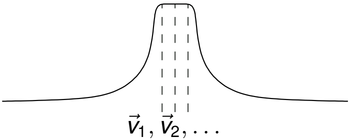


The spectral theorem and eigendecompositions

Any symmetric matrix

$$A = \sum_{i=1}^{k} \lambda_i \; \vec{v}_i \otimes \vec{v}_i$$

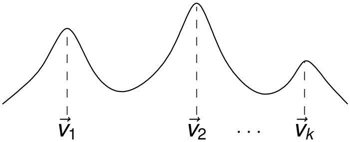
Decomposition is unique only if all eigenvalues λ_i are distinct.



Special 3rd-order tensor

$$B = \sum_{i=1}^{k} \lambda_i \; \vec{v}_i \otimes \vec{v}_i \otimes \vec{v}_i$$

If decomposition exists, then it's always unique (even if λ_i all same).



The spectral theorem and eigendecompositions

Any symmetric matrix

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

Decomposition is unique only if all eigenvalues λ_i are distinct.

Special 3rd-order tensor

$$B = \sum_{i=1}^{k} \lambda_i \, \vec{v}_i \otimes \vec{v}_i \otimes \vec{v}_i$$

If decomposition exists, then it's always unique (even if λ_i all same).

Uniqueness of orthogonal decomposition (+low-rank structure) implies that P_θ and T_θ uniquely determine θ.

The spectral theorem and eigendecompositions

Any symmetric matrix

$$A = \sum_{i=1}^{k} \lambda_i \; \vec{v}_i \otimes \vec{v}_i$$

Decomposition is unique only if all eigenvalues λ_i are distinct.

Special 3rd-order tensor

$$B = \sum_{i=1}^{k} \lambda_i \; \vec{v}_i \otimes \vec{v}_i \otimes \vec{v}_i$$

If decomposition exists, then it's always unique (even if λ_i all same).

Uniqueness of orthogonal decomposition (+low-rank structure) implies that P_θ and T_θ uniquely determine θ.

Q2. How to solve the moment equations?

Q2. How to solve the moment equations?

Solve moment equations via optimization problem

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

Q2. How to solve the moment equations?

Solve moment equations via optimization problem

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

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

Q2. How to solve the moment equations?

Solve moment equations via optimization problem

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

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

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

$$\min_{\lambda,\vec{v}} \|\lambda \vec{v} \otimes \vec{v} \otimes \vec{v} - \widehat{B}\|^2 \tag{$\ddagger$}$$

(after change-of-coord. system via $\widehat{P}$: $\widehat{T} \longrightarrow \widehat{B}$).

Q2. How to solve the moment equations?

Solve moment equations via optimization problem

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

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

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

$$\max_{\|\vec{u}\| \leq 1} \sum_{i,j,k} \widehat{B}_{i,j,k}\, u_i u_j u_k \qquad (\ddagger)$$

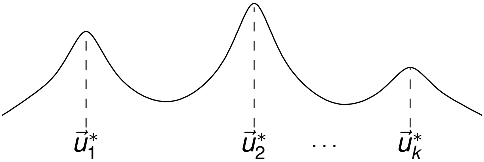
Q2. How to solve the moment equations?

Solve moment equations via optimization problem

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

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

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



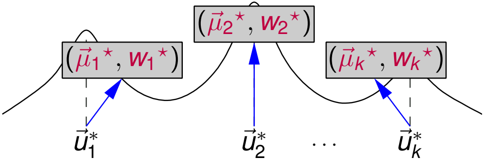
Q2. How to solve the moment equations?

Solve moment equations via optimization problem

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

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

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



Can approximate *all* local optima, each corresp. to a topic.
$\longrightarrow$ Near-optimal solution to $(\dagger)$. ∎

Variational argument

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

Variational argument

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

Lemma

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

$$\max_{\vec{u} \in \mathbb{R}^d} \; T_\theta(\vec{u}, \vec{u}, \vec{u}) \quad s.t. \quad P_\theta(\vec{u}, \vec{u}) \le 1 \tag{‡}$$

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

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

Variational argument

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

Lemma

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

$$\max_{\vec{u} \in \mathbb{R}^d} T_\theta(\vec{u}, \vec{u}, \vec{u}) \quad s.t. \quad P_\theta(\vec{u}, \vec{u}) \le 1 \tag{‡}$$

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

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

$\therefore \{(\vec{\mu}_i, w_i) : i \in [k]\}$ uniquely determined by P_θ and T_θ.

Implementation of topic model estimator

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

Implementation of topic model estimator

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

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

Implementation of topic model estimator

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

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

Implicit representation of $\widehat{T}$:

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

where $\vec{h} \in \mathbb{N}^d$ is (sparse) histogram vector for a document.

Implementation of topic model estimator

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

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

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

$$\widehat{T}(\vec{u}) \approx \frac{1}{|S|} \sum_{\vec{h} \in S} \left(\vec{h} \otimes \vec{h} \otimes \vec{h} \right)(\vec{u}) = \frac{1}{|S|} \sum_{\vec{h} \in S} (\vec{h}^\top \vec{u})^2 \, \vec{h}$$

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

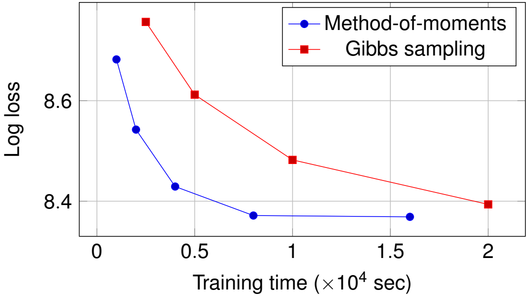
Illustrative empirical results

- Corpus: 300000 New York Times articles.
- Vocabulary size: 102660 words.
- Set number of topics $k := 50$.

