Learning latent variable models using tensor decompositions

Daniel Hsu

Columbia University

January 27, 2017

Subject matter

Learning algorithms

for **latent variable models**

based on **decompositions of moment tensors**.

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.**

Team Relocations Keep N.F.L. Moving Up Financially

The Chargers' announced move to Los Angeles will add even more money for owners amid growing uncertainties facing the league.

By KEN BELSON

Jan. 12, 2017

Example #1: summarizing a corpus of documents

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

Team Relocations Keep N.F.L. Moving Up Financially

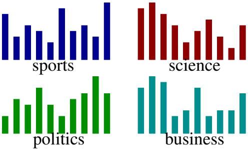
The Chargers' announced move to Los Angeles will add even more money for owners amid growing uncertainties facing the league.

By KEN BELSON

Jan. 12, 2017

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

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

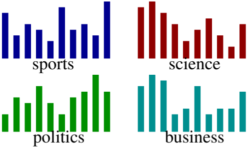


K topics (distributions over vocab words).

Document $\equiv$ mixture of topics.

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

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

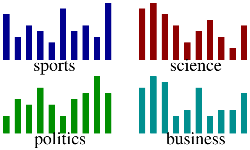


K topics (distributions over vocab words).
Document $\equiv$ mixture of topics.
Word tokens in doc. $\overset{\text{iid}}{\sim}$ mixture distribution.

E.g.,

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

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., 

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

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 $t_x \in \{1, 2, \ldots, K\}$.

Team	Relocations	Keep	N.F.L.	Moving	Up	Financially
1	1	1	1	4	4	4

Labels / annotations

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

Team	Relocations	Keep	N.F.L.	Moving	Up	Financially
1	1	1	1	4	4	4

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 $t_x \in \{1, 2, \ldots, K\}$.

Team	Relocations	Keep	N.F.L.	Moving	Up	Financially
1	1	1	1	4	4	4

  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.

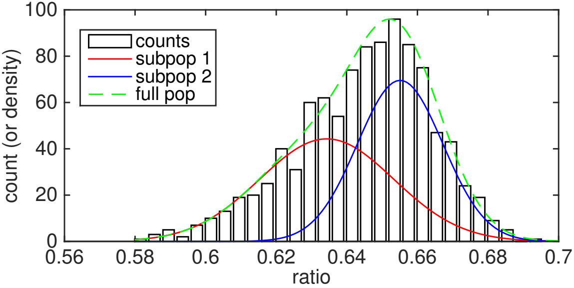
Example #2: subpopulations in data

Data studied by Pearson (1894):
ratio of forehead-width to body-length for 1000 crabs.

Example #2: subpopulations in data

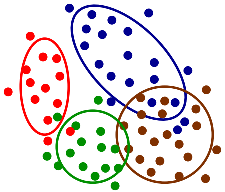
Data studied by Pearson (1894):
ratio of forehead-width to body-length for 1000 crabs.

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



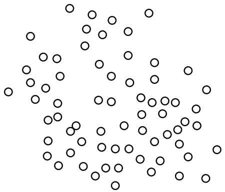
Gaussian mixture model

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



Gaussian mixture model

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



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):

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

Maximum likelihood estimation

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

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

  - **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):

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

  - **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., $\boldsymbol{\Sigma}_t \succeq \sigma^2 \boldsymbol{I}$).
- Similar to Euclidean K-means problem, which is NP-hard (Dasgupta, 2008; Aloise, Deshpande, Hansen, & Popat, 2009; Mahajan, Nimbhorkar, & Varadarajan, 2009; Vattani, 2009; Awasthi, Charikar, Krishnaswamy, & Sinop, 2015).

Parameter learning objective

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

Parameter learning objective

Suppose iid sample of size n is generated by distribution from model with (unknown) parameters $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ $\quad$ ($p = \#$ 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 (as $n \to \infty$):
  - For $K = 2$ (and $\pi_t = 1/2$, $\boldsymbol{\Sigma}_t = \boldsymbol{I}$): EM is consistent (Xu, H., & Maleki, 2016; Daskalakis, Tzamos, & Zampetakis, 2016).

Parameter learning objective

Suppose iid sample of size n is generated by distribution from model with (unknown) parameters $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ $\quad (p = \#$ 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 (as $n \to \infty$):
  - For $K = 2$ (and $\pi_t = 1/2$, $\boldsymbol{\Sigma}_t = \boldsymbol{I}$): EM is consistent (Xu, H., & Maleki, 2016; Daskalakis, Tzamos, & Zampetakis, 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 $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ $\quad (p = \#$ 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 (as $n \to \infty$):
  - For $K = 2$ (and $\pi_t = 1/2$, $\boldsymbol{\Sigma}_t = \boldsymbol{I}$): EM is consistent (Xu, H., & Maleki, 2016; Daskalakis, Tzamos, & Zampetakis, 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 $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ $\quad (p = \#$ 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 $\mathrm{poly}(p, 1/\epsilon, 1/\delta, \dots)$ sample size and running time.

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

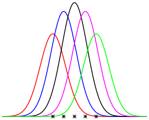
    Practitioners often use EM with many (random) restarts