Learning latent variable models using tensor decompositions

Daniel Hsu

Computer Science Department & Data Science Institute
Columbia University

Machine Learning Summer School
June 29-30, 2018

El tema (subject matter)

Learning algorithms

for **latent variable models**

based on **decompositions of moment tensors**.

El tema (subject matter)

Learning algorithms **(parameter estimation)**

for **latent variable models**

based on **decompositions of moment tensors**.

“Method-of-moments” **(Pearson, 1894)**

Example #1: summarizing a corpus of documents

Observation: **documents express one or more thematic topics.**

Politics Ensnare Mohamed Salah and Switzerland at the World Cup

By Rory Smith, James Montague and Tariq Panja

June 24, 2018

MOSCOW — The World Cup was thrust into the combustible mix of politics and soccer — dangerous ground that world soccer takes great pains to avoid — as a growing number of disciplinary proceedings and a star player's threatened retirement brought several sensitive international flash points to the tournament's doorstep this weekend.

Example #1: summarizing a corpus of documents

Observation: **documents express one or more thematic topics.**

Politics Ensnare Mohamed Salah and Switzerland at the World Cup

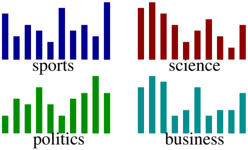
By Rory Smith, James Montague and Tariq Panja

June 24, 2018

MOSCOW — The World Cup was thrust into the combustible mix of politics and soccer — dangerous ground that world soccer takes great pains to avoid — as a growing number of disciplinary proceedings and a star player's threatened retirement brought several sensitive international flash points to the tournament's doorstep this weekend.

- What topics are expressed in a corpus of documents?
- How prevalent is each topic in the corpus?

Topic model (e.g., latent Dirichlet allocation)

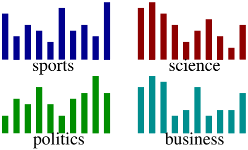


K topics (distributions over vocab words).

Document $\equiv$ mixture of topics.

Word tokens in doc. $\overset{\text{iid}}{\sim}$ mixture distribution.

Topic model (e.g., latent Dirichlet allocation)



K topics (distributions over vocab words).

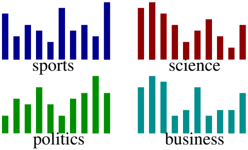
Document $\equiv$ mixture of topics.

Word tokens in doc. $\overset{\text{iid}}{\sim}$ mixture distribution.

E.g.,

$$\overset{\text{iid}}{\sim} \quad 0.7 \times \boldsymbol{P}_{\text{sports}} + 0.3 \times \boldsymbol{P}_{\text{politics}} .$$

Topic model (e.g., latent Dirichlet allocation)



K topics (distributions over vocab words).

Document $\equiv$ mixture of topics.

Word tokens in doc. $\overset{\text{iid}}{\sim}$ mixture distribution.

E.g.,

$$\overset{\text{iid}}{\sim} \quad 0.7 \times \boldsymbol{P}_{\text{sports}} + 0.3 \times \boldsymbol{P}_{\text{politics}}.$$

Given corpus of documents (and "hyper-parameters", e.g., K), produce estimates of **model parameters**, e.g.:

- Distribution $\boldsymbol{P}_t$ over vocab words, for each $t \in [K]$.
- Weight w_t of topic t in document corpus, for each $t \in [K]$.

Labels / annotations

- Suppose each word token x in document is *annotated* with source topic $t_x \in \{1, 2, \ldots, K\}$.

Politics	Ensnare	Mohamed_Salah	and	Switzerland	at
3	3	1	5	3	5

Labels / annotations

- Suppose each word token x in document is *annotated* with source topic $t_x \in \{1, 2, \ldots, K\}$.

Politics	Ensnare	Mohamed_Salah	and	Switzerland	at
3	3	1	5	3	5

Then estimating the $\{(\boldsymbol{P}_t, w_t)\}_{t=1}^{K}$ can be done “directly”.

Labels / annotations

- Suppose each word token x in document is *annotated* with source topic $t_x \in \{1, 2, \ldots, K\}$.

Politics	Ensnare	Mohamed_Salah	and	Switzerland	at
3	3	1	5	3	5

  Then estimating the $\{(\boldsymbol{P}_t, w_t)\}_{t=1}^{K}$ can be done "directly".

- Unfortunately, we often don't have such annotations (i.e., data are *unlabeled* / topics are *hidden*).

  "Direct" approach to estimation unavailable.

Example #2: subpopulations in data

Data studied by Pearson (1894):

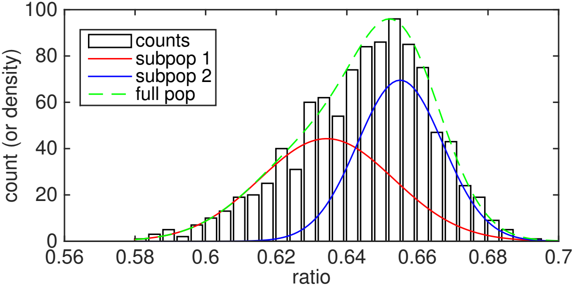
ratio of forehead-width to body-length for 1000 crabs.

Example #2: subpopulations in data

Data studied by Pearson (1894):

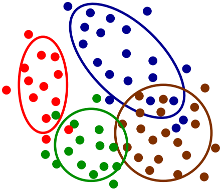
ratio of forehead-width to body-length for 1000 crabs.

Sample may be comprised of different sub-species of crabs.



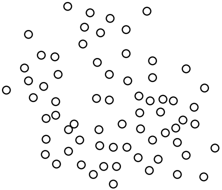
Gaussian mixture model

$$\begin{aligned} H &\sim \text{Categorical}(\pi_1, \pi_2, \ldots, \pi_K)\,; \\ \boldsymbol{X} \mid H = t &\sim \text{Normal}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)\,, \quad t \in [K]\,. \end{aligned}$$



Gaussian mixture model

$$\begin{aligned} H &\sim \text{Categorical}(\pi_1, \pi_2, \ldots, \pi_K)\,; \\ \boldsymbol{X} \mid H = t &\sim \text{Normal}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)\,, \quad t \in [K]\,. \end{aligned}$$



Estimate **mean vector**, **covariance matrix**, and **mixing weight** of each subpopulation from *unlabeled data*.

Maximum likelihood estimation

- No “direct” estimators when some variables are hidden.

Maximum likelihood estimation

- No "direct" estimators when some variables are hidden.
- **Maximum likelihood estimator** (MLE):

$$\boldsymbol{\theta}_{\mathsf{MLE}} := \arg\max_{\boldsymbol{\theta} \in \Theta} \log \Pr_{\boldsymbol{\theta}}(\mathsf{data}) .$$

Maximum likelihood estimation

- No “direct” estimators when some variables are hidden.
- **Maximum likelihood estimator** (MLE):

$$\boldsymbol{\theta}_{\mathsf{MLE}} := \operatorname*{arg\,max}_{\boldsymbol{\theta} \in \Theta} \log \Pr_{\boldsymbol{\theta}}(\mathsf{data}) .$$

    - **Note**: log-likelihood is not necessarily concave function of $\boldsymbol{\theta}$.

Maximum likelihood estimation

- No "direct" estimators when some variables are hidden.
- **Maximum likelihood estimator** (MLE):

$$\boldsymbol{\theta}_{\mathsf{MLE}} := \arg\max_{\boldsymbol{\theta}\in\Theta} \log \Pr_{\boldsymbol{\theta}}(\mathsf{data}) .$$

  - **Note**: log-likelihood is not necessarily concave function of $\boldsymbol{\theta}$.
  - For latent variable models, often use local optimization, most notably via Expectation-Maximization (EM) (Dempster, Laird, & Rubin, 1977).

MLE for Gaussian mixture models

Given data $\{\boldsymbol{x}_i\}_{i=1}^{n}$, find $\{(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, \pi_t)\}_{t=1}^{K}$ to maximize

$$\sum_{i=1}^{n} \log \left(\sum_{t=1}^{K} \pi_t \cdot \frac{1}{\det(\boldsymbol{\Sigma}_t)^{1/2}} \exp \left\{ -\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_t)^{\top} \boldsymbol{\Sigma}_t^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_t) \right\} \right) .$$

MLE for Gaussian mixture models

Given data $\{\boldsymbol{x}_i\}_{i=1}^{n}$, find $\{(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, \pi_t)\}_{t=1}^{K}$ to maximize

$$\sum_{i=1}^{n} \log \left(\sum_{t=1}^{K} \pi_t \cdot \frac{1}{\det(\boldsymbol{\Sigma}_t)^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_t)^\top \boldsymbol{\Sigma}_t^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_t) \right\} \right) .$$

- Sensible with restrictions on $\boldsymbol{\Sigma}_t$ (e.g., $\boldsymbol{\Sigma}_t \succeq \sigma^2 \boldsymbol{I}$).

MLE for Gaussian mixture models

Given data $\{\boldsymbol{x}_i\}_{i=1}^n$, find $\{(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, \pi_t)\}_{t=1}^K$ to maximize

$$\sum_{i=1}^n \log \left(\sum_{t=1}^K \pi_t \cdot \frac{1}{\det(\boldsymbol{\Sigma}_t)^{1/2}} \exp \left\{ -\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_t)^\top \boldsymbol{\Sigma}_t^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_t) \right\} \right) .$$

- Sensible with restrictions on $\boldsymbol{\Sigma}_t$ (e.g., $\boldsymbol{\Sigma}_t \succeq \sigma^2 \boldsymbol{I}$).
- But NP-hard to maximize (Tosh and Dasgupta, 2018):
  Can't expect efficient algorithms to work for all data sets.

Parameter learning objective

Suppose iid sample of size n is generated by distribution from model with (unknown) parameters $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$. ($p = \#$ params)

Parameter learning objective

Suppose iid sample of size n is generated by distribution from model with (unknown) parameters $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$.   ($p = \#$ params)

Task: Produce estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ such that

$$\mathbb{E}\,\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| \;\to\; 0 \quad \text{as} \quad n \;\to\; \infty$$

(i.e., $\hat{\boldsymbol{\theta}}$ is *consistent*).

Parameter learning objective

Suppose iid sample of size n is generated by distribution from model with (unknown) parameters $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$. $(p = \#$ params)

Task: Produce estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ such that

$$\mathbb{E}\,\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| \;\to\; 0 \quad \text{as} \quad n \;\to\; \infty$$

(i.e., $\hat{\boldsymbol{\theta}}$ is *consistent*).

- E.g., for spherical Gaussian mixtures:
  - For $K = 2$ (and $\pi_t = 1/2$, $\boldsymbol{\Sigma}_t = \boldsymbol{I}$): EM is consistent (Xu, H., & Maleki, 2016).

Parameter learning objective

Suppose iid sample of size n is generated by distribution from model with (unknown) parameters $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$. $(p = \#$ params)

Task: Produce estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ such that

$$\mathbb{E}\,\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| \;\to\; 0 \quad \text{as} \quad n \;\to\; \infty$$

(i.e., $\hat{\boldsymbol{\theta}}$ is *consistent*).

- E.g., for spherical Gaussian mixtures:
  - For $K = 2$ (and $\pi_t = 1/2$, $\boldsymbol{\Sigma}_t = \boldsymbol{I}$): EM is consistent (Xu, H., & Maleki, 2016).
  - Larger K: easily trapped in local maxima, far from global max (Jin, Zhang, Balakrishnan, Wainwright, & Jordan, 2016).

Parameter learning objective

Suppose iid sample of size n is generated by distribution from model with (unknown) parameters $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$. $(p = \#$ params)

Task: Produce estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ such that

$$\mathbb{E}\,\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| \;\to\; 0 \quad \text{as} \quad n \;\to\; \infty$$

(i.e., $\hat{\boldsymbol{\theta}}$ is *consistent*).

- E.g., for spherical Gaussian mixtures:
    - For $K = 2$ (and $\pi_t = 1/2$, $\boldsymbol{\Sigma}_t = \boldsymbol{I}$): EM is consistent (Xu, H., & Maleki, 2016).
    - Larger K: easily trapped in local maxima, far from global max (Jin, Zhang, Balakrishnan, Wainwright, & Jordan, 2016).

      Practitioners often use EM with many (random) restarts . . . but may take a long time to get near the global max.

Parameter learning objective

Suppose iid sample of size n is generated by distribution from model with (unknown) parameters $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$. ($p = \#$ params)

Task: Produce estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ such that

$$\Pr\left(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| \leq \epsilon\right) \ \geq \ 1 - \delta$$

with $\operatorname{poly}(p, 1/\epsilon, 1/\delta, \dots)$ sample size and running time.

- E.g., for spherical Gaussian mixtures:
  - For $K = 2$ (and $\pi_t = 1/2$, $\boldsymbol{\Sigma}_t = \boldsymbol{I}$): EM is consistent (Xu, H., & Maleki, 2016).
  - Larger K: easily trapped in local maxima, far from global max (Jin, Zhang, Balakrishnan, Wainwright, & Jordan, 2016).

    Practitioners often use EM with many (random) restarts ... but may take a long time to get near the global max.

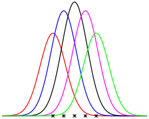
Barriers

Hard to learn model parameters,
even when data is generated by a model distribution.

Barriers

Hard to learn model parameters,
even when data is generated by a model distribution.

Cryptographic hardness
(e.g., Mossel & Roch, 2006)



Information-theoretic hardness
(e.g., Moitra & Valiant, 2010)

May require $2^{\Omega(K)}$ running time or $2^{\Omega(K)}$ sample size.

Ways around the barriers

- **Separation conditions**.

  E.g., assume mixture component distributions are far apart.
  (Dasgupta, 1999; Arora & Kannan, 2001; Vempala & Wang, 2002; . . .)

Ways around the barriers

- **Separation conditions**.

  E.g., assume mixture component distributions are far apart.
  (Dasgupta, 1999; Arora & Kannan, 2001; Vempala & Wang, 2002; . . .)

- **Structural assumptions**.

  E.g., sparsity, anchor words.
  (Spielman, Wang, & Wright, 2012; Arora, Ge, & Moitra, 2012; . . .)

Ways around the barriers

- **Separation conditions**.

  E.g., assume mixture component distributions are far apart.

  (Dasgupta, 1999; Arora & Kannan, 2001; Vempala & Wang, 2002; . . .)

- **Structural assumptions**.

  E.g., sparsity, anchor words.

  (Spielman, Wang, & Wright, 2012; Arora, Ge, & Moitra, 2012; . . .)

- **Non-degeneracy conditions**.

  E.g., assume $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_K$ are in general position.

Ways around the barriers

- **Separation conditions**.

  E.g., assume mixture component distributions are far apart.
  (Dasgupta, 1999; Arora & Kannan, 2001; Vempala & Wang, 2002; ...)

- **Structural assumptions**.

  E.g., sparsity, anchor words.
  (Spielman, Wang, & Wright, 2012; Arora, Ge, & Moitra, 2012; ...)

- **Non-degeneracy conditions**.

  E.g., assume $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K$ are in general position.

This lecture: learning algorithms for non-degenerate instances via *method-of-moments*.

Method-of-moments at a glance

1. Determine function of model parameters $\boldsymbol{\theta}$ estimatable from observable data:

$$\mathbb{E}_{\boldsymbol{\theta}}[f(\boldsymbol{X})] \qquad (\text{“moments”})\,.$$

2. Form estimates of moments using data (e.g., iid sample):

$$\widehat{\mathbb{E}}[f(\boldsymbol{X})] \qquad (\text{“empirical moments”})\,.$$

3. Approximately solve equations for parameters $\boldsymbol{\theta}$:

$$\mathbb{E}_{\boldsymbol{\theta}}[f(\boldsymbol{X})] \;=\; \widehat{\mathbb{E}}[f(\boldsymbol{X})]\,.$$

4. (“Fine-tune” estimated parameters with local optimization.)

Method-of-moments at a glance

1. Determine function of model parameters $\boldsymbol{\theta}$ estimatable from observable data:

$$\mathbb{E}_{\boldsymbol{\theta}}[f(\boldsymbol{X})] \qquad (\text{“moments”})\,.$$

Which moments?

2. Form estimates of moments using data (e.g., iid sample):

$$\widehat{\mathbb{E}}[f(\boldsymbol{X})] \qquad (\text{“empirical moments”})\,.$$

3. Approximately solve equations for parameters $\boldsymbol{\theta}$:

$$\mathbb{E}_{\boldsymbol{\theta}}[f(\boldsymbol{X})] \;=\; \widehat{\mathbb{E}}[f(\boldsymbol{X})]\,.$$

How?

4. (“Fine-tune” estimated parameters with local optimization.)

Method-of-moments at a glance

1. Determine function of model parameters $\boldsymbol{\theta}$ estimatable from observable data:

$$\mathbb{E}_{\boldsymbol{\theta}}[f(\boldsymbol{X})] \qquad (\text{“moments”})\,.$$

Which moments? Often low-order moments suffice.

2. Form estimates of moments using data (e.g., iid sample):

$$\widehat{\mathbb{E}}[f(\boldsymbol{X})] \qquad (\text{“empirical moments”})\,.$$

3. Approximately solve equations for parameters $\boldsymbol{\theta}$:

$$\mathbb{E}_{\boldsymbol{\theta}}[f(\boldsymbol{X})] \;=\; \widehat{\mathbb{E}}[f(\boldsymbol{X})]\,.$$

How? Algorithms for tensor decomposition.

4. (“Fine-tune” estimated parameters with local optimization.)

A simple example of the method-of-moments

Let $X \sim \mathrm{Normal}(\mu, \sigma^2)$. How to estimate σ^2 from iid sample?

A simple example of the method-of-moments

Let $X \sim \mathrm{Normal}(\mu, \sigma^2)$. How to estimate σ^2 from iid sample?

- Consider first- and second-moments: $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$.

A simple example of the method-of-moments

Let $X \sim \mathrm{Normal}(\mu, \sigma^2)$. How to estimate σ^2 from iid sample?

- Consider first- and second-moments: $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$.
- Formula for σ^2 in terms of moments:

$$\mathbb{E}[X^2] - \mathbb{E}[X]^2 \;=\; \left(\sigma^2 + \mu^2\right) - \mu^2 \;=\; \sigma^2 \,.$$

A simple example of the method-of-moments

Let $X \sim \mathrm{Normal}(\mu, \sigma^2)$. How to estimate σ^2 from iid sample?