Illustrative empirical results

- Corpus: 300000 New York Times articles.
- Vocabulary size: 102660 words.
- Set number of topics $k := 50$.

Predictive performance of straightforward implementation: $\approx$ 4–8$\times$ speed-up over Gibbs sampling.



Illustrative empirical results

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

Econ.	**Baseball**	**Edu.**	**Health care**	**Golf**
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

Illustrative empirical results

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

Invest.	**Election**	**auto race**	**Child's Lit.**	**Afghan War**
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

Illustrative empirical results

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

Web	Antitrust	TV	Movies	Music
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.

Recap

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

$$\text{moments}_\theta = \widehat{\text{moments}}_S.$$

Recap

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

$$\text{moments}_{\theta} = \widehat{\text{moments}}_{S}.$$

Q1. Which moments should we use?

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

Recap

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

$$\text{moments}_{\theta} = \widehat{\text{moments}}_{S}.$$

Q1. Which moments should we use?

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

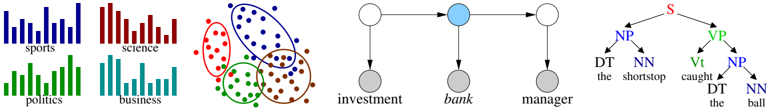
Q2. How do we (approx.) solve these moment equations?

Local optimization based on orthogonal tensor decompositions.

Structure in latent variable models

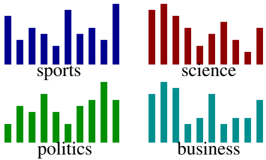
"Eigen-structure" found in low-order moments
for many other models of
high-dimensional data

$$\sum_{i=1}^{k} \lambda_i \; \vec{v}_i \otimes \vec{v}_i \otimes \vec{v}_i$$



Latent Dirichlet Allocation and Mixtures of Gaussians

Latent Dirichlet Allocation **(Blei-Ng-Jordan, '02)** **topic model**:



k topics (distributions over d words).

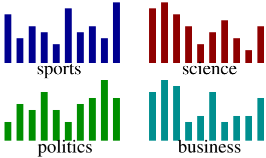
Each document $\leftrightarrow$ mixture of topics.

Doc.'s mixing weights $\sim$ Dirichlet($\vec{\alpha}$).

Words in doc. $\sim_{\text{iid}}$ mixture dist.

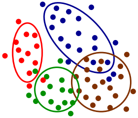
Latent Dirichlet Allocation and Mixtures of Gaussians

Latent Dirichlet Allocation (Blei-Ng-Jordan, '02) topic model:



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

Mixtures of Gaussians (Pearson, 1894)



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

Finding the relevant eigenstructure

In both LDA and mixtures of axis-aligned Gaussians:

$$f\left(\leq 2^{\text{nd}}\text{-order moments}_\theta\right) = \sum w_t \, \vec{\mu}_t \otimes \vec{\mu}_t$$
$$g\left(\leq 3^{\text{rd}}\text{-order moments}_\theta\right) = \sum w_t \, \vec{\mu}_t \otimes \vec{\mu}_t \otimes \vec{\mu}_t$$

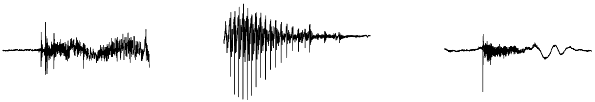
for suitable f and g based on additional model structure.

Hidden Markov Models (HMMs)

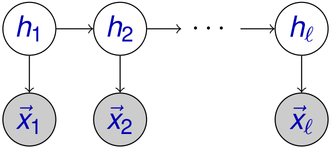
$$
\begin{tikzcd}
h_1 \arrow[r] \arrow[d] & h_2 \arrow[r] \arrow[d] & \cdots \arrow[r] & h_\ell \arrow[d] \\
\vec{x}_1 & \vec{x}_2 & & \vec{x}_\ell
\end{tikzcd}
$$

Workhorse statistical model for sequence data

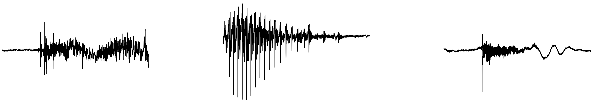
/k/ /a/ /t/



Hidden Markov Models (HMMs)



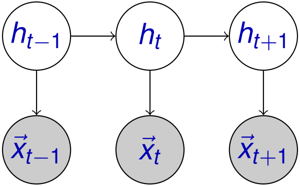
Workhorse statistical model for sequence data



- Hidden state variables $h_1 \to h_2 \to \cdots$ form a *Markov chain*.
- Observation $\vec{x}_t$ at time t depends only on hidden state h_t at time t.

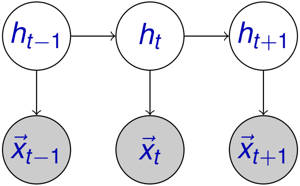
Learning HMMs

Correlations between past, present, and future



Learning HMMs

Correlations between past, present, and future



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

$$\mathbb{E}_\theta[\ \vec{x}_{t-1} \otimes \vec{x}_t \otimes \vec{x}_{t+1}\ ].$$

Where to read more

Tensor decompositions for learning latent variable models
A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, M. Telgarsky
Journal of Machine Learning Research, 2014.

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