- Consider first- and second-moments: $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$.
- Formula for σ^2 in terms of moments:

$$\mathbb{E}[X^2] - \mathbb{E}[X]^2 \;=\; \left(\sigma^2 + \mu^2\right) - \mu^2 \;=\; \sigma^2\,.$$

- Form *estimates* of $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ from iid sample $\{x_i\}_{i=1}^n$: e.g.,

$$\widehat{\mathbb{E}}[X] \;:=\; \frac{1}{n}\sum_{i=1}^n x_i\,, \qquad \widehat{\mathbb{E}}[X^2] \;:=\; \frac{1}{n}\sum_{i=1}^n x_i^2\,.$$

A simple example of the method-of-moments

Let $X \sim \mathrm{Normal}(\mu, \sigma^2)$. How to estimate σ^2 from iid sample?

- Consider first- and second-moments: $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$.
- Formula for σ^2 in terms of moments:

$$\mathbb{E}[X^2] - \mathbb{E}[X]^2 \;=\; \left(\sigma^2 + \mu^2\right) - \mu^2 \;=\; \sigma^2 \,.$$

- Form *estimates* of $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ from iid sample $\{x_i\}_{i=1}^n$: e.g.,

$$\widehat{\mathbb{E}}[X] \;:=\; \frac{1}{n}\sum_{i=1}^n x_i \,, \qquad \widehat{\mathbb{E}}[X^2] \;:=\; \frac{1}{n}\sum_{i=1}^n x_i^2 \,.$$

- Then estimate σ^2 with

$$\hat{\sigma}^2 \;:=\; \widehat{\mathbb{E}}[X^2] - \widehat{\mathbb{E}}[X]^2 \,.$$

A simple example of the method-of-moments

Let $X \sim \mathrm{Normal}(\mu, \sigma^2)$. How to estimate σ^2 from iid sample?

- Consider first- and second-moments: $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$.
- Formula for σ^2 in terms of moments:

$$\mathbb{E}[X^2] - \mathbb{E}[X]^2 \;=\; \left(\sigma^2 + \mu^2\right) - \mu^2 \;=\; \sigma^2 \,.$$

- Form *estimates* of $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ from iid sample $\{x_i\}_{i=1}^n$: e.g.,

$$\widehat{\mathbb{E}}[X] \;:=\; \frac{1}{n}\sum_{i=1}^n x_i \,, \qquad \widehat{\mathbb{E}}[X^2] \;:=\; \frac{1}{n}\sum_{i=1}^n x_i^2 \,.$$

- Then estimate σ^2 with

$$\hat{\sigma}^2 \;:=\; \widehat{\mathbb{E}}[X^2] - \widehat{\mathbb{E}}[X]^2 \,.$$

We'll follow this same basic recipe for much richer models!

Outline

1. Topic model for single-topic documents.
    - Identifiability.
    - Parameter recovery via orthogonal tensor decomposition.
2. Moment decompositions for other models.
    - Mixtures of Gaussians and linear regressions.
    - Multi-view models (e.g., HMMs).
    - Other models (e.g., single-index models).
3. Error analysis.

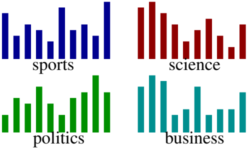
Other models amenable to moment tensor decomposition

- Models for independent components analysis (Comon, 1994; Frieze, Jerrum, & Kannan, 1996; Arora, Ge, Moitra & Sachdeva, 2012; Anandkumar, Foster, H., Kakade, & Liu, 2012, 2015; Belkin, Rademacher, & Voss, 2013; etc.)
- Latent Dirichlet Allocation (Anandkumar, Foster, H., Kakade, & Liu, 2012, 2015; Anderson, Goyal, & Rademacher, 2013)
- Mixed-membership stochastic blockmodels (Anandkumar, Ge, H., & Kakade, 2013, 2014)
- Simple probabilistic grammars (H., Kakade, & Liang, 2012)
- Noisy-or networks (Halpern & Sontag, 2013; Jernite, Halpern & Sontag, 2013; Arora, Ge, Ma, & Risteski, 2016)
- Indian buffet process (Tung & Smola, 2014)
- Mixed multinomial logit model (Oh & Shah, 2014)
- Dawid-Skene model (Zhang, Chen, Zhou, & Jordan, 2014)
- Multi-task bandits (Azar, Lazaric, & Brunskill, 2013)
- Partially obs. MDPs (Azizzadenesheli, Lazaric, & Anandkumar, 2016)
- . . .

1. Topic model for single-topic documents

Topic model

General topic model (e.g., Latent Dirichlet Allocation)



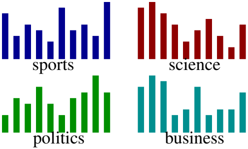
K topics (dists. over words) $\{\boldsymbol{P}_t\}_{t=1}^{K}$.

Document $\equiv$ mixture of topics (**hidden**).

Word tokens in doc. $\overset{\text{iid}}{\sim}$ mixture distribution.

Topic model

Topic model for single-topic documents



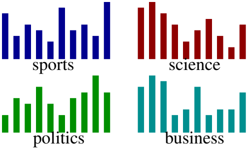
K topics (dists. over words) $\{\boldsymbol{P}_t\}_{t=1}^{K}$.

Pick topic t with prob. w_t (**hidden**).

Word tokens in doc. $\overset{\text{iid}}{\sim}$ $\boldsymbol{P}_t$.

Topic model

Topic model for single-topic documents



K topics (dists. over words) $\{\boldsymbol{P}_t\}_{t=1}^{K}$.

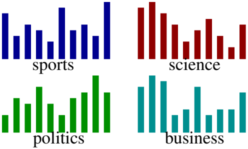
Pick topic t with prob. w_t (**hidden**).

Word tokens in doc. $\overset{\text{iid}}{\sim}$ $\boldsymbol{P}_t$.

Given iid sample of documents of length L,
produce estimates of **model parameters** $\{(\boldsymbol{P}_t, w_t)\}_{t=1}^{K}$.

Topic model

Topic model for single-topic documents



K topics (dists. over words) $\{\boldsymbol{P}_t\}_{t=1}^{K}$.

Pick topic t with prob. w_t (**hidden**).

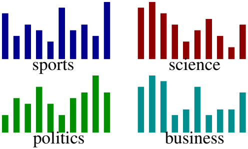
Word tokens in doc. $\overset{\text{iid}}{\sim}$ $\boldsymbol{P}_t$.

Given iid sample of documents of length L,
produce estimates of **model parameters** $\{(\boldsymbol{P}_t, w_t)\}_{t=1}^{K}$.

How long must the documents be?

Topic model

Topic model for single-topic documents



K topics (dists. over words) $\{\boldsymbol{P}_t\}_{t=1}^{K}$.
Pick topic t with prob. w_t (**hidden**).
Word tokens in doc. $\overset{\text{iid}}{\sim}$ $\boldsymbol{P}_t$.

Given iid sample of documents of length L,
produce estimates of **model parameters** $\{(\boldsymbol{P}_t, w_t)\}_{t=1}^{K}$.

How long must the documents be?

(Answering this question leads to efficient algorithms for estimating parameters!)

Identifiability

Generative process:
Pick $t \sim \mathrm{Categorical}(w_1, w_2, \ldots, w_K)$.
Given t, pick L words from $\boldsymbol{P}_t$.

Identifiability

Generative process:
Pick $t \sim \mathrm{Categorical}(w_1, w_2, \ldots, w_K)$.
Given t, pick L words from $\boldsymbol{P}_t$.

- $L = 1$: random document (single word) $\sim \sum_{t=1}^{K} w_t \boldsymbol{P}_t$.

  Are parameters $\{(\boldsymbol{P}_t, w_t)\}_{t=1}^{K}$ identifiable from single-word documents?

Identifiability

Generative process:
Pick $t \sim \mathrm{Categorical}(w_1, w_2, \ldots, w_K)$.
Given t, pick L words from $\boldsymbol{P}_t$.

- $L = 1$: random document (single word) $\sim \sum_{t=1}^{K} w_t \boldsymbol{P}_t$.

  Are parameters $\{(\boldsymbol{P}_t, w_t)\}_{t=1}^{K}$ identifiable from single-word documents?

  No.

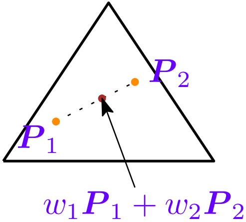
Identifiability

Generative process:
Pick $t \sim \mathrm{Categorical}(w_1, w_2, \ldots, w_K)$.
Given t, pick L words from $\boldsymbol{P}_t$.

- $L = 1$: random document (single word) $\sim \sum_{t=1}^{K} w_t \boldsymbol{P}_t$.

  Are parameters $\{(\boldsymbol{P}_t, w_t)\}_{t=1}^{K}$ identifiable from single-word documents?

  No.



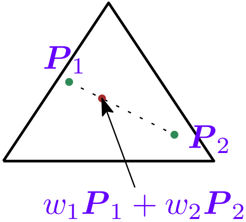
Identifiability

Generative process:
Pick $t \sim \mathrm{Categorical}(w_1, w_2, \ldots, w_K)$.
Given t, pick L words from $\boldsymbol{P}_t$.

- $L = 1$: random document (single word) $\sim \sum_{t=1}^{K} w_t \boldsymbol{P}_t$.

  Are parameters $\{(\boldsymbol{P}_t, w_t)\}_{t=1}^{K}$ identifiable from single-word documents?

  No.



Identifiability: $L = 2$

Generative process:
Pick $t \sim \mathrm{Categorical}(w_1, w_2, \ldots, w_K)$.
Given t, pick L words from $\boldsymbol{P}_t$.

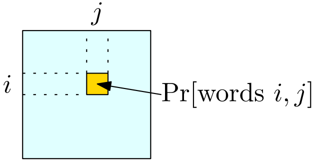
- $L = 2$:

Identifiability: $L = 2$

Generative process:
Pick $t \sim \mathrm{Categorical}(w_1, w_2, \ldots, w_K)$.
Given t, pick L words from $\boldsymbol{P}_t$.

- $L = 2$:
  Regard $\boldsymbol{P}_t$ as probability vector (ith entry of $\boldsymbol{P}_t$ is $\Pr[\text{word } i]$).
  Joint distribution of word pairs (for topic t) is given by matrix:

  $\boldsymbol{P}_t \boldsymbol{P}_t^\top =$

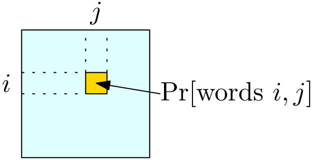
  Random document $\sim \sum_{t=1}^{K} w_t \boldsymbol{P}_t \boldsymbol{P}_t^\top$.

Identifiability: $L = 2$

Generative process:
Pick $t \sim \mathrm{Categorical}(w_1, w_2, \dots, w_K)$.
Given t, pick L words from $\boldsymbol{P}_t$.

- $L = 2$:
  Regard $\boldsymbol{P}_t$ as probability vector (ith entry of $\boldsymbol{P}_t$ is $\Pr[\text{word } i]$).
  Joint distribution of word pairs (for topic t) is given by matrix:

$\boldsymbol{P}_t \boldsymbol{P}_t^\top =$



Random document $\sim \sum_{t=1}^{K} w_t \boldsymbol{P}_t \boldsymbol{P}_t^\top$.

Are parameters $\{(\boldsymbol{P}_t, w_t)\}_{t=1}^{K}$ identifiable from word pairs?

Simple observation

Suppose distribution of word pairs (as a matrix) can be written as

$$\boldsymbol{M} \ = \ \boldsymbol{A}\boldsymbol{A}^\top .$$

Simple observation

Suppose distribution of word pairs (as a matrix) can be written as

$$\boldsymbol{M} \;=\; \boldsymbol{A}\boldsymbol{A}^{\top}.$$

Then it can also be written as

$$\boldsymbol{M} \;=\; (\boldsymbol{A}\boldsymbol{R})(\boldsymbol{A}\boldsymbol{R})^{\top}$$

for any orthogonal matrix $\boldsymbol{R}$ (because $\boldsymbol{R}^{\top}\boldsymbol{R} = \boldsymbol{I}$).

Identifiability: $L = 2$ counterexample

Parameters $\{(\boldsymbol{P}_1, w_1), (\boldsymbol{P}_2, w_2)\}$ and $\{(\widetilde{\boldsymbol{P}}_1, \tilde{w}_1), (\widetilde{\boldsymbol{P}}_2, \tilde{w}_2)\}$

$$\begin{aligned}
(\boldsymbol{P}_1, w_1) &= \left(\begin{bmatrix}0.40\\0.60\end{bmatrix}, 0.5\right), \quad (\boldsymbol{P}_2, w_2) = \left(\begin{bmatrix}0.60\\0.40\end{bmatrix}, 0.5\right); \\
(\widetilde{\boldsymbol{P}}_1, \tilde{w}_1) &= \left(\begin{bmatrix}0.55\\0.45\end{bmatrix}, 0.8\right), \quad (\widetilde{\boldsymbol{P}}_2, \tilde{w}_2) = \left(\begin{bmatrix}0.30\\0.70\end{bmatrix}, 0.2\right)
\end{aligned}$$

Identifiability: $L = 2$ counterexample

Parameters $\{(\boldsymbol{P}_1, w_1), (\boldsymbol{P}_2, w_2)\}$ and $\{(\widetilde{\boldsymbol{P}}_1, \tilde{w}_1), (\widetilde{\boldsymbol{P}}_2, \tilde{w}_2)\}$

$$(\boldsymbol{P}_1, w_1) \;=\; \left(\begin{bmatrix}0.40\\0.60\end{bmatrix}, 0.5\right), \quad (\boldsymbol{P}_2, w_2) \;=\; \left(\begin{bmatrix}0.60\\0.40\end{bmatrix}, 0.5\right);$$

$$(\widetilde{\boldsymbol{P}}_1, \tilde{w}_1) \;=\; \left(\begin{bmatrix}0.55\\0.45\end{bmatrix}, 0.8\right), \quad (\widetilde{\boldsymbol{P}}_2, \tilde{w}_2) \;=\; \left(\begin{bmatrix}0.30\\0.70\end{bmatrix}, 0.2\right)$$

satisfy

$$w_1 \boldsymbol{P}_1 \boldsymbol{P}_1^\top + w_2 \boldsymbol{P}_2 \boldsymbol{P}_2^\top \;=\; \tilde{w}_1 \widetilde{\boldsymbol{P}}_1 \widetilde{\boldsymbol{P}}_1^\top + \tilde{w}_2 \widetilde{\boldsymbol{P}}_2 \widetilde{\boldsymbol{P}}_2^\top \;=\; \begin{bmatrix}0.26 & 0.24\\0.24 & 0.26\end{bmatrix}.$$

Identifiability: $L = 2$ counterexample

Parameters $\{(\boldsymbol{P}_1, w_1), (\boldsymbol{P}_2, w_2)\}$ and $\{(\widetilde{\boldsymbol{P}}_1, \tilde{w}_1), (\widetilde{\boldsymbol{P}}_2, \tilde{w}_2)\}$

$$(\boldsymbol{P}_1, w_1) \ = \ \left(\begin{bmatrix} 0.40 \\ 0.60 \end{bmatrix}, 0.5 \right), \quad (\boldsymbol{P}_2, w_2) \ = \ \left(\begin{bmatrix} 0.60 \\ 0.40 \end{bmatrix}, 0.5 \right);$$

$$(\widetilde{\boldsymbol{P}}_1, \tilde{w}_1) \ = \ \left(\begin{bmatrix} 0.55 \\ 0.45 \end{bmatrix}, 0.8 \right), \quad (\widetilde{\boldsymbol{P}}_2, \tilde{w}_2) \ = \ \left(\begin{bmatrix} 0.30 \\ 0.70 \end{bmatrix}, 0.2 \right)$$

satisfy

$$w_1 \boldsymbol{P}_1 \boldsymbol{P}_1^\top + w_2 \boldsymbol{P}_2 \boldsymbol{P}_2^\top \ = \ \tilde{w}_1 \widetilde{\boldsymbol{P}}_1 \widetilde{\boldsymbol{P}}_1^\top + \tilde{w}_2 \widetilde{\boldsymbol{P}}_2 \widetilde{\boldsymbol{P}}_2^\top \ = \ \begin{bmatrix} 0.26 & 0.24 \\ 0.24 & 0.26 \end{bmatrix}.$$

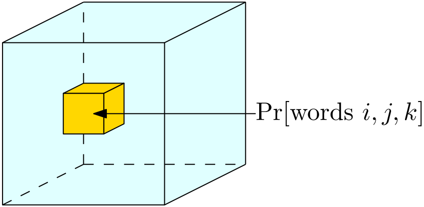
Cannot identify parameters from length-two documents.

Identifiability: $L = 3$

Documents of length $L = 3$

Joint distribution of word triple (for topic t) is given by *tensor*:

$\boldsymbol{P}_t \otimes \boldsymbol{P}_t \otimes \boldsymbol{P}_t =$



Random document $\sim \sum_{t=1}^{K} w_t \boldsymbol{P}_t \otimes \boldsymbol{P}_t \otimes \boldsymbol{P}_t$.

Identifiability from documents of length three

Claim: If $\{\boldsymbol{P}_t\}_{t=1}^K$ are linearly independent & all $w_t > 0$, then parameters $\{(\boldsymbol{P}_t, w_t)\}_{t=1}^K$ are identifiable from word triples.

Identifiability from documents of length three

Claim: If $\{\boldsymbol{P}_t\}_{t=1}^{K}$ are linearly independent & all $w_t > 0$, then parameters $\{(\boldsymbol{P}_t, w_t)\}_{t=1}^{K}$ are identifiable from word triples.

- Claim implied by uniqueness of certain *tensor decompositions*.

Identifiability from documents of length three

Claim: If $\{\boldsymbol{P}_t\}_{t=1}^K$ are linearly independent & all $w_t > 0$, then parameters $\{(\boldsymbol{P}_t, w_t)\}_{t=1}^K$ are identifiable from word triples.

- Claim implied by uniqueness of certain *tensor decompositions*.
- Proof is *constructive*: i.e., comes with an algorithm!

Identifiability from documents of length three

Claim: If $\{\boldsymbol{P}_t\}_{t=1}^{K}$ are linearly independent & all $w_t > 0$, then parameters $\{(\boldsymbol{P}_t, w_t)\}_{t=1}^{K}$ are identifiable from word triples.

- Claim implied by uniqueness of certain *tensor decompositions*.
- Proof is *constructive*: i.e., comes with an algorithm!

Next: Brief overview of tensors.

Tensors of order two

Matrices (tensors of order two): $\boldsymbol{M} \in \mathbb{R}^{d \times d}$.

- Regard as *bi-linear function* $\boldsymbol{M} \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$:

$$\begin{aligned} \boldsymbol{M}(a\boldsymbol{x} + \boldsymbol{x}', \boldsymbol{y}) &= a\boldsymbol{M}(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{M}(\boldsymbol{x}', \boldsymbol{y}) \,; \\ \boldsymbol{M}(\boldsymbol{x}, a\boldsymbol{y} + \boldsymbol{y}') &= a\boldsymbol{M}(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{M}(\boldsymbol{x}, \boldsymbol{y}') \,. \end{aligned}$$

Tensors of order two

Matrices (tensors of order two): $\boldsymbol{M} \in \mathbb{R}^{d \times d}$.

- Regard as *bi-linear function* $\boldsymbol{M} \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$:

$$\begin{aligned} \boldsymbol{M}(a\boldsymbol{x} + \boldsymbol{x}', \boldsymbol{y}) &= a\boldsymbol{M}(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{M}(\boldsymbol{x}', \boldsymbol{y}) ; \\ \boldsymbol{M}(\boldsymbol{x}, a\boldsymbol{y} + \boldsymbol{y}') &= a\boldsymbol{M}(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{M}(\boldsymbol{x}, \boldsymbol{y}') . \end{aligned}$$

- Can describe $\boldsymbol{M}$ by d^2 values $\boldsymbol{M}(\boldsymbol{e}_i, \boldsymbol{e}_j) =: M_{i,j}$.

  ($\boldsymbol{e}_i$ is ith coordinate basis vector.)

Tensors of order two

Matrices (tensors of order two): $\boldsymbol{M} \in \mathbb{R}^{d \times d}$.

- Regard as *bi-linear function* $\boldsymbol{M} \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$:

$$\begin{aligned} \boldsymbol{M}(a\boldsymbol{x} + \boldsymbol{x}', \boldsymbol{y}) &= a\boldsymbol{M}(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{M}(\boldsymbol{x}', \boldsymbol{y})\,; \\ \boldsymbol{M}(\boldsymbol{x}, a\boldsymbol{y} + \boldsymbol{y}') &= a\boldsymbol{M}(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{M}(\boldsymbol{x}, \boldsymbol{y}')\,. \end{aligned}$$

- Can describe $\boldsymbol{M}$ by d^2 values $\boldsymbol{M}(\boldsymbol{e}_i, \boldsymbol{e}_j) =: M_{i,j}$.

  ($\boldsymbol{e}_i$ is ith coordinate basis vector.)

- Formula using matrix representation:

$$\boldsymbol{M}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{y} = \sum_{i,j} M_{i,j} \cdot x_i y_j\,.$$

Tensors of order two

Matrices (tensors of order two): $\boldsymbol{M} \in \mathbb{R}^{d \times d}$.

- Regard as *bi-linear function* $\boldsymbol{M} \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$:

$$\begin{aligned} \boldsymbol{M}(a\boldsymbol{x} + \boldsymbol{x}', \boldsymbol{y}) &= a\boldsymbol{M}(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{M}(\boldsymbol{x}', \boldsymbol{y}) \,; \\ \boldsymbol{M}(\boldsymbol{x}, a\boldsymbol{y} + \boldsymbol{y}') &= a\boldsymbol{M}(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{M}(\boldsymbol{x}, \boldsymbol{y}') \,. \end{aligned}$$

- Can describe $\boldsymbol{M}$ by d^2 values $\boldsymbol{M}(\boldsymbol{e}_i, \boldsymbol{e}_j) =: M_{i,j}$.

  ($\boldsymbol{e}_i$ is ith coordinate basis vector.)

- Formula using matrix representation:

$$\boldsymbol{M}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{y} = \sum_{i,j} M_{i,j} \cdot x_i y_j \,.$$

Tensors are *multi-linear* generalization.

Tensors of order p

p-**linear functions**: $\boldsymbol{T} \colon \mathbb{R}^d \times \mathbb{R}^d \times \cdots \times \mathbb{R}^d \to \mathbb{R}$.

Tensors of order p

p-**linear functions**: $\boldsymbol{T} \colon \mathbb{R}^d \times \mathbb{R}^d \times \cdots \times \mathbb{R}^d \to \mathbb{R}$.

- Can describe $\boldsymbol{T}$ by d^p values $\boldsymbol{T}(\boldsymbol{e}_{i_1}, \boldsymbol{e}_{i_2}, \ldots, \boldsymbol{e}_{i_p}) =: T_{i_1,i_2,\ldots,i_p}$.

Tensors of order p

p-**linear functions**: $\boldsymbol{T} \colon \mathbb{R}^d \times \mathbb{R}^d \times \cdots \times \mathbb{R}^d \to \mathbb{R}$.

- Can describe $\boldsymbol{T}$ by d^p values $\boldsymbol{T}(\boldsymbol{e}_{i_1}, \boldsymbol{e}_{i_2}, \dots, \boldsymbol{e}_{i_p}) =: T_{i_1,i_2,\dots,i_p}$.
- Identify $\boldsymbol{T}$ with multi-index array $\boldsymbol{T} \in \mathbb{R}^{d \times d \times \cdots \times d}$.

Tensors of order p

p-**linear functions**: $\boldsymbol{T}\colon \mathbb{R}^d \times \mathbb{R}^d \times \cdots \times \mathbb{R}^d \to \mathbb{R}$.

- Can describe $\boldsymbol{T}$ by d^p values $\boldsymbol{T}(\boldsymbol{e}_{i_1}, \boldsymbol{e}_{i_2}, \dots, \boldsymbol{e}_{i_p}) =: T_{i_1,i_2,\dots,i_p}$.
- Identify $\boldsymbol{T}$ with multi-index array $\boldsymbol{T} \in \mathbb{R}^{d \times d \times \cdots \times d}$.

  Formula for function value:

$$\boldsymbol{T}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(p)}) \;=\; \sum_{i_1,i_2,\dots,i_p} T_{i_1,i_2,\dots,i_p} \cdot x^{(1)}_{i_1} x^{(2)}_{i_2} \cdots x^{(p)}_{i_p} \,.$$

Tensors of order p

p-**linear functions**: $\boldsymbol{T}\colon \mathbb{R}^d \times \mathbb{R}^d \times \cdots \times \mathbb{R}^d \to \mathbb{R}$.

- Can describe $\boldsymbol{T}$ by d^p values $\boldsymbol{T}(\boldsymbol{e}_{i_1}, \boldsymbol{e}_{i_2}, \dots, \boldsymbol{e}_{i_p}) =: T_{i_1,i_2,\dots,i_p}$.
- Identify $\boldsymbol{T}$ with multi-index array $\boldsymbol{T} \in \mathbb{R}^{d\times d\times\cdots\times d}$.

  Formula for function value:

$$\boldsymbol{T}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(p)}) \;=\; \sum_{i_1,i_2,\dots,i_p} T_{i_1,i_2,\dots,i_p} \cdot x^{(1)}_{i_1} x^{(2)}_{i_2} \cdots x^{(p)}_{i_p} \,.$$

- *Rank*-1 *tensor*: $\boldsymbol{T} = \boldsymbol{v}^{(1)} \otimes \boldsymbol{v}^{(2)} \otimes \cdots \otimes \boldsymbol{v}^{(p)}$,

$$\boldsymbol{T}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(p)}) \;=\; \langle \boldsymbol{v}^{(1)}, \boldsymbol{x}^{(1)} \rangle \langle \boldsymbol{v}^{(2)}, \boldsymbol{x}^{(2)} \rangle \cdots \langle \boldsymbol{v}^{(p)}, \boldsymbol{x}^{(p)} \rangle \,.$$

Tensors of order p

p-**linear functions**: $\boldsymbol{T} \colon \mathbb{R}^d \times \mathbb{R}^d \times \cdots \times \mathbb{R}^d \to \mathbb{R}$.

- Can describe $\boldsymbol{T}$ by d^p values $\boldsymbol{T}(\boldsymbol{e}_{i_1}, \boldsymbol{e}_{i_2}, \dots, \boldsymbol{e}_{i_p}) =: T_{i_1,i_2,\dots,i_p}$.
- Identify $\boldsymbol{T}$ with multi-index array $\boldsymbol{T} \in \mathbb{R}^{d \times d \times \cdots \times d}$.

  Formula for function value:

$$\boldsymbol{T}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(p)}) \; = \; \sum_{i_1,i_2,\dots,i_p} T_{i_1,i_2,\dots,i_p} \cdot x_{i_1}^{(1)} x_{i_2}^{(2)} \cdots x_{i_p}^{(p)} \, .$$

- *Rank*-1 *tensor*: $\boldsymbol{T} = \boldsymbol{v}^{(1)} \otimes \boldsymbol{v}^{(2)} \otimes \cdots \otimes \boldsymbol{v}^{(p)}$,

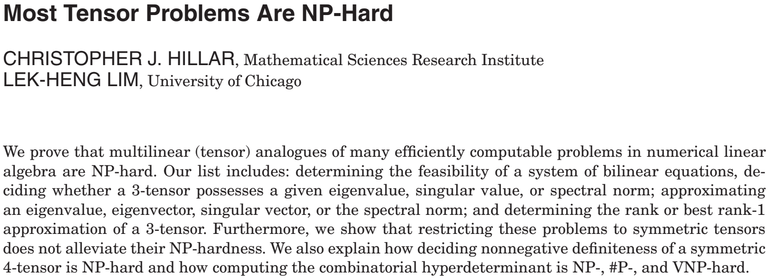
$$\boldsymbol{T}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(p)}) \; = \; \langle \boldsymbol{v}^{(1)}, \boldsymbol{x}^{(1)} \rangle \langle \boldsymbol{v}^{(2)}, \boldsymbol{x}^{(2)} \rangle \cdots \langle \boldsymbol{v}^{(p)}, \boldsymbol{x}^{(p)} \rangle \, .$$

  Symmetric rank-1 *tensor*: $\boldsymbol{T} = \boldsymbol{v}^{\otimes p} = \boldsymbol{v} \otimes \boldsymbol{v} \otimes \cdots \otimes \boldsymbol{v}$,

$$\boldsymbol{T}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(p)}) \; = \; \langle \boldsymbol{v}, \boldsymbol{x}^{(1)} \rangle \langle \boldsymbol{v}, \boldsymbol{x}^{(2)} \rangle \cdots \langle \boldsymbol{v}, \boldsymbol{x}^{(p)} \rangle \, .$$

Usual caveat

(Hillar & Lim, 2013)



Most Tensor Problems Are NP-Hard

CHRISTOPHER J. HILLAR, Mathematical Sciences Research Institute
LEK-HENG LIM, University of Chicago

We prove that multilinear (tensor) analogues of many efficiently computable problems in numerical linear algebra are NP-hard. Our list includes: determining the feasibility of a system of bilinear equations, deciding whether a 3-tensor possesses a given eigenvalue, singular value, or spectral norm; approximating an eigenvalue, eigenvector, singular vector, or the spectral norm; and determining the rank or best rank-1 approximation of a 3-tensor. Furthermore, we show that restricting these problems to symmetric tensors does not alleviate their NP-hardness. We also explain how deciding nonnegative definiteness of a symmetric 4-tensor is NP-hard and how computing the combinatorial hyperdeterminant is NP-, #P-, and VNP-hard.

Example: rank

- Rank of $\boldsymbol{T}$: smallest r s.t. $\boldsymbol{T}$ is sum of r rank-1 tensors.

Example: rank

- Rank of $\boldsymbol{T}$: smallest r s.t. $\boldsymbol{T}$ is sum of r rank-1 tensors.
  (Computing this is NP-hard.)

Example: rank

- Rank of $\boldsymbol{T}$: smallest r s.t. $\boldsymbol{T}$ is sum of r rank-1 tensors.
  (Computing this is NP-hard.)

- "Border rank" of $\boldsymbol{T}$: smallest r s.t. there exists sequence $(\boldsymbol{T}_k)_{k \in \mathbb{N}}$ of rank-r tensors with $\lim\limits_{k \to \infty} \boldsymbol{T}_k = \boldsymbol{T}$.

Example: rank

- Rank of $\boldsymbol{T}$: smallest r s.t. $\boldsymbol{T}$ is sum of r rank-1 tensors.
  (Computing this is NP-hard.)
- "Border rank" of $\boldsymbol{T}$: smallest r s.t. there exists sequence $(\boldsymbol{T}_k)_{k\in\mathbb{N}}$ of rank-r tensors with $\lim_{k\to\infty}\boldsymbol{T}_k = \boldsymbol{T}$.
- Rank is not same as border rank!

  Define

  $$\boldsymbol{T} \;:=\; \boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{y}+\boldsymbol{x}\otimes\boldsymbol{y}\otimes\boldsymbol{x}+\boldsymbol{y}\otimes\boldsymbol{x}\otimes\boldsymbol{x}\,,$$

  which has rank 3.

Example: rank

- Rank of $\boldsymbol{T}$: smallest r s.t. $\boldsymbol{T}$ is sum of r rank-1 tensors.
  (Computing this is NP-hard.)
- "Border rank" of $\boldsymbol{T}$: smallest r s.t. there exists sequence $(\boldsymbol{T}_k)_{k\in\mathbb{N}}$ of rank-r tensors with $\lim_{k\to\infty} \boldsymbol{T}_k = \boldsymbol{T}$.
- Rank is not same as border rank!

  Define

  $$\boldsymbol{T} \;:=\; \boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{y} + \boldsymbol{x}\otimes\boldsymbol{y}\otimes\boldsymbol{x} + \boldsymbol{y}\otimes\boldsymbol{x}\otimes\boldsymbol{x}\,,$$

  which has rank 3.

  Define

  $$\boldsymbol{T}_{1/\epsilon} \;:=\; \frac{1}{\epsilon}(\boldsymbol{x}+\epsilon\boldsymbol{y})\otimes(\boldsymbol{x}+\epsilon\boldsymbol{y})\otimes(\boldsymbol{x}+\epsilon\boldsymbol{y}) - \frac{1}{\epsilon}\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}\,,$$

  which have rank 2.

Example: rank

- Rank of $\boldsymbol{T}$: smallest r s.t. $\boldsymbol{T}$ is sum of r rank-1 tensors.
  (Computing this is NP-hard.)
- "Border rank" of $\boldsymbol{T}$: smallest r s.t. there exists sequence $(\boldsymbol{T}_k)_{k\in\mathbb{N}}$ of rank-r tensors with $\lim_{k\to\infty} \boldsymbol{T}_k = \boldsymbol{T}$.
- Rank is not same as border rank!

  Define

  $$\boldsymbol{T} \;:=\; \boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{y}+\boldsymbol{x}\otimes\boldsymbol{y}\otimes\boldsymbol{x}+\boldsymbol{y}\otimes\boldsymbol{x}\otimes\boldsymbol{x}\,,$$

  which has rank 3.

  Define

  $$\boldsymbol{T}_{1/\epsilon} \;:=\; \frac{1}{\epsilon}(\boldsymbol{x}+\epsilon\boldsymbol{y})\otimes(\boldsymbol{x}+\epsilon\boldsymbol{y})\otimes(\boldsymbol{x}+\epsilon\boldsymbol{y})-\frac{1}{\epsilon}\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}\,,$$

  which have rank 2.

  For $\epsilon = 1/k$, have $\lim_{k\to\infty} \boldsymbol{T}_k = \boldsymbol{T}$.

Aside: eigenvalue decomposition

Recall: every symmetric matrix $\boldsymbol{M} \in \mathbb{R}^{d \times d}$ of rank K has an *eigen-decomposition* (which can be efficiently computed):

$$\boldsymbol{M} \;=\; \sum_{t=1}^{K} \lambda_t \boldsymbol{v}_t \boldsymbol{v}_t^{\top} \,,$$

Aside: eigenvalue decomposition

Recall: every symmetric matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ of rank K has an *eigen-decomposition* (which can be efficiently computed):

$$\boldsymbol{M} \;=\; \sum_{t=1}^{K} \lambda_t \boldsymbol{v}_t \boldsymbol{v}_t^\top \,,$$

- $\{\lambda_t\}_{t=1}^{K}$ are *eigenvalues*,
- $\{\boldsymbol{v}_t\}_{t=1}^{K}$ are the corresponding *eigenvectors*, which are orthonormal (i.e., orthogonal & unit length).
- Decomposition is unique iff $\{\lambda_t\}_{t=1}^{K}$ are distinct.
  (Up to sign of $\boldsymbol{v}_t$s.)

Aside: eigenvalue decomposition

Recall: every symmetric matrix $\boldsymbol{M} \in \mathbb{R}^{d \times d}$ of rank K has an *eigen-decomposition* (which can be efficiently computed):

$$\boldsymbol{M} = \sum_{t=1}^{K} \lambda_t \boldsymbol{v}_t \boldsymbol{v}_t^\top ,$$

- $\{\lambda_t\}_{t=1}^{K}$ are *eigenvalues*,
- $\{\boldsymbol{v}_t\}_{t=1}^{K}$ are the corresponding *eigenvectors*, which are orthonormal (i.e., orthogonal & unit length).
- Decomposition is unique iff $\{\lambda_t\}_{t=1}^{K}$ are distinct.
  (Up to sign of $\boldsymbol{v}_t$s.)

For (symmetric) tensors of order $p \geq 3$:
an analogous decomposition is **not** guaranteed to exist.

Reduction to orthonormal case

Suppose we have (estimates of) moments of the form

$$\boldsymbol{M} = \sum_{t=1}^{K} \boldsymbol{v}_t \otimes \boldsymbol{v}_t \,, \qquad \text{(e.g., word pairs)}$$

and

$$\boldsymbol{T} = \sum_{t=1}^{K} \lambda_t \cdot \boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t \,. \qquad \text{(e.g., word triples)}$$

Here, we assume $\{\boldsymbol{v}_t\}_{t=1}^{K}$ are linearly independent, and $\{\lambda_t\}_{t=1}^{K}$ are positive.

Reduction to orthonormal case

Suppose we have (estimates of) moments of the form

$$\boldsymbol{M} \;=\; \sum_{t=1}^{K} \boldsymbol{v}_t \otimes \boldsymbol{v}_t\,, \qquad \text{(e.g., word pairs)}$$

$$\text{and} \qquad \boldsymbol{T} \;=\; \sum_{t=1}^{K} \lambda_t \cdot \boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t\,. \qquad \text{(e.g., word triples)}$$

Here, we assume $\{\boldsymbol{v}_t\}_{t=1}^{K}$ are linearly independent, and $\{\lambda_t\}_{t=1}^{K}$ are positive.

- $\boldsymbol{M}$ is positive semidefinite of rank K.

Reduction to orthonormal case

Suppose we have (estimates of) moments of the form

$$\boldsymbol{M} \;=\; \sum_{t=1}^{K} \boldsymbol{v}_t \otimes \boldsymbol{v}_t \,, \qquad \text{(e.g., word pairs)}$$

and

$$\boldsymbol{T} \;=\; \sum_{t=1}^{K} \lambda_t \cdot \boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t \,. \qquad \text{(e.g., word triples)}$$

Here, we assume $\{\boldsymbol{v}_t\}_{t=1}^{K}$ are linearly independent, and $\{\lambda_t\}_{t=1}^{K}$ are positive.

- $\boldsymbol{M}$ is positive semidefinite of rank K.
- $\boldsymbol{M}$ determines inner product system on $\operatorname{span}\{\boldsymbol{v}_t\}_{t=1}^{K}$ s.t. $\{\boldsymbol{v}_t\}_{t=1}^{K}$ are **orthonormal**:

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\boldsymbol{M}} \;:=\; \boldsymbol{x}^{\top} \boldsymbol{M}^{\dagger} \boldsymbol{y} \,.$$

Reduction to orthonormal case

Suppose we have (estimates of) moments of the form

$$\boldsymbol{M} \;=\; \sum_{t=1}^{K} \boldsymbol{v}_t \otimes \boldsymbol{v}_t \,, \qquad \text{(e.g., word pairs)}$$

and

$$\boldsymbol{T} \;=\; \sum_{t=1}^{K} \lambda_t \cdot \boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t \,. \qquad \text{(e.g., word triples)}$$

Here, we assume $\{\boldsymbol{v}_t\}_{t=1}^{K}$ are linearly independent, and $\{\lambda_t\}_{t=1}^{K}$ are positive.

- $\boldsymbol{M}$ is positive semidefinite of rank K.
- $\boldsymbol{M}$ determines inner product system on $\operatorname{span}\{\boldsymbol{v}_t\}_{t=1}^{K}$ s.t. $\{\boldsymbol{v}_t\}_{t=1}^{K}$ are **orthonormal**:

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{M} \;:=\; \boldsymbol{x}^\top \boldsymbol{M}^\dagger \boldsymbol{y} \,.$$

- $\therefore$ Can assume $d = K$ and $\{\boldsymbol{v}_t\}_{t=1}^{d}$ are orthonormal.
  (Similar to PCA; called "whitening" in signal processing context.)

Orthogonally decomposable tensors $(d = K)$

Goal: Given tensor $\boldsymbol{T} = \sum_{t=1}^{d} \lambda_t \cdot \boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t \in \mathbb{R}^{d \times d \times d}$ where:

- $\{\boldsymbol{v}_t\}_{t=1}^{d}$ are orthonormal;
- all $\lambda_t > 0$;

approximately recover $\{(\boldsymbol{v}_t, \lambda_t)\}_{t=1}^{d}$.

Exact orthogonally decomposable tensor

(Zhang & Golub, 2001)

Matching moments:

$$\{(\hat{\boldsymbol{v}}_t, \hat{\lambda}_t)\}_{t=1}^{d} \;:=\; \operatorname*{arg\,min}_{\{(\boldsymbol{x}_t, \sigma_t)\}_{t=1}^{d}} \left\| \boldsymbol{T} - \sum_{t=1}^{d} \sigma_t \cdot \boldsymbol{x}_t \otimes \boldsymbol{x}_t \otimes \boldsymbol{x}_t \right\|_F^2 .$$

(Here, $\| \cdot \|_F$ is "Frobenius norm", just like for matrices.)

Exact orthogonally decomposable tensor

(Zhang & Golub, 2001)

Matching moments:

$$\{(\hat{\boldsymbol{v}}_t, \hat{\lambda}_t)\}_{t=1}^{d} \;:=\; \operatorname*{arg\,min}_{\{(\boldsymbol{x}_t, \sigma_t)\}_{t=1}^{d}} \left\| \boldsymbol{T} - \sum_{t=1}^{d} \sigma_t \cdot \boldsymbol{x}_t \otimes \boldsymbol{x}_t \otimes \boldsymbol{x}_t \right\|_F^2 .$$

(Here, $\|\cdot\|_F$ is "Frobenius norm", just like for matrices.)

- Greedy approach:
  - Find best rank-1 approximation:

$$(\hat{\boldsymbol{v}}, \hat{\lambda}) \;:=\; \operatorname*{arg\,min}_{\|\boldsymbol{x}\|=1, \sigma \geq 0} \|\boldsymbol{T} - \sigma \cdot \boldsymbol{x} \otimes \boldsymbol{x} \otimes \boldsymbol{x}\|_F^2 .$$

  - "Deflate" $\boldsymbol{T} := \boldsymbol{T} - \hat{\lambda} \cdot \hat{\boldsymbol{v}} \otimes \hat{\boldsymbol{v}} \otimes \hat{\boldsymbol{v}}$ and repeat.

Exact orthogonally decomposable tensor

(Zhang & Golub, 2001)

Matching moments:

$$\{(\hat{\boldsymbol{v}}_t, \hat{\lambda}_t)\}_{t=1}^{d} \;:=\; \operatorname*{arg\,min}_{\{(\boldsymbol{x}_t,\sigma_t)\}_{t=1}^{d}} \left\| \boldsymbol{T} - \sum_{t=1}^{d} \sigma_t \cdot \boldsymbol{x}_t \otimes \boldsymbol{x}_t \otimes \boldsymbol{x}_t \right\|_F^2 .$$

(Here, $\|\cdot\|_F$ is "Frobenius norm", just like for matrices.)

- Greedy approach:
  - Find best rank-1 approximation:

$$\hat{\boldsymbol{v}} \;:=\; \operatorname*{arg\,max}_{\|\boldsymbol{x}\|=1} \boldsymbol{T}(\boldsymbol{x},\boldsymbol{x},\boldsymbol{x}) \,, \quad \hat{\lambda} \;:=\; \boldsymbol{T}(\hat{\boldsymbol{v}},\hat{\boldsymbol{v}},\hat{\boldsymbol{v}}) \,.$$

  - "Deflate" $\boldsymbol{T} := \boldsymbol{T} - \hat{\lambda} \cdot \hat{\boldsymbol{v}} \otimes \hat{\boldsymbol{v}} \otimes \hat{\boldsymbol{v}}$ and repeat.

Rank-1 approximation problem

Claim: Local maximizers of the function

$$\boldsymbol{x} \;\mapsto\; \boldsymbol{T}(\boldsymbol{x},\boldsymbol{x},\boldsymbol{x}) \;=\; \sum_{i,j,k} T_{i,j,k} \cdot x_i x_j x_k$$

(over the unit ball) are $\{\boldsymbol{v}_t\}_{t=1}^{d}$, and

$$\boldsymbol{T}(\boldsymbol{v}_t,\boldsymbol{v}_t,\boldsymbol{v}_t) \;=\; \lambda_t\,, \qquad t \in [d]\,.$$

Rank-1 approximation problem

Claim: Local maximizers of the function

$$\boldsymbol{x} \ \mapsto \ \boldsymbol{T}(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{x}) \ = \ \sum_{i,j,k} T_{i,j,k} \cdot x_i x_j x_k \ = \ \sum_{t=1}^{d} \lambda_t \cdot \langle \boldsymbol{v}_t, \boldsymbol{x} \rangle^3$$

(over the unit ball) are $\{\boldsymbol{v}_t\}_{t=1}^{d}$, and

$$\boldsymbol{T}(\boldsymbol{v}_t, \boldsymbol{v}_t, \boldsymbol{v}_t) \ = \ \lambda_t \, , \qquad t \in [d] \, .$$

Rank-1 approximation problem

Claim: Local maximizers of the function

$$\boldsymbol{x} \;\mapsto\; \boldsymbol{T}(\boldsymbol{x},\boldsymbol{x},\boldsymbol{x}) \;=\; \sum_{i,j,k} T_{i,j,k} \cdot x_i x_j x_k \;=\; \sum_{t=1}^{d} \lambda_t \cdot \langle \boldsymbol{v}_t, \boldsymbol{x} \rangle^3$$

(over the unit ball) are $\{\boldsymbol{v}_t\}_{t=1}^{d}$, and

$$\boldsymbol{T}(\boldsymbol{v}_t,\boldsymbol{v}_t,\boldsymbol{v}_t) \;=\; \lambda_t \,, \qquad t \in [d]\,.$$

Corollary: decomposition of $\boldsymbol{T}$ as $\sum_{t=1}^{K} \lambda_t \cdot \boldsymbol{v}_t^{\otimes 3}$ is *unique*!

Proof

By linearity and orthogonality:

$$\boldsymbol{T}(\boldsymbol{v}_t, \boldsymbol{v}_t, \boldsymbol{v}_t) \;=\; \sum_{s=1}^{d} (\lambda_s \cdot \boldsymbol{v}_s^{\otimes 3})(\boldsymbol{v}_t, \boldsymbol{v}_t, \boldsymbol{v}_t)$$

Proof

By linearity and orthogonality:

$$\boldsymbol{T}(\boldsymbol{v}_t, \boldsymbol{v}_t, \boldsymbol{v}_t) \;=\; \sum_{s=1}^{d} (\lambda_s \cdot \boldsymbol{v}_s^{\otimes 3})(\boldsymbol{v}_t, \boldsymbol{v}_t, \boldsymbol{v}_t) \;=\; \sum_{s=1}^{d} \begin{cases} \lambda_s & \text{if } s = t \\ 0 & \text{if } s \neq t \end{cases}$$

Proof

By linearity and orthogonality:

$$\boldsymbol{T}(\boldsymbol{v}_t, \boldsymbol{v}_t, \boldsymbol{v}_t) \;=\; \sum_{s=1}^{d} (\lambda_s \cdot \boldsymbol{v}_s^{\otimes 3})(\boldsymbol{v}_t, \boldsymbol{v}_t, \boldsymbol{v}_t) \;=\; \sum_{s=1}^{d} \begin{cases} \lambda_s & \text{if } s = t \\ 0 & \text{if } s \neq t \end{cases} \;=\; \lambda_t \,.$$

Proof

By linearity and orthogonality:

$$\boldsymbol{T}(\boldsymbol{v}_t, \boldsymbol{v}_t, \boldsymbol{v}_t) \;=\; \sum_{s=1}^{d} (\lambda_s \cdot \boldsymbol{v}_s^{\otimes 3})(\boldsymbol{v}_t, \boldsymbol{v}_t, \boldsymbol{v}_t) \;=\; \sum_{s=1}^{d} \begin{cases} \lambda_s & \text{if } s = t \\ 0 & \text{if } s \neq t \end{cases} \;=\; \lambda_t \, .$$

WLOG assume $\boldsymbol{v}_t = \boldsymbol{e}_t$, so optimization problem is

$$\max_{\boldsymbol{x} \in \mathbb{R}^d} \sum_{t=1}^{d} \lambda_t x_t^3 \quad \text{s.t.} \quad \sum_{t=1}^{d} x_t^2 \leq 1 \, .$$

Proof

By linearity and orthogonality:

$$\boldsymbol{T}(\boldsymbol{v}_t, \boldsymbol{v}_t, \boldsymbol{v}_t) \;=\; \sum_{s=1}^{d} (\lambda_s \cdot \boldsymbol{v}_s^{\otimes 3})(\boldsymbol{v}_t, \boldsymbol{v}_t, \boldsymbol{v}_t) \;=\; \sum_{s=1}^{d} \begin{cases} \lambda_s & \text{if } s = t \\ 0 & \text{if } s \neq t \end{cases} \;=\; \lambda_t \,.$$

WLOG assume $\boldsymbol{v}_t = \boldsymbol{e}_t$, so optimization problem is

$$\max_{\boldsymbol{x} \in \mathbb{R}^d} \sum_{t=1}^{d} \lambda_t x_t^3 \quad \text{s.t.} \quad \sum_{t=1}^{d} x_t^2 \leq 1 \,.$$

If both x_1 and x_2 are non-zero, then

$$\lambda_1 x_1^3 + \lambda_2 x_2^3 \;<\; \lambda_1 x_1^2 + \lambda_2 x_2^2$$

Proof

By linearity and orthogonality:

$$\boldsymbol{T}(\boldsymbol{v}_t, \boldsymbol{v}_t, \boldsymbol{v}_t) \;=\; \sum_{s=1}^{d} (\lambda_s \cdot \boldsymbol{v}_s^{\otimes 3})(\boldsymbol{v}_t, \boldsymbol{v}_t, \boldsymbol{v}_t) \;=\; \sum_{s=1}^{d} \begin{cases} \lambda_s & \text{if } s = t \\ 0 & \text{if } s \neq t \end{cases} \;=\; \lambda_t \,.$$

WLOG assume $\boldsymbol{v}_t = \boldsymbol{e}_t$, so optimization problem is

$$\max_{\boldsymbol{x} \in \mathbb{R}^d} \sum_{t=1}^{d} \lambda_t x_t^3 \quad \text{s.t.} \quad \sum_{t=1}^{d} x_t^2 \leq 1 \,.$$

If both x_1 and x_2 are non-zero, then

$$\lambda_1 x_1^3 + \lambda_2 x_2^3 \;<\; \lambda_1 x_1^2 + \lambda_2 x_2^2 \;\leq\; \max\{\lambda_1, \lambda_2\} \,.$$

Proof

By linearity and orthogonality:

$$\boldsymbol{T}(\boldsymbol{v}_t, \boldsymbol{v}_t, \boldsymbol{v}_t) \;=\; \sum_{s=1}^{d} (\lambda_s \cdot \boldsymbol{v}_s^{\otimes 3})(\boldsymbol{v}_t, \boldsymbol{v}_t, \boldsymbol{v}_t) \;=\; \sum_{s=1}^{d} \begin{cases} \lambda_s & \text{if } s = t \\ 0 & \text{if } s \neq t \end{cases} \;=\; \lambda_t \,.$$

WLOG assume $\boldsymbol{v}_t = \boldsymbol{e}_t$, so optimization problem is

$$\max_{\boldsymbol{x} \in \mathbb{R}^d} \sum_{t=1}^{d} \lambda_t x_t^3 \quad \text{s.t.} \quad \sum_{t=1}^{d} x_t^2 \leq 1 \,.$$

If both x_1 and x_2 are non-zero, then

$$\lambda_1 x_1^3 + \lambda_2 x_2^3 \;<\; \lambda_1 x_1^2 + \lambda_2 x_2^2 \;\leq\; \max\{\lambda_1, \lambda_2\} \,.$$

So better to put all energy on a single coordinate.
$\therefore$ Local maximizers are $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d$.

Proof

By linearity and orthogonality:

$$\boldsymbol{T}(\boldsymbol{v}_t, \boldsymbol{v}_t, \boldsymbol{v}_t) \;=\; \sum_{s=1}^{d} (\lambda_s \cdot \boldsymbol{v}_s^{\otimes 3})(\boldsymbol{v}_t, \boldsymbol{v}_t, \boldsymbol{v}_t) \;=\; \sum_{s=1}^{d} \begin{cases} \lambda_s & \text{if } s = t \\ 0 & \text{if } s \neq t \end{cases} \;=\; \lambda_t \,.$$

WLOG assume $\boldsymbol{v}_t = \boldsymbol{e}_t$, so optimization problem is

$$\max_{\boldsymbol{x} \in \mathbb{R}^d} \sum_{t=1}^{d} \lambda_t x_t^3 \quad \text{s.t.} \quad \sum_{t=1}^{d} x_t^2 \leq 1 \,.$$

If both x_1 and x_2 are non-zero, then

$$\lambda_1 x_1^3 + \lambda_2 x_2^3 \;<\; \lambda_1 x_1^2 + \lambda_2 x_2^2 \;\leq\; \max\{\lambda_1, \lambda_2\} \,.$$

So better to put all energy on a single coordinate.
$\therefore$ Local maximizers are $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_d$. $\square$

Uniqueness of orthogonal decompositions

What we have seen so far:

1. When components $\{\boldsymbol{v}_t\}_{t=1}^{d}$ are linearly independent:
   - Reduce decomposition problem to *orthogonal tensor decomposition*, where components are orthonormal.

Uniqueness of orthogonal decompositions

What we have seen so far:

1. When components $\{\boldsymbol{v}_t\}_{t=1}^d$ are linearly independent:
   - Reduce decomposition problem to *orthogonal tensor decomposition*, where components are orthonormal.

2. For orthogonally decomposable tensors $\boldsymbol{T}$, local maximizers of the function

   $$\boldsymbol{x} \mapsto \boldsymbol{T}(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{x})$$

   (over the unit ball) are $\{\boldsymbol{v}_t\}_{t=1}^d$.

Uniqueness of orthogonal decompositions

What we have seen so far:

1. When components $\{\boldsymbol{v}_t\}_{t=1}^{d}$ are linearly independent:
    - Reduce decomposition problem to *orthogonal tensor decomposition*, where components are orthonormal.
2. For orthogonally decomposable tensors $\boldsymbol{T}$, local maximizers of the function

   $$\boldsymbol{x} \mapsto \boldsymbol{T}(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{x})$$

   (over the unit ball) are $\{\boldsymbol{v}_t\}_{t=1}^{d}$.

Algorithm: use gradient ascent to find all of the local maximizers, which are exactly $\boldsymbol{v}_t$.

(Can use "deflation" to remove components from $\boldsymbol{T}$ that you've already found.)

Application to topic model parameters

Probabilities of word triples as third-order tensor:

$$\boldsymbol{T} \;=\; \sum_{t=1}^{K} w_t \boldsymbol{P}_t \otimes \boldsymbol{P}_t \otimes \boldsymbol{P}_t \;=\; \sum_{t=1}^{K} \boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t$$

for $\boldsymbol{v}_t = w_t^{1/3} \boldsymbol{P}_t$.

Application to topic model parameters

Probabilities of word triples as third-order tensor:

$$\boldsymbol{T} = \sum_{t=1}^{K} w_t \boldsymbol{P}_t \otimes \boldsymbol{P}_t \otimes \boldsymbol{P}_t = \sum_{t=1}^{K} \boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t$$

for $\boldsymbol{v}_t = w_t^{1/3} \boldsymbol{P}_t$.

- About linear independence condition on $\{\boldsymbol{v}_t\}_{t=1}^{K}$:

$$\begin{aligned} & \{\boldsymbol{v}_t\}_{t=1}^{K} \text{ are linearly independent} \\ \Leftrightarrow \quad & \{\boldsymbol{P}_t\}_{t=1}^{K} \text{ are linearly independent and all } w_t > 0. \end{aligned}$$

Application to topic model parameters

Probabilities of word triples as third-order tensor:

$$\boldsymbol{T} = \sum_{t=1}^{K} w_t \boldsymbol{P}_t \otimes \boldsymbol{P}_t \otimes \boldsymbol{P}_t = \sum_{t=1}^{K} \boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t$$

for $\boldsymbol{v}_t = w_t^{1/3} \boldsymbol{P}_t$.

- About linear independence condition on $\{\boldsymbol{v}_t\}_{t=1}^K$:

$$\begin{aligned} & \{\boldsymbol{v}_t\}_{t=1}^K \text{ are linearly independent} \\ \Leftrightarrow\ & \{\boldsymbol{P}_t\}_{t=1}^K \text{ are linearly independent and all } w_t > 0. \end{aligned}$$

- Can recover $\{\boldsymbol{P}_t\}_{t=1}^K$ from $\{c_t \boldsymbol{v}_t\}_{t=1}^K$ for any $c_t \neq 0$.

Application to topic model parameters

Probabilities of word triples as third-order tensor:

$$\boldsymbol{T} = \sum_{t=1}^{K} w_t \boldsymbol{P}_t \otimes \boldsymbol{P}_t \otimes \boldsymbol{P}_t = \sum_{t=1}^{K} \boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t$$

for $\boldsymbol{v}_t = w_t^{1/3} \boldsymbol{P}_t$.

- About linear independence condition on $\{\boldsymbol{v}_t\}_{t=1}^K$:

$$\begin{aligned} & \{\boldsymbol{v}_t\}_{t=1}^K \text{ are linearly independent} \\ \Leftrightarrow \quad & \{\boldsymbol{P}_t\}_{t=1}^K \text{ are linearly independent and all } w_t > 0. \end{aligned}$$

- Can recover $\{\boldsymbol{P}_t\}_{t=1}^K$ from $\{c_t \boldsymbol{v}_t\}_{t=1}^K$ for any $c_t \neq 0$.
- Can recover $\{(\boldsymbol{P}_t, w_t)\}_{t=1}^K$ from $\{\boldsymbol{P}_t\}_{t=1}^K$ and $\boldsymbol{T}$.

$\square$

Recap

- Parameters of topic model for single-topic documents (satisfying linear independence condition) can be efficiently recovered from distribution of three-word documents.

Recap

- Parameters of topic model for single-topic documents (satisfying linear independence condition) can be efficiently recovered from distribution of three-word documents.
- Two-word documents not sufficient (without further assumptions).

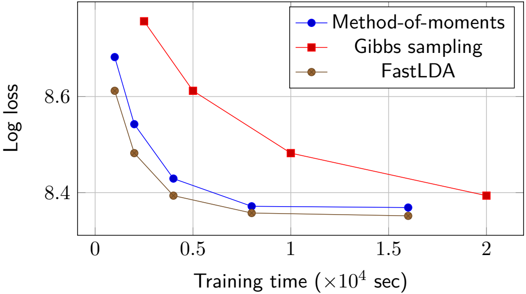
Recap

- Parameters of topic model for single-topic documents (satisfying linear independence condition) can be efficiently recovered from distribution of three-word documents.
- Two-word documents not sufficient (without further assumptions).
- Variational characterization of orthogonally decomposable tensors leads to simple and efficient algorithms!

Illustrative empirical results

- Corpus: $300,000$ New York Times articles.
- Vocabulary size: $102,660$ words.
- Set number of topics $K := 50$.

Model predictive performance:
≈ 4–$8\times$ speed-up over Gibbs sampling for LDA;
comparable to "FastLDA" (Porteous, Newman, Ihler, Asuncion, Smyth, & Welling, 2008).



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.

Computation

Caveat: forming and computing with a third-order tensor $\boldsymbol{T}$ generally requires cubic space.

Computation

Caveat: forming and computing with a third-order tensor $\boldsymbol{T}$ generally requires cubic space.

- Fortunately, the tensor we often work with is an *empirical estimate* of a $\boldsymbol{T}$: e.g.,

  $$\widehat{\boldsymbol{T}} \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathrm{data}_i \,,$$

  where data_i is a tensor involving only the i-th data point.

Computation

Caveat: forming and computing with a third-order tensor $\boldsymbol{T}$ generally requires cubic space.

- Fortunately, the tensor we often work with is an *empirical estimate* of a $\boldsymbol{T}$: e.g.,

  $$\widehat{\boldsymbol{T}} \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathrm{data}_i \,,$$

  where data_i is a tensor involving only the i-th data point.
- Our algorithms will only involve $\widehat{\boldsymbol{T}}$ through *evaluations* of $\widehat{\boldsymbol{T}}$ at (several) given arguments, say, $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$.

Computation

Caveat: forming and computing with a third-order tensor $\boldsymbol{T}$ generally requires cubic space.

- Fortunately, the tensor we often work with is an *empirical estimate* of a $\boldsymbol{T}$: e.g.,

  $$\widehat{\boldsymbol{T}} \ = \ \frac{1}{n}\sum_{i=1}^{n} \mathrm{data}_i \,,$$

  where data_i is a tensor involving only the i-th data point.
- Our algorithms will only involve $\widehat{\boldsymbol{T}}$ through *evaluations* of $\widehat{\boldsymbol{T}}$ at (several) given arguments, say, $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$.

  By linearity:

  $$\widehat{\boldsymbol{T}}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \ = \ \frac{1}{n}\sum_{i=1}^{n} \mathrm{data}_i(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \,.$$

Computation

Caveat: forming and computing with a third-order tensor $\boldsymbol{T}$ generally requires cubic space.

- Fortunately, the tensor we often work with is an *empirical estimate* of a $\boldsymbol{T}$: e.g.,

  $$\widehat{\boldsymbol{T}} \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathrm{data}_i \,,$$

  where data_i is a tensor involving only the i-th data point.
- Our algorithms will only involve $\widehat{\boldsymbol{T}}$ through *evaluations* of $\widehat{\boldsymbol{T}}$ at (several) given arguments, say, $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$.

  By linearity:

  $$\widehat{\boldsymbol{T}}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathrm{data}_i(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \,.$$

- **Often**: $\mathrm{data}_i(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ is easy to compute, even without forming any tensors!

Computation

Caveat: forming and computing with a third-order tensor $\boldsymbol{T}$ generally requires cubic space.

- Fortunately, the tensor we often work with is an *empirical estimate* of a $\boldsymbol{T}$: e.g.,

$$\widehat{\boldsymbol{T}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{data}_i \,,$$

  where data_i is a tensor involving only the i-th data point.

- Our algorithms will only involve $\widehat{\boldsymbol{T}}$ through *evaluations* of $\widehat{\boldsymbol{T}}$ at (several) given arguments, say, $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$.

  By linearity:

$$\widehat{\boldsymbol{T}}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{data}_i(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \,.$$

- **Often**: $\mathrm{data}_i(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ is easy to compute, even without forming any tensors! $\longrightarrow$ Linear time/space algorithms.

Learning algorithms

- Estimation via **method-of-moments**:
  1. *Estimate* distribution of three-word documents $\rightarrow$ $\widehat{\boldsymbol{T}}$ (*empirical moment tensor*).
  2. *Approximately decompose* $\widehat{\boldsymbol{T}}$ $\rightarrow$ estimates $\{(\widehat{\boldsymbol{P}}_t, \hat{w}_t)\}_{t=1}^{K}$.

Learning algorithms

- Estimation via **method-of-moments**:
  1. *Estimate* distribution of three-word documents $\rightarrow$ $\widehat{\boldsymbol{T}}$ (*empirical moment tensor*).
  2. *Approximately decompose* $\widehat{\boldsymbol{T}}$ $\rightarrow$ estimates $\{(\widehat{\boldsymbol{P}}_t, \hat{w}_t)\}_{t=1}^{K}$.
- **Issues**:
  1. Accuracy of *moment estimates*?
  2. Robustness of *(approximate) tensor decomposition*?
  3. *Generality* beyond simple topic models?

Learning algorithms

- Estimation via **method-of-moments**:
  1. *Estimate* distribution of three-word documents $\rightarrow$ $\widehat{\boldsymbol{T}}$ (*empirical moment tensor*).
  2. *Approximately decompose* $\widehat{\boldsymbol{T}}$ $\rightarrow$ estimates $\{(\widehat{\boldsymbol{P}}_t, \hat{w}_t)\}_{t=1}^{K}$.

- **Issues**:
  1. Accuracy of *moment estimates*?

     Can more reliably estimate lower-order moments; distribution-specific sample complexity bounds.
  2. Robustness of *(approximate) tensor decomposition*?

     In some sense, more stable than matrix eigen-decomposition (Mu, H., & Goldfarb, 2015)!
  3. *Generality* beyond simple topic models?

Learning algorithms

- Estimation via **method-of-moments**:
  1. *Estimate* distribution of three-word documents $\rightarrow$ $\widehat{\boldsymbol{T}}$ (*empirical moment tensor*).
  2. *Approximately decompose* $\widehat{\boldsymbol{T}}$ $\rightarrow$ estimates $\{(\widehat{\boldsymbol{P}}_t, \hat{w}_t)\}_{t=1}^{K}$.
- **Issues**:
  1. Accuracy of *moment estimates*?

     Can more reliably estimate lower-order moments; distribution-specific sample complexity bounds.
  2. Robustness of *(approximate) tensor decomposition*?

     In some sense, more stable than matrix eigen-decomposition (Mu, H., & Goldfarb, 2015)!
  3. *Generality* beyond simple topic models?

Next: Moment decompositions for other models.

2. Moment decompositions for other models

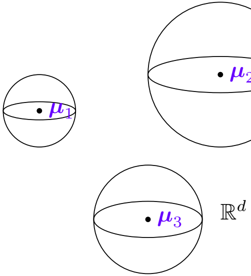
Moment decompositions

Some examples of usable moment decompositions.

1. Two classical mixture models.
2. Models with multi-view structure.
3. Single-index models.

Mixture model #1: Mixtures of spherical Gaussians

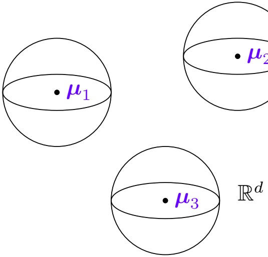
$$\begin{aligned} H &\sim \text{Categorical}(\pi_1, \pi_2, \ldots, \pi_K) \quad \text{(hidden)}\,; \\ \boldsymbol{X} \mid H = t &\sim \text{Normal}(\boldsymbol{\mu}_t, \sigma_t^2 \boldsymbol{I}_d)\,, \quad t \in [K]\,. \end{aligned}$$



Mixture model #1: Mixtures of spherical Gaussians

$$\begin{aligned} H &\sim \text{Categorical}(\pi_1, \pi_2, \ldots, \pi_K) \quad \text{(hidden)}\,; \\ \boldsymbol{X} \mid H = t &\sim \text{Normal}(\boldsymbol{\mu}_t, \sigma^2 \boldsymbol{I}_d)\,, \quad t \in [K]\,. \end{aligned}$$

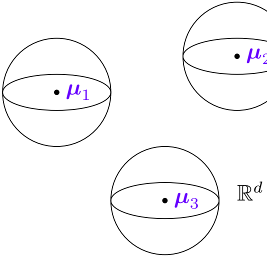
(For simplicity, restrict $\sigma_1 = \sigma_2 = \cdots = \sigma_K = \sigma$.)



Mixture model #1: Mixtures of spherical Gaussians

$$\begin{aligned} H &\sim \mathrm{Categorical}(\pi_1, \pi_2, \dots, \pi_K) \quad \text{(hidden)}\,; \\ \boldsymbol{X} \mid H = t &\sim \mathrm{Normal}(\boldsymbol{\mu}_t, \sigma^2 \boldsymbol{I}_d)\,, \quad t \in [K]\,. \end{aligned}$$

(For simplicity, restrict $\sigma_1 = \sigma_2 = \cdots = \sigma_K = \sigma$.)



Generative process:

$$\boldsymbol{X} = \boldsymbol{Y} + \sigma \boldsymbol{Z}$$

where $\Pr(\boldsymbol{Y} = \boldsymbol{\mu}_t) = \pi_t$, and
$\boldsymbol{Z} \sim \mathrm{Normal}(\boldsymbol{0}, \boldsymbol{I}_d)$ (indep. of $\boldsymbol{Y}$).

Using moments for spherical Gaussian mixtures

We'll see two ways to use low-order moments.

Using moments for spherical Gaussian mixtures

We'll see two ways to use low-order moments.

First- and second-order moments:

$$\mathbb{E}(\boldsymbol{X}) \ \in \ \mathbb{R}^d \qquad \text{and} \qquad \mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X}) \ \in \ \mathbb{R}^{d \times d}.$$

Using moments for spherical Gaussian mixtures

We'll see two ways to use low-order moments.

First- and second-order moments:

$$\mathbb{E}(\boldsymbol{X}) \; \in \; \mathbb{R}^d \qquad \text{and} \qquad \mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X}) \; \in \; \mathbb{R}^{d \times d} .$$

Claim (Vempala & Wang, 2002):
Span of top K eigenvectors of $\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X})$ contains $\{\boldsymbol{\mu}_t\}_{t=1}^K$.
(K-dimensional Principal Component Analysis (PCA) subspace.)

Proof

Key fact: k-dimensional PCA subspace (based on $\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X})$) captures as much of overall variance as any other k-dim. subspace.

Proof

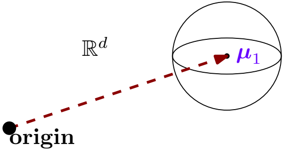
Key fact: k-dimensional PCA subspace (based on $\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X})$) captures as much of overall variance as any other k-dim. subspace.

- $K = 1$ (just a single Gaussian):
  What is the 1-dimensional PCA subspace?

Proof

Key fact: k-dimensional PCA subspace (based on $\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X})$) captures as much of overall variance as any other k-dim. subspace.

- $K = 1$ (just a single Gaussian):
  What is the 1-dimensional PCA subspace?

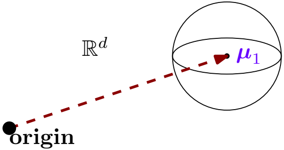


$$\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X}) \;=\; \boldsymbol{\mu}_1 \otimes \boldsymbol{\mu}_1 + \sigma^2 \boldsymbol{I}_d \,.$$

Proof

Key fact: k-dimensional PCA subspace (based on $\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X})$) captures as much of overall variance as any other k-dim. subspace.

- $K = 1$ (just a single Gaussian):
  What is the 1-dimensional PCA subspace?



$$\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X}) \;=\; \boldsymbol{\mu}_1 \otimes \boldsymbol{\mu}_1 + \sigma^2 \boldsymbol{I}_d \,.$$

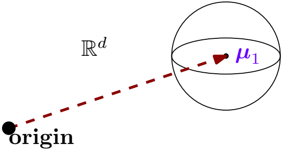
Variance in direction $\boldsymbol{v}$ (with $\|\boldsymbol{v}\| = 1$):

$\boldsymbol{v}^\top \mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X}) \boldsymbol{v}$

Proof

Key fact: k-dimensional PCA subspace (based on $\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X})$) captures as much of overall variance as any other k-dim. subspace.

- $K = 1$ (just a single Gaussian):
  What is the 1-dimensional PCA subspace?



$$\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X}) \;=\; \boldsymbol{\mu}_1 \otimes \boldsymbol{\mu}_1 + \sigma^2 \boldsymbol{I}_d\,.$$

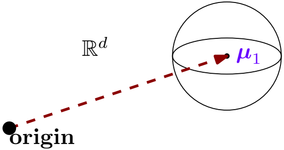
Variance in direction $\boldsymbol{v}$ (with $\|\boldsymbol{v}\| = 1$):

$$\boldsymbol{v}^\top \mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X}) \boldsymbol{v} \;=\; \boldsymbol{v}^\top (\boldsymbol{\mu}_1 \otimes \boldsymbol{\mu}_1 + \sigma^2 \boldsymbol{I}_d)\, \boldsymbol{v}$$

Proof

Key fact: k-dimensional PCA subspace (based on $\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X})$) captures as much of overall variance as any other k-dim. subspace.

- $K = 1$ (just a single Gaussian):
  What is the 1-dimensional PCA subspace?



$$\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X}) \;=\; \boldsymbol{\mu}_1 \otimes \boldsymbol{\mu}_1 + \sigma^2 \boldsymbol{I}_d \,.$$

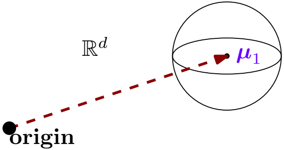
Variance in direction $\boldsymbol{v}$ (with $\|\boldsymbol{v}\| = 1$):

$$\boldsymbol{v}^\top \mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X}) \boldsymbol{v} \;=\; \boldsymbol{v}^\top (\boldsymbol{\mu}_1 \otimes \boldsymbol{\mu}_1 + \sigma^2 \boldsymbol{I}_d)\, \boldsymbol{v} \;=\; (\boldsymbol{v}^\top \boldsymbol{\mu}_1)^2 + \sigma^2 \,.$$

Proof

Key fact: k-dimensional PCA subspace (based on $\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X})$) captures as much of overall variance as any other k-dim. subspace.

- $K = 1$ (just a single Gaussian):
  What is the 1-dimensional PCA subspace?



$$\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X}) \;=\; \boldsymbol{\mu}_1 \otimes \boldsymbol{\mu}_1 + \sigma^2 \boldsymbol{I}_d \,.$$

Variance in direction $\boldsymbol{v}$ (with $\|\boldsymbol{v}\| = 1$):

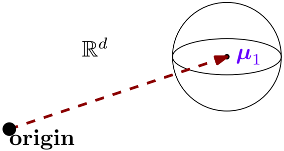
$$\boldsymbol{v}^\top \mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X}) \boldsymbol{v} \;=\; \boldsymbol{v}^\top (\boldsymbol{\mu}_1 \otimes \boldsymbol{\mu}_1 + \sigma^2 \boldsymbol{I}_d) \,\boldsymbol{v} \;=\; (\boldsymbol{v}^\top \boldsymbol{\mu}_1)^2 + \sigma^2 \,.$$

Best direction (1-dim. PCA subspace): $\boldsymbol{v} = \pm \boldsymbol{\mu}_1 / \|\boldsymbol{\mu}_1\|$.

Proof (continued)

Key fact: k-dimensional PCA subspace (based on $\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X})$) captures as much of overall variance as any other k-dim. subspace.

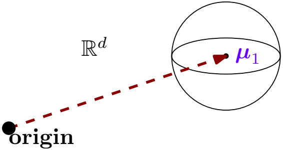
- $K = 1$ (just a single Gaussian):
  What is the k-dimensional PCA subspace?



Proof (continued)

Key fact: k-dimensional PCA subspace (based on $\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X})$) captures as much of overall variance as any other k-dim. subspace.

- $K = 1$ (just a single Gaussian):
  What is the k-dimensional PCA subspace?



Answer: any k-dim. subspace containing $\boldsymbol{\mu}_1$.

Proof (continued)

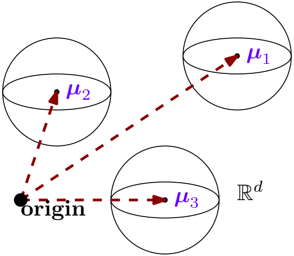
Key fact: k-dimensional PCA subspace (based on $\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X})$) captures as much of overall variance as any other k-dim. subspace.

- General K (mixture of K Gaussians):
  What is the K-dimensional PCA subspace?

Proof (continued)

Key fact: k-dimensional PCA subspace (based on $\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X})$) captures as much of overall variance as any other k-dim. subspace.

- General K (mixture of K Gaussians):
  What is the K-dimensional PCA subspace?

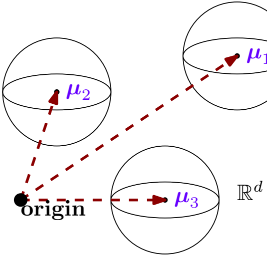


$$\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X}) = \sum_{t=1}^{K} \pi_t \cdot \boldsymbol{\mu}_t \otimes \boldsymbol{\mu}_t + \sigma^2 \boldsymbol{I}_d \,.$$

Proof (continued)

Key fact: k-dimensional PCA subspace (based on $\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X})$) captures as much of overall variance as any other k-dim. subspace.

- General K (mixture of K Gaussians):
  What is the K-dimensional PCA subspace?



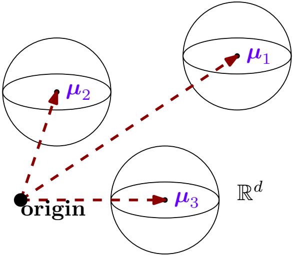
$$\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X}) = \sum_{t=1}^{K} \pi_t \cdot \boldsymbol{\mu}_t \otimes \boldsymbol{\mu}_t + \sigma^2 \boldsymbol{I}_d \,.$$

Answer: any K-dim. subspace containing $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$. □

Proof (continued)

Key fact: k-dimensional PCA subspace (based on $\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X})$) captures as much of overall variance as any other k-dim. subspace.

- General K (mixture of K Gaussians):
  What is the K-dimensional PCA subspace?



$$\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X}) = \sum_{t=1}^{K} \pi_t \cdot \boldsymbol{\mu}_t \otimes \boldsymbol{\mu}_t + \sigma^2 \boldsymbol{I}_d \,.$$

Answer: any K-dim. subspace containing $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$. ☐

How does this help with learning mixtures of Gaussians?

Use of moments for mixtures of spherical Gaussians

Separation (Dasgupta, 1999):

$\#$ standard deviations between component means

$$\text{sep} \ := \ \min_{i \neq j} \frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|}{\sigma} \, .$$

Use of moments for mixtures of spherical Gaussians

Separation (Dasgupta, 1999):
standard deviations between component means

$$\mathsf{sep} \ := \ \min_{i \neq j} \frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|}{\sigma} \, .$$

- (Dasgupta & Schulman, 2000):
  Distance-based clustering (e.g., EM) works when $\mathsf{sep} \gtrsim d^{1/4}$.

Use of moments for mixtures of spherical Gaussians

Separation (Dasgupta, 1999):
standard deviations between component means

$$\text{sep} \;:=\; \min_{i \neq j} \frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|}{\sigma} \,.$$

- (Dasgupta & Schulman, 2000):
  Distance-based clustering (e.g., EM) works when sep $\gtrsim d^{1/4}$.
- (Vempala & Wang, 2002):
  Problem becomes K-dimensional via PCA (assume $K \leq d$).
  Required separation reduced to sep $\gtrsim K^{1/4}$.

Use of moments for mixtures of spherical Gaussians

Separation (Dasgupta, 1999):
$\#$ standard deviations between component means

$$\text{sep} \; := \; \min_{i \neq j} \frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|}{\sigma} \, .$$

- (Dasgupta & Schulman, 2000):
  Distance-based clustering (e.g., EM) works when sep $\gtrsim d^{1/4}$.
- (Vempala & Wang, 2002):
  Problem becomes K-dimensional via PCA (assume $K \leq d$).
  Required separation reduced to sep $\gtrsim K^{1/4}$.

Third-order moments identify the mixture distribution when $\{\boldsymbol{\mu}_t\}_{t=1}^{K}$ are lin. indpt.; sep may be arbitrarily close to zero.

Use of moments for mixtures of spherical Gaussians

Separation (Dasgupta, 1999):

standard deviations between component means

$$\text{sep} \;:=\; \min_{i \neq j} \frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|}{\sigma}\,.$$

- (Dasgupta & Schulman, 2000):
  Distance-based clustering (e.g., EM) works when sep $\gtrsim d^{1/4}$.
- (Vempala & Wang, 2002):
  Problem becomes K-dimensional via PCA (assume $K \leq d$).
  Required separation reduced to sep $\gtrsim K^{1/4}$.

Third-order moments identify the mixture distribution when $\{\boldsymbol{\mu}_t\}_{t=1}^{K}$ are lin. indpt.; sep may be arbitrarily close to zero.

(Belkin & Sinha, 2010; Moitra & Valiant, 2010):

General Gaussians & no minimum sep, but Kth-order moments.

Third-order moments of spherical Gaussian mixtures

Generative process:

$$\boldsymbol{X} \; = \; \boldsymbol{Y} + \sigma \boldsymbol{Z}$$

where $\Pr(\boldsymbol{Y} = \boldsymbol{\mu}_t) = \pi_t$, and $\boldsymbol{Z} \sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{I}_d)$, $\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z}$.

Third-order moment tensor:

$$\mathbb{E}\left(\boldsymbol{X}^{\otimes 3}\right) \; = \; \mathbb{E}\left(\{\boldsymbol{Y} + \sigma \boldsymbol{Z}\}^{\otimes 3}\right)$$

Third-order moments of spherical Gaussian mixtures

Generative process:

$$\boldsymbol{X} \;=\; \boldsymbol{Y} + \sigma \boldsymbol{Z}$$

where $\Pr(\boldsymbol{Y} = \boldsymbol{\mu}_t) = \pi_t$, and $\boldsymbol{Z} \sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{I}_d)$, $\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z}$.

Third-order moment tensor:

$$\begin{aligned}\mathbb{E}\left(\boldsymbol{X}^{\otimes 3}\right) \;&=\; \mathbb{E}\left(\{\boldsymbol{Y} + \sigma \boldsymbol{Z}\}^{\otimes 3}\right) \\ \;&=\; \mathbb{E}\left(\boldsymbol{Y}^{\otimes 3}\right) + \sigma^2\, \mathbb{E}\left(\boldsymbol{Y} \otimes \boldsymbol{Z} \otimes \boldsymbol{Z} + \boldsymbol{Z} \otimes \boldsymbol{Y} \otimes \boldsymbol{Z} + \boldsymbol{Z} \otimes \boldsymbol{Z} \otimes \boldsymbol{Y}\right)\end{aligned}$$

Third-order moments of spherical Gaussian mixtures

Generative process:

$$\boldsymbol{X} \;=\; \boldsymbol{Y} + \sigma \boldsymbol{Z}$$

where $\Pr(\boldsymbol{Y} = \boldsymbol{\mu}_t) = \pi_t$, and $\boldsymbol{Z} \sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{I}_d)$, $\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z}$.

Third-order moment tensor:

$$\begin{aligned}
\mathbb{E}\left(\boldsymbol{X}^{\otimes 3}\right) \;&=\; \mathbb{E}\left(\{\boldsymbol{Y} + \sigma \boldsymbol{Z}\}^{\otimes 3}\right) \\
&=\; \mathbb{E}\left(\boldsymbol{Y}^{\otimes 3}\right) + \sigma^2\,\mathbb{E}\left(\boldsymbol{Y} \otimes \boldsymbol{Z} \otimes \boldsymbol{Z} + \boldsymbol{Z} \otimes \boldsymbol{Y} \otimes \boldsymbol{Z} + \boldsymbol{Z} \otimes \boldsymbol{Z} \otimes \boldsymbol{Y}\right) \\
&=\; \sum_{t=1}^{K} \pi_t \cdot \boldsymbol{\mu}_t^{\otimes 3} + \sigma^2 \tau(\boldsymbol{\mu})\,.
\end{aligned}$$

(Above, $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{X})$ and $\tau(\boldsymbol{\mu})$ is a third-order tensor involving only $\boldsymbol{\mu}$.)

Third-order moments of spherical Gaussian mixtures

Generative process:

$$\boldsymbol{X} \;=\; \boldsymbol{Y} + \sigma \boldsymbol{Z}$$

where $\Pr(\boldsymbol{Y} = \boldsymbol{\mu}_t) = \pi_t$, and $\boldsymbol{Z} \sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{I}_d)$, $\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z}$.

Third-order moment tensor:

$$\begin{aligned}
\mathbb{E}\left(\boldsymbol{X}^{\otimes 3}\right) \;&=\; \mathbb{E}\left(\{\boldsymbol{Y} + \sigma \boldsymbol{Z}\}^{\otimes 3}\right) \\
&=\; \mathbb{E}\left(\boldsymbol{Y}^{\otimes 3}\right) + \sigma^2\, \mathbb{E}\left(\boldsymbol{Y} \otimes \boldsymbol{Z} \otimes \boldsymbol{Z} + \boldsymbol{Z} \otimes \boldsymbol{Y} \otimes \boldsymbol{Z} + \boldsymbol{Z} \otimes \boldsymbol{Z} \otimes \boldsymbol{Y}\right) \\
&=\; \sum_{t=1}^{K} \pi_t \cdot \boldsymbol{\mu}_t^{\otimes 3} + \sigma^2 \tau(\boldsymbol{\mu})\,.
\end{aligned}$$

(Above, $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{X})$ and $\tau(\boldsymbol{\mu})$ is a third-order tensor involving only $\boldsymbol{\mu}$.)

Exercise: find explicit formula for $\tau(\boldsymbol{\mu})$.

Tensor decomposition for spherical Gaussian mixtures

(H. & Kakade, 2013)

$$\mathbb{E}\left(\boldsymbol{X}^{\otimes 3}\right) \;=\; \sum_{t=1}^{K} \pi_t \cdot \boldsymbol{\mu}_t^{\otimes 3} + \sigma^2 \tau(\boldsymbol{\mu})\,.$$

Tensor decomposition for spherical Gaussian mixtures

(H. & Kakade, 2013)

$$\mathbb{E}\left(\boldsymbol{X}^{\otimes 3}\right) \;=\; \sum_{t=1}^{K} \pi_t \cdot \boldsymbol{\mu}_t^{\otimes 3} + \sigma^2 \tau(\boldsymbol{\mu})\,.$$

Claim: $\boldsymbol{\mu}$ & σ^2 are simple functions of $\mathbb{E}(\boldsymbol{X})$ & $\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X})$.

Tensor decomposition for spherical Gaussian mixtures

(H. & Kakade, 2013)

$$\mathbb{E}\left(\boldsymbol{X}^{\otimes 3}\right) = \sum_{t=1}^{K} \pi_t \cdot \boldsymbol{\mu}_t^{\otimes 3} + \sigma^2 \tau(\boldsymbol{\mu}) .$$

Claim: $\boldsymbol{\mu}$ & σ^2 are simple functions of $\mathbb{E}(\boldsymbol{X})$ & $\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X})$.

Claim: If $\{\boldsymbol{\mu}_t\}_{t=1}^{K}$ are linearly independent and all $\pi_t > 0$, then $\{(\boldsymbol{\mu}_t, \pi_t)\}_{t=1}^{K}$ are identifiable from

$$\boldsymbol{T} := \mathbb{E}(\boldsymbol{X}^{\otimes 3}) - \sigma^2 \tau(\boldsymbol{\mu}) = \sum_{t=1}^{K} \pi_t \cdot \boldsymbol{\mu}_t^{\otimes 3} .$$

Tensor decomposition for spherical Gaussian mixtures

(H. & Kakade, 2013)

$$\mathbb{E}\left(\boldsymbol{X}^{\otimes 3}\right) \;=\; \sum_{t=1}^{K} \pi_t \cdot \boldsymbol{\mu}_t^{\otimes 3} + \sigma^2 \tau(\boldsymbol{\mu})\,.$$

Claim: $\boldsymbol{\mu}$ & σ^2 are simple functions of $\mathbb{E}(\boldsymbol{X})$ & $\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X})$.

Claim: If $\{\boldsymbol{\mu}_t\}_{t=1}^K$ are linearly independent and all $\pi_t > 0$, then $\{(\boldsymbol{\mu}_t, \pi_t)\}_{t=1}^K$ are identifiable from

$$\boldsymbol{T} \;:=\; \mathbb{E}(\boldsymbol{X}^{\otimes 3}) - \sigma^2 \tau(\boldsymbol{\mu}) \;=\; \sum_{t=1}^{K} \pi_t \cdot \boldsymbol{\mu}_t^{\otimes 3}\,.$$

Can use tensor decomposition to recover $\{(\boldsymbol{\mu}_t, \pi_t)\}_{t=1}^K$ from $\boldsymbol{T}$.

Even more Gaussian mixtures

Note: Linear independence condition on $\{\boldsymbol{\mu}_t\}_{t=1}^{K}$ requires $K \leq d$.

Even more Gaussian mixtures

Note: Linear independence condition on $\{\boldsymbol{\mu}_t\}_{t=1}^{K}$ requires $K \leq d$.

- (Anderson, Belkin, Goyal, Rademacher, & Voss, 2014), (Bhaskara, Charikar, Moitra, & Vijayaraghavan, 2014)
  Mixtures of $d^{O(1)}$ Gaussians (w/ simple or known covariance) via **smoothed analysis** and $O(1)$-order moments.

Even more Gaussian mixtures

Note: Linear independence condition on $\{\boldsymbol{\mu}_t\}_{t=1}^{K}$ requires $K \leq d$.

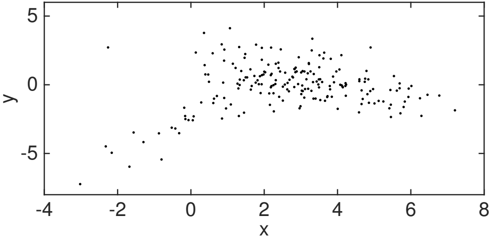
- (Anderson, Belkin, Goyal, Rademacher, & Voss, 2014), (Bhaskara, Charikar, Moitra, & Vijayaraghavan, 2014)
  Mixtures of $d^{O(1)}$ Gaussians (w/ simple or known covariance) via **smoothed analysis** and $O(1)$-order moments.
- (Ge, Huang, & Kakade, 2015)
  Also with **unknown covariances of arbitrary shape**.

Mixture model #2: Mixtures of linear regressions

$$\begin{aligned} H &\sim \mathrm{Categorical}(\pi_1, \pi_2, \ldots, \pi_K) \quad \textsf{(hidden)}\,; \\ \boldsymbol{X} &\sim \mathrm{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\,; \\ Y \mid H = t, \boldsymbol{X} = \boldsymbol{x} &\sim \mathrm{Normal}(\langle \boldsymbol{\beta}_t, \boldsymbol{x} \rangle, \sigma^2)\,. \end{aligned}$$

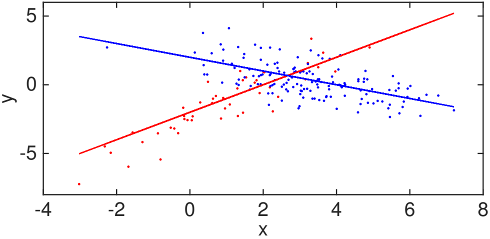
Mixture model #2: Mixtures of linear regressions

$$\begin{aligned} H &\sim \mathrm{Categorical}(\pi_1, \pi_2, \ldots, \pi_K) \quad \text{(hidden)}\,; \\ \boldsymbol{X} &\sim \mathrm{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\,; \\ Y \mid H = t, \boldsymbol{X} = \boldsymbol{x} &\sim \mathrm{Normal}(\langle \boldsymbol{\beta}_t, \boldsymbol{x} \rangle, \sigma^2)\,. \end{aligned}$$



Mixture model #2: Mixtures of linear regressions

$$\begin{aligned} H &\sim \text{Categorical}(\pi_1, \pi_2, \ldots, \pi_K) \quad \textsf{(hidden)}\,; \\ \boldsymbol{X} &\sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\,; \\ Y \mid H = t, \boldsymbol{X} = \boldsymbol{x} &\sim \text{Normal}(\langle \boldsymbol{\beta}_t, \boldsymbol{x} \rangle, \sigma^2)\,. \end{aligned}$$



Use of moments for mixtures of linear regressions

Second-order moments (assume $\boldsymbol{X} \sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{I}_d)$):

$$\mathbb{E}(Y^2 \boldsymbol{X} \boldsymbol{X}^\top) \;=\; 2 \sum_{t=1}^{K} \pi_t \cdot \boldsymbol{\beta}_t \boldsymbol{\beta}_t^\top + \left(\sigma^2 + \sum_{t=1}^{K} \pi_t \cdot \|\boldsymbol{\beta}_t\|^2 \right) \boldsymbol{I}_d \, .$$

Use of moments for mixtures of linear regressions

Second-order moments (assume $\boldsymbol{X} \sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{I}_d)$):

$$\mathbb{E}(Y^2 \boldsymbol{X}\boldsymbol{X}^\top) \;=\; 2\sum_{t=1}^{K} \pi_t \cdot \boldsymbol{\beta}_t \boldsymbol{\beta}_t^\top + \left(\sigma^2 + \sum_{t=1}^{K} \pi_t \cdot \|\boldsymbol{\beta}_t\|^2 \right) \boldsymbol{I}_d \,.$$

- Span of top K eigenvectors of $\mathbb{E}(Y^2 \boldsymbol{X}\boldsymbol{X}^\top)$ contains $\{\boldsymbol{\beta}_t\}_{t=1}^{K}$.

Use of moments for mixtures of linear regressions

Second-order moments (assume $\boldsymbol{X} \sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{I}_d)$):

$$\mathbb{E}(Y^2 \boldsymbol{X}\boldsymbol{X}^\top) \; = \; 2 \sum_{t=1}^{K} \pi_t \cdot \boldsymbol{\beta}_t \boldsymbol{\beta}_t^\top + \left(\sigma^2 + \sum_{t=1}^{K} \pi_t \cdot \|\boldsymbol{\beta}_t\|^2 \right) \boldsymbol{I}_d \, .$$

- Span of top K eigenvectors of $\mathbb{E}(Y^2 \boldsymbol{X}\boldsymbol{X}^\top)$ contains $\{\boldsymbol{\beta}_t\}_{t=1}^K$.
- Using Stein's identity (1973), similar approach works for GLMs (Sun, Ioannidis, & Montanari, 2013).

Use of moments for mixtures of linear regressions

Second-order moments (assume $\boldsymbol{X} \sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{I}_d)$):

$$\mathbb{E}(Y^2 \boldsymbol{X}\boldsymbol{X}^\top) \;=\; 2\sum_{t=1}^{K} \pi_t \cdot \boldsymbol{\beta}_t \boldsymbol{\beta}_t^\top + \left(\sigma^2 + \sum_{t=1}^{K} \pi_t \cdot \|\boldsymbol{\beta}_t\|^2 \right) \boldsymbol{I}_d \,.$$

- Span of top K eigenvectors of $\mathbb{E}(Y^2 \boldsymbol{X}\boldsymbol{X}^\top)$ contains $\{\boldsymbol{\beta}_t\}_{t=1}^{K}$.
- Using Stein's identity (1973), similar approach works for GLMs (Sun, Ioannidis, & Montanari, 2013).

Tensor decomposition approach:
Can recover parameters $\{(\boldsymbol{\beta}_t, \pi_t)\}_{t=1}^{K}$ with higher-order moments (Chaganty & Liang, 2013; Yi, Caramanis, & Sanghavi, 2014, 2016).

Use of moments for mixtures of linear regressions

Second-order moments (assume $\boldsymbol{X} \sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{I}_d)$):

$$\mathbb{E}(Y^2 \boldsymbol{X}\boldsymbol{X}^\top) \;=\; 2\sum_{t=1}^{K} \pi_t \cdot \boldsymbol{\beta}_t \boldsymbol{\beta}_t^\top + \left(\sigma^2 + \sum_{t=1}^{K} \pi_t \cdot \|\boldsymbol{\beta}_t\|^2\right) \boldsymbol{I}_d \, .$$

- Span of top K eigenvectors of $\mathbb{E}(Y^2 \boldsymbol{X}\boldsymbol{X}^\top)$ contains $\{\boldsymbol{\beta}_t\}_{t=1}^{K}$.
- Using Stein's identity (1973), similar approach works for GLMs (Sun, Ioannidis, & Montanari, 2013).

Tensor decomposition approach:

Can recover parameters $\{(\boldsymbol{\beta}_t, \pi_t)\}_{t=1}^{K}$ with higher-order moments (Chaganty & Liang, 2013; Yi, Caramanis, & Sanghavi, 2014, 2016).

Also for GLMs, via Stein's identity (Sedghi & Anandkumar, 2014).

Recap: mixtures of Gaussians and linear regressions

- Parameters of Gaussian mixture models and related models (satisfying linear independence condition) can be efficiently recovered from $O(1)$-order moments.

Recap: mixtures of Gaussians and linear regressions

- Parameters of Gaussian mixture models and related models (satisfying linear independence condition) can be efficiently recovered from $O(1)$-order moments.
- Exploit distributional properties to determine usable moments.

Recap: mixtures of Gaussians and linear regressions

- Parameters of Gaussian mixture models and related models (satisfying linear independence condition) can be efficiently recovered from $O(1)$-order moments.
- Exploit distributional properties to determine usable moments.
- *Smoothed analysis*: avoid linear independence condition for "most" mixture distributions.

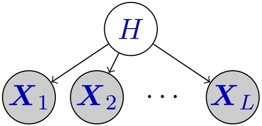
Recap: mixtures of Gaussians and linear regressions

- Parameters of Gaussian mixture models and related models (satisfying linear independence condition) can be efficiently recovered from $O(1)$-order moments.
- Exploit distributional properties to determine usable moments.
- *Smoothed analysis*: avoid linear independence condition for "most" mixture distributions.

Next: Multi-view approach to finding usable moments.

Multi-view interpretation of topic model

Recall: Topic model for single-topic documents



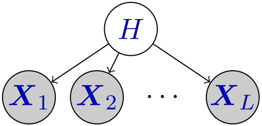
K topics (dists. over words) $\{\boldsymbol{P}_t\}_{t=1}^{K}$.

Pick topic $H = t$ with prob. w_t (**hidden**).

Word tokens $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_L \overset{\text{iid}}{\sim} \boldsymbol{P}_H$.

Multi-view interpretation of topic model

Recall: Topic model for single-topic documents



K topics (dists. over words) $\{\boldsymbol{P}_t\}_{t=1}^K$.

Pick topic $H = t$ with prob. w_t (**hidden**).

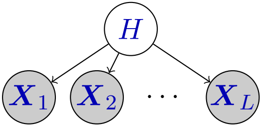
Word tokens $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_L \overset{\text{iid}}{\sim} \boldsymbol{P}_H$.

Key property:

$\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_L$ conditionally independent given H.

Multi-view interpretation of topic model

Recall: Topic model for single-topic documents



K topics (dists. over words) $\{\boldsymbol{P}_t\}_{t=1}^{K}$.

Pick topic $H = t$ with prob. w_t (**hidden**).

Word tokens $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_L \overset{\text{iid}}{\sim} \boldsymbol{P}_H$.

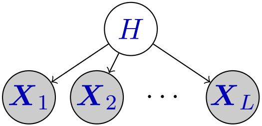
Key property:

$\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_L$ conditionally independent given H.

Each word token $\boldsymbol{X}_i$ provides new "view" of hidden variable H.

Multi-view interpretation of topic model

Recall: Topic model for single-topic documents



K topics (dists. over words) $\{\boldsymbol{P}_t\}_{t=1}^{K}$.
Pick topic $H = t$ with prob. w_t (**hidden**).
Word tokens $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_L \overset{\text{iid}}{\sim} \boldsymbol{P}_H$.

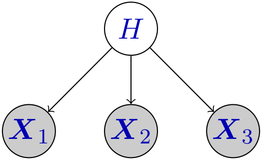
Key property:
$\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_L$ conditionally independent given H.

Each word token $\boldsymbol{X}_i$ provides new "view" of hidden variable H.

Some previous analyses:

- (Blum & Mitchell, 1998)
  Co-training in semi-supervised learning.
- (Chaudhuri, Kakade, Livescu, & Sridharan, 2009)
  Multi-view Gaussian mixture models.

Multi-view mixture model

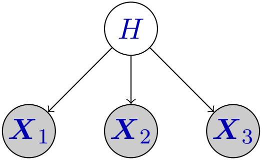


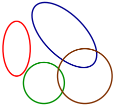
View 1: $\boldsymbol{X}_1$

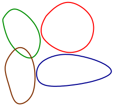
View 2: $\boldsymbol{X}_2$

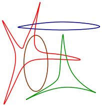
View 3: $\boldsymbol{X}_3$

Multi-view mixture model

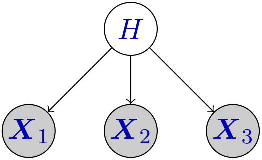



View 1: $\boldsymbol{X}_1$


View 2: $\boldsymbol{X}_2$


View 3: $\boldsymbol{X}_3$

Multi-view mixture model

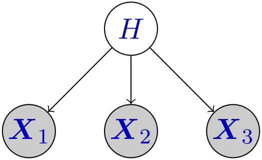


$$\mathbb{E}\left(\boldsymbol{X}_1 \otimes \boldsymbol{X}_2 \otimes \boldsymbol{X}_3\right) = \sum_{t=1}^{K} \pi_t \cdot \boldsymbol{\mu}_t^{(1)} \otimes \boldsymbol{\mu}_t^{(2)} \otimes \boldsymbol{\mu}_t^{(3)}$$

$$\text{where } \boldsymbol{\mu}_t^{(i)} = \mathbb{E}[\boldsymbol{X}_i \mid H = t],$$

$$\pi_t = \Pr(H = t).$$

Multi-view mixture model



$$\begin{aligned}\mathbb{E}\left(\boldsymbol{X}_1 \otimes \boldsymbol{X}_2 \otimes \boldsymbol{X}_3\right) &= \sum_{t=1}^{K} \pi_t \cdot \boldsymbol{\mu}_t^{(1)} \otimes \boldsymbol{\mu}_t^{(2)} \otimes \boldsymbol{\mu}_t^{(3)} \\ \text{where } \boldsymbol{\mu}_t^{(i)} &= \mathbb{E}[\boldsymbol{X}_i \mid H = t], \\ \pi_t &= \Pr(H = t).\end{aligned}$$

Tensor decomposition approach works in this asymmetric case as long as $\{\boldsymbol{\mu}_t^{(j)}\}_{t=1}^{K}$ are lin. indpt. for each j, and all $\pi_t > 0$.

Examples of multi-view mixture models

(Mossel & Roch, 2006; Anandkumar, H., & Kakade, 2012)

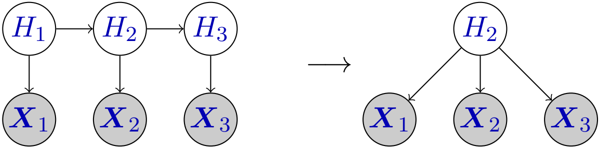
1. Mixtures of high-dimensional product distributions.
   (E.g., mixtures of axis-aligned Gaussians, other topic models.)

Examples of multi-view mixture models

(Mossel & Roch, 2006; Anandkumar, H., & Kakade, 2012)

1. Mixtures of high-dimensional product distributions.
   (E.g., mixtures of axis-aligned Gaussians, other topic models.)

2. Hidden Markov models.

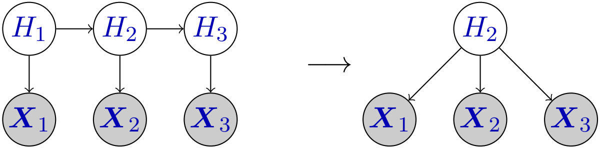


Examples of multi-view mixture models

(Mossel & Roch, 2006; Anandkumar, H., & Kakade, 2012)

1. Mixtures of high-dimensional product distributions.
   (E.g., mixtures of axis-aligned Gaussians, other topic models.)

2. Hidden Markov models.



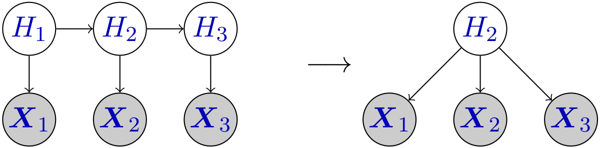
3. Phylogenetic trees.
   - $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$: genes of three extant species.
   - H: LCA of most closely related pair of species.

Examples of multi-view mixture models

(Mossel & Roch, 2006; Anandkumar, H., & Kakade, 2012)

1. Mixtures of high-dimensional product distributions.
   (E.g., mixtures of axis-aligned Gaussians, other topic models.)

2. Hidden Markov models.



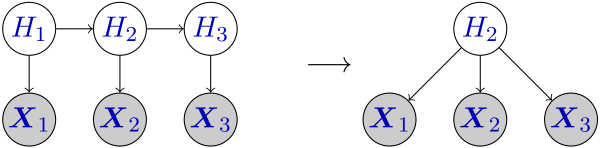
3. Phylogenetic trees.
   - $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$: genes of three extant species.
   - H: LCA of most closely related pair of species.

4. . . .

Examples of multi-view mixture models

(Mossel & Roch, 2006; Anandkumar, H., & Kakade, 2012)

1. Mixtures of high-dimensional product distributions.
   (E.g., mixtures of axis-aligned Gaussians, other topic models.)
2. Hidden Markov models.



3. Phylogenetic trees.
   - $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$: genes of three extant species.
   - H: LCA of most closely related pair of species.
4. . . .

Next: Single index models.

Single-index models

$$\begin{aligned} \boldsymbol{X} &\sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{I})\,; \\ Y \mid \boldsymbol{X} = \boldsymbol{x} &\sim \mathrm{Normal}(g(\langle \boldsymbol{\beta}, \boldsymbol{x} \rangle), \sigma^2)\,. \end{aligned}$$

Here, $g\colon \mathbb{R} \to \mathbb{R}$ is the *link function*.

Single-index models

$$\begin{aligned} \boldsymbol{X} &\sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{I})\,; \\ Y \mid \boldsymbol{X} = \boldsymbol{x} &\sim \mathrm{Normal}(g(\langle \boldsymbol{\beta}, \boldsymbol{x} \rangle), \sigma^2)\,. \end{aligned}$$

Here, $g \colon \mathbb{R} \to \mathbb{R}$ is the *link function*.

- **Phase retrieval** (real signals): assume $g(z) = z^2$.
- **1-bit compressed sensing**: assume $g(z) = \mathrm{sign}(z)$.
- **Isotonic regression**: assume g is monotone (e.g., $g' \geq 0$).
- **Convex regression**: assume g is convex (e.g., $g'' \geq 0$).
- . . .

Single-index models

$$\begin{aligned} \boldsymbol{X} &\sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{I})\,; \\ Y \mid \boldsymbol{X} = \boldsymbol{x} &\sim \mathrm{Normal}(g(\langle \boldsymbol{\beta}, \boldsymbol{x} \rangle), \sigma^2)\,. \end{aligned}$$

Here, $g\colon \mathbb{R} \to \mathbb{R}$ is the *link function*.

- **Phase retrieval** (real signals): assume $g(z) = z^2$.
- **1-bit compressed sensing**: assume $g(z) = \mathrm{sign}(z)$.
- **Isotonic regression**: assume g is monotone (e.g., $g' \geq 0$).
- **Convex regression**: assume g is convex (e.g., $g'' \geq 0$).
- . . .

When g is unknown, model is generally called **single-index model**.

Single-index models

$$\begin{aligned} \boldsymbol{X} &\sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{I}) \, ; \\ Y \mid \boldsymbol{X} = \boldsymbol{x} &\sim \mathrm{Normal}(g(\langle \boldsymbol{\beta}, \boldsymbol{x} \rangle), \sigma^2) \, . \end{aligned}$$

Here, $g \colon \mathbb{R} \to \mathbb{R}$ is the *link function*.

- **Phase retrieval** (real signals): assume $g(z) = z^2$.
- **1-bit compressed sensing**: assume $g(z) = \mathrm{sign}(z)$.
- **Isotonic regression**: assume g is monotone (e.g., $g' \geq 0$).
- **Convex regression**: assume g is convex (e.g., $g'' \geq 0$).
- . . .

When g is unknown, model is generally called **single-index model**.

Semi-parametric estimation: regard g as nuisance parameter; focus on estimating $\boldsymbol{\beta}$.

Aside: symmetric tensors and homogeneous polynomials

Recall formula for tensor function value:

$$\boldsymbol{T}(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(p)}) \;=\; \sum_{i_1,\ldots,i_p} T_{i_1,\ldots,i_p} \cdot x^{(1)}_{i_1} \cdots x^{(p)}_{i_p} \,.$$

Aside: symmetric tensors and homogeneous polynomials

Recall formula for tensor function value:

$$\boldsymbol{T}(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(p)}) \;=\; \sum_{i_1,\dots,i_p} T_{i_1,\dots,i_p} \cdot x^{(1)}_{i_1} \cdots x^{(p)}_{i_p} \,.$$

If $\boldsymbol{T}$ is symmetric (i.e., $T_{i_1,\dots,i_p} = T_{\pi(i_1),\dots,\pi(i_p)}$ for any permutation π), then evaluating at $\boldsymbol{x}^{(1)} = \cdots = \boldsymbol{x}^{(p)} = \boldsymbol{x}$ gives

$$\boldsymbol{T}(\boldsymbol{x}, \dots, \boldsymbol{x}) \;=\; c_p \sum_{i_1 \leq \cdots \leq i_p} T_{i_1,\dots,i_p} \cdot x_{i_1} \cdots x_{i_p} \,,$$

which is just the formula for a degree-p homogeneous polynomial.

Aside: symmetric tensors and homogeneous polynomials

Recall formula for tensor function value:

$$\boldsymbol{T}(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(p)}) \;=\; \sum_{i_1,\dots,i_p} T_{i_1,\dots,i_p} \cdot x^{(1)}_{i_1} \cdots x^{(p)}_{i_p} \,.$$

If $\boldsymbol{T}$ is symmetric (i.e., $T_{i_1,\dots,i_p} = T_{\pi(i_1),\dots,\pi(i_p)}$ for any permutation π), then evaluating at $\boldsymbol{x}^{(1)} = \cdots = \boldsymbol{x}^{(p)} = \boldsymbol{x}$ gives

$$\boldsymbol{T}(\boldsymbol{x}, \dots, \boldsymbol{x}) \;=\; c_p \sum_{i_1 \leq \cdots \leq i_p} T_{i_1,\dots,i_p} \cdot x_{i_1} \cdots x_{i_p} \,,$$

which is just the formula for a degree-p homogeneous polynomial.

p-th order symmetric tensors $\simeq$ degree-p homogeneous polynomials.

Using orthogonal polynomials

(Dudeja & H., 2018)

Let $H_p \colon \mathbb{R} \to \mathbb{R}$ denote the degree-p Hermite polynomial.

Assume (for $Z \sim \mathrm{Normal}(0,1)$):

- $\mathbb{E}[g(Z)^2] = 1$ (normalization—this is WLOG);
- $\mathbb{E}[g'(Z)^2] \geq \epsilon$ (necessary for identifiability);
- g is smooth and $\mathbb{E}[g''(Z)^2] = O(1)$.

Using orthogonal polynomials

(Dudeja & H., 2018)

Let $H_p \colon \mathbb{R} \to \mathbb{R}$ denote the degree-p Hermite polynomial.

Assume (for $Z \sim \mathrm{Normal}(0,1)$):

- $\mathbb{E}[g(Z)^2] = 1$ (normalization—this is WLOG);
- $\mathbb{E}[g'(Z)^2] \geq \epsilon$ (necessary for identifiability);
- g is smooth and $\mathbb{E}[g''(Z)^2] = O(1)$.

There exists $p = O(1/\epsilon)$ such that

$$\mathbb{E}[Y H_p(\langle \boldsymbol{v}, \boldsymbol{X} \rangle)] \ = \ (\lambda \boldsymbol{\beta}^{\otimes p})(\boldsymbol{v}), \quad \boldsymbol{v} \in \mathbb{R}^d$$

for some $\lambda \neq 0$ with $|\lambda| = \Omega(\epsilon/\sqrt{p})$.

Using orthogonal polynomials

(Dudeja & H., 2018)

Let $H_p \colon \mathbb{R} \to \mathbb{R}$ denote the degree-p Hermite polynomial.

Assume (for $Z \sim \mathrm{Normal}(0,1)$):

- $\mathbb{E}[g(Z)^2] = 1$ (normalization—this is WLOG);
- $\mathbb{E}[g'(Z)^2] \geq \epsilon$ (necessary for identifiability);
- g is smooth and $\mathbb{E}[g''(Z)^2] = O(1)$.

There exists $p = O(1/\epsilon)$ such that

$$\mathbb{E}[Y H_p(\langle \boldsymbol{v}, \boldsymbol{X} \rangle)] \;=\; (\lambda \boldsymbol{\beta}^{\otimes p})(\boldsymbol{v}), \quad \boldsymbol{v} \in \mathbb{R}^d$$

for some $\lambda \neq 0$ with $|\lambda| = \Omega(\epsilon/\sqrt{p})$.

$\Rightarrow$ Get efficient algorithms for semi-parametric estimation of single-index model parameters, for very general link functions.

Recap

- Parameters of many latent variable models (satisfying non-degeneracy conditions) can be efficiently recovered from $O(1)$-order moments.

Recap

- Parameters of many latent variable models (satisfying non-degeneracy conditions) can be efficiently recovered from $O(1)$-order moments.
- Exploit distributional properties, multi-view structure, and other structure to determine usable moments.

Recap

- Parameters of many latent variable models (satisfying non-degeneracy conditions) can be efficiently recovered from $O(1)$-order moments.

- Exploit distributional properties, multi-view structure, and other structure to determine usable moments.

- Estimation via **method-of-moments**:
    1. *Estimate* moments → empirical moment tensor $\widehat{\boldsymbol{T}}$.
    2. *Approximately decompose* $\widehat{\boldsymbol{T}}$ → parameter estimate $\hat{\boldsymbol{\theta}}$.

3. Error analysis

Moment estimates

Estimation of $\mathbb{E}[\boldsymbol{X}^{\otimes 3}]$ (say) from iid sample $\{\boldsymbol{x}_i\}_{i=1}^{n}$:

$$\widehat{\mathbb{E}}[\boldsymbol{X}^{\otimes 3}] \;:=\; \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i^{\otimes 3}\,.$$

Moment estimates

Estimation of $\mathbb{E}[\boldsymbol{X}^{\otimes 3}]$ (say) from iid sample $\{\boldsymbol{x}_i\}_{i=1}^{n}$:

$$\widehat{\mathbb{E}}[\boldsymbol{X}^{\otimes 3}] \ := \ \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i^{\otimes 3} \, .$$

Inevitably expect error of order $n^{-1/2}$ in some norm, e.g.,

$$\|\boldsymbol{T}\| \ := \ \sup_{\|\boldsymbol{x}\|=\|\boldsymbol{y}\|=\|\boldsymbol{z}\|=1} \boldsymbol{T}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) \quad \text{(injective/``spectral'' norm)}\,,$$

$$\|\boldsymbol{T}\|_F \ := \ \left(\sum_{i,j,k} T_{i,j,k}^2\right)^{1/2} \quad \text{(Frobenius norm)}\,.$$

Nearly orthogonally decomposable tensor

(Mu, H., & Goldfarb, 2015)

Let $\varepsilon = \|\boldsymbol{E}\|$ for $\boldsymbol{E} := \widehat{\boldsymbol{T}} - \boldsymbol{T}$.

Claim: Let $\hat{\boldsymbol{v}} := \arg\max_{\|\boldsymbol{x}\|=1} \widehat{\boldsymbol{T}}(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{x})$ and $\hat{\lambda} := \widehat{\boldsymbol{T}}(\hat{\boldsymbol{v}}, \hat{\boldsymbol{v}}, \hat{\boldsymbol{v}})$.

Then

$$|\hat{\lambda} - \lambda_t| \;\leq\; \varepsilon\,, \qquad \|\hat{\boldsymbol{v}} - \boldsymbol{v}_t\| \;\leq\; O\left(\frac{\varepsilon}{\lambda_t} + \left(\frac{\varepsilon}{\lambda_t}\right)^2\right)$$

for some $t \in [d]$ with $\lambda_t \geq \max_{t'} \lambda_{t'} - 2\varepsilon$.

Nearly orthogonally decomposable tensor

(Mu, H., & Goldfarb, 2015)

Let $\varepsilon = \|\boldsymbol{E}\|$ for $\boldsymbol{E} := \widehat{\boldsymbol{T}} - \boldsymbol{T}$.

Claim: Let $\hat{\boldsymbol{v}} := \arg\max_{\|\boldsymbol{x}\|=1} \widehat{\boldsymbol{T}}(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{x})$ and $\hat{\lambda} := \widehat{\boldsymbol{T}}(\hat{\boldsymbol{v}}, \hat{\boldsymbol{v}}, \hat{\boldsymbol{v}})$.

Then

$$|\hat{\lambda} - \lambda_t| \;\leq\; \varepsilon\,, \qquad \|\hat{\boldsymbol{v}} - \boldsymbol{v}_t\| \;\leq\; O\left(\frac{\varepsilon}{\lambda_t} + \left(\frac{\varepsilon}{\lambda_t}\right)^2\right)$$

for some $t \in [d]$ with $\lambda_t \geq \max_{t'} \lambda_{t'} - 2\varepsilon$.

Many efficient algorithms for solving this approximately, when ε is small enough, like $1/d$ or $1/\sqrt{d}$ (e.g., Anandkumar, Ge, H., Kakade, & Telgarsky, 2014; Ma, Shi, & Steurer, 2016).

Recall: greedy decomposition

(Zhang & Golub, 2001)

Matching moments:

$$\{(\hat{\boldsymbol{v}}_t, \hat{\lambda}_t)\}_{t=1}^{d} \; := \; \operatorname*{arg\,min}_{\{(\boldsymbol{x}_t, \sigma_t)\}_{t=1}^{d}} \left\| \boldsymbol{T} - \sum_{t=1}^{d} \sigma_t \cdot \boldsymbol{x}_t \otimes \boldsymbol{x}_t \otimes \boldsymbol{x}_t \right\|_F^2 .$$

Recall: greedy decomposition

(Zhang & Golub, 2001)

Matching moments:

$$\{(\hat{\boldsymbol{v}}_t, \hat{\lambda}_t)\}_{t=1}^d \;:=\; \operatorname*{arg\,min}_{\{(\boldsymbol{x}_t,\sigma_t)\}_{t=1}^d} \left\| \boldsymbol{T} - \sum_{t=1}^{d} \sigma_t \cdot \boldsymbol{x}_t \otimes \boldsymbol{x}_t \otimes \boldsymbol{x}_t \right\|_F^2 .$$

- Greedy approach:
  - Find best rank-1 approximation:

$$(\hat{\boldsymbol{v}}, \hat{\lambda}) \;:=\; \operatorname*{arg\,min}_{\|\boldsymbol{x}\|=1, \sigma \geq 0} \|\boldsymbol{T} - \sigma \cdot \boldsymbol{x} \otimes \boldsymbol{x} \otimes \boldsymbol{x}\|_F^2 .$$

  - "Deflate" $\boldsymbol{T} := \boldsymbol{T} - \hat{\lambda} \cdot \hat{\boldsymbol{v}} \otimes \hat{\boldsymbol{v}} \otimes \hat{\boldsymbol{v}}$ and repeat.

Recall: greedy decomposition

(Zhang & Golub, 2001)

Matching moments:

$$\{(\hat{\boldsymbol{v}}_t, \hat{\lambda}_t)\}_{t=1}^{d} \;:=\; \operatorname*{arg\,min}_{\{(\boldsymbol{x}_t, \sigma_t)\}_{t=1}^{d}} \left\| \boldsymbol{T} - \sum_{t=1}^{d} \sigma_t \cdot \boldsymbol{x}_t \otimes \boldsymbol{x}_t \otimes \boldsymbol{x}_t \right\|_F^2 .$$

- Greedy approach:
    - Find best rank-1 approximation:

$$\hat{\boldsymbol{v}} \;:=\; \operatorname*{arg\,max}_{\|\boldsymbol{x}\|=1} \boldsymbol{T}(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{x}) , \quad \hat{\lambda} \;:=\; \boldsymbol{T}(\hat{\boldsymbol{v}}, \hat{\boldsymbol{v}}, \hat{\boldsymbol{v}}) .$$

    - "Deflate" $\boldsymbol{T} := \boldsymbol{T} - \hat{\lambda} \cdot \hat{\boldsymbol{v}} \otimes \hat{\boldsymbol{v}} \otimes \hat{\boldsymbol{v}}$ and repeat.

Errors from deflation

(For simplicity, assume $\lambda_t = 1$ for all t, so $\boldsymbol{T} = \sum_t \boldsymbol{v}_t^{\otimes 3}$.)

First greedy step:
Rank-1 approx. $\hat{\boldsymbol{v}}_1^{\otimes 3}$ to $\widehat{\boldsymbol{T}}$ satisfies $\|\hat{\boldsymbol{v}}_1 - \boldsymbol{v}_1\| \leq \varepsilon$ (say).

Errors from deflation

(For simplicity, assume $\lambda_t = 1$ for all t, so $\boldsymbol{T} = \sum_t \boldsymbol{v}_t^{\otimes 3}$.)

First greedy step:
Rank-1 approx. $\hat{\boldsymbol{v}}_1^{\otimes 3}$ to $\widehat{\boldsymbol{T}}$ satisfies $\|\hat{\boldsymbol{v}}_1 - \boldsymbol{v}_1\| \leq \varepsilon$ (say).

Deflation: To find next $\boldsymbol{v}_t$, use

$$\begin{aligned}\widehat{\boldsymbol{T}} - \hat{\boldsymbol{v}}_1^{\otimes 3} &= \boldsymbol{T} + \boldsymbol{E} - \hat{\boldsymbol{v}}_1^{\otimes 3} \\ &= \sum_{t=\mathbf{2}}^{d} \boldsymbol{v}_t^{\otimes 3} + \boldsymbol{E} + \left(\boldsymbol{v}_1^{\otimes 3} - \hat{\boldsymbol{v}}_1^{\otimes 3} \right).\end{aligned}$$

Errors from deflation

(For simplicity, assume $\lambda_t = 1$ for all t, so $\boldsymbol{T} = \sum_t \boldsymbol{v}_t^{\otimes 3}$.)

First greedy step:
Rank-1 approx. $\hat{\boldsymbol{v}}_1^{\otimes 3}$ to $\widehat{\boldsymbol{T}}$ satisfies $\|\hat{\boldsymbol{v}}_1 - \boldsymbol{v}_1\| \leq \varepsilon$ (say).

Deflation: To find next $\boldsymbol{v}_t$, use

$$\begin{aligned} \widehat{\boldsymbol{T}} - \hat{\boldsymbol{v}}_1^{\otimes 3} &= \boldsymbol{T} + \boldsymbol{E} - \hat{\boldsymbol{v}}_1^{\otimes 3} \\ &= \sum_{t=\mathbf{2}}^{d} \boldsymbol{v}_t^{\otimes 3} + \boldsymbol{E} + \left(\boldsymbol{v}_1^{\otimes 3} - \hat{\boldsymbol{v}}_1^{\otimes 3} \right) . \end{aligned}$$

Now error seems to have doubled (i.e., of size 2ε) ...

Effect of deflation errors

For any unit vector $\boldsymbol{x}$ orthogonal to $\boldsymbol{v}_1$:

$$\left\| \frac{1}{3} \nabla_{\boldsymbol{x}} \left\{ \left(\boldsymbol{v}_1^{\otimes 3} - \hat{\boldsymbol{v}}_1^{\otimes 3} \right) (\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{x}) \right\} \right\| \quad = \quad \left\| \langle \boldsymbol{v}_1, \boldsymbol{x} \rangle^2 \boldsymbol{v}_1 - \langle \hat{\boldsymbol{v}}_1, \boldsymbol{x} \rangle^2 \hat{\boldsymbol{v}}_1 \right\|$$

Effect of deflation errors

For any unit vector $\boldsymbol{x}$ orthogonal to $\boldsymbol{v}_1$:

$$\begin{aligned} \left\| \frac{1}{3} \nabla_{\boldsymbol{x}} \left\{ \left(\boldsymbol{v}_1^{\otimes 3} - \hat{\boldsymbol{v}}_1^{\otimes 3} \right) (\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{x}) \right\} \right\| &= \left\| \langle \boldsymbol{v}_1, \boldsymbol{x} \rangle^2 \boldsymbol{v}_1 - \langle \hat{\boldsymbol{v}}_1, \boldsymbol{x} \rangle^2 \hat{\boldsymbol{v}}_1 \right\| \\ &= \langle \hat{\boldsymbol{v}}_1, \boldsymbol{x} \rangle^2 \end{aligned}$$

Effect of deflation errors

For any unit vector $\boldsymbol{x}$ orthogonal to $\boldsymbol{v}_1$:

$$\begin{aligned}\left\|\frac{1}{3}\nabla_{\boldsymbol{x}}\left\{\left(\boldsymbol{v}_1^{\otimes 3}-\hat{\boldsymbol{v}}_1^{\otimes 3}\right)(\boldsymbol{x},\boldsymbol{x},\boldsymbol{x})\right\}\right\| &= \left\|\langle\boldsymbol{v}_1,\boldsymbol{x}\rangle^2\boldsymbol{v}_1-\langle\hat{\boldsymbol{v}}_1,\boldsymbol{x}\rangle^2\hat{\boldsymbol{v}}_1\right\| \\ &= \langle\hat{\boldsymbol{v}}_1,\boldsymbol{x}\rangle^2 \\ &\leq \|\boldsymbol{v}_1-\hat{\boldsymbol{v}}_1\|^2 \;\leq\; \varepsilon^2\,.\end{aligned}$$

Effect of deflation errors

For any unit vector $\boldsymbol{x}$ orthogonal to $\boldsymbol{v}_1$:

$$\begin{aligned}\left\| \frac{1}{3} \nabla_{\boldsymbol{x}} \left\{ \left(\boldsymbol{v}_1^{\otimes 3} - \hat{\boldsymbol{v}}_1^{\otimes 3} \right) (\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{x}) \right\} \right\| &= \left\| \langle \boldsymbol{v}_1, \boldsymbol{x} \rangle^2 \boldsymbol{v}_1 - \langle \hat{\boldsymbol{v}}_1, \boldsymbol{x} \rangle^2 \hat{\boldsymbol{v}}_1 \right\| \\ &= \langle \hat{\boldsymbol{v}}_1, \boldsymbol{x} \rangle^2 \\ &\leq \| \boldsymbol{v}_1 - \hat{\boldsymbol{v}}_1 \|^2 \ \leq \ \varepsilon^2 \,.\end{aligned}$$

So effect of errors (original and from deflation) $\boldsymbol{E} + \left(\boldsymbol{v}_1^{\otimes 3} - \hat{\boldsymbol{v}}_1^{\otimes 3} \right)$ in directions orthogonal to $\boldsymbol{v}_1$ is $(1 + o(1))\varepsilon$ rather than 2ε.

Effect of deflation errors

For any unit vector $\boldsymbol{x}$ orthogonal to $\boldsymbol{v}_1$:

$$\begin{aligned} \left\| \frac{1}{3} \nabla_{\boldsymbol{x}} \left\{ \left(\boldsymbol{v}_1^{\otimes 3} - \hat{\boldsymbol{v}}_1^{\otimes 3} \right) (\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{x}) \right\} \right\| &= \left\| \langle \boldsymbol{v}_1, \boldsymbol{x} \rangle^2 \boldsymbol{v}_1 - \langle \hat{\boldsymbol{v}}_1, \boldsymbol{x} \rangle^2 \hat{\boldsymbol{v}}_1 \right\| \\ &= \langle \hat{\boldsymbol{v}}_1, \boldsymbol{x} \rangle^2 \\ &\leq \| \boldsymbol{v}_1 - \hat{\boldsymbol{v}}_1 \|^2 \leq \varepsilon^2 . \end{aligned}$$

So effect of errors (original and from deflation) $\boldsymbol{E} + \left(\boldsymbol{v}_1^{\otimes 3} - \hat{\boldsymbol{v}}_1^{\otimes 3} \right)$ in directions orthogonal to $\boldsymbol{v}_1$ is $(1 + o(1))\varepsilon$ rather than 2ε.

- Deflation errors have lower-order effect on finding other $\boldsymbol{v}_t$.

  (Analogous statement for deflation with matrices does not hold.)

Summary

- Using method-of-moments with low-order moments, can efficiently estimate parameters for many models.
    - Exploit distributional properties, multi-view structure, and other structure to determine usable moments tensors.
    - Some efficient algorithms for carrying out the tensor decomposition to obtain parameter estimates.

Summary

- Using method-of-moments with low-order moments, can efficiently estimate parameters for many models.
  - Exploit distributional properties, multi-view structure, and other structure to determine usable moments tensors.
  - Some efficient algorithms for carrying out the tensor decomposition to obtain parameter estimates.
- Many issues to resolve!
  - Handle model misspecification, increase robustness.
  - General methodology.
  - Incorporate general prior knowledge and interactive feedback.

Acknowledgements

Collaborators: Anima Anandkumar (Caltech), Rishabh Dudeja (Columbia), Dean Foster (Amazon), Rong Ge (Duke), Don Goldfarb (Columbia), Sham Kakade (UW), Percy Liang (Stanford), Yi-Kai Liu (NIST), Cun Mu (Jet), Matus Telgarsky (UIUC), Tong Zhang (Tencent)

Further reading:

- Anandkumar, Ge, H., Kakade, & Telgarsky.
  Tensor decompositions for learning latent variable models.
  Journal of Machine Learning Research, 15(Aug):2773–2831, 2014.
  `https://goo.gl/F8HudN`

¡Gracias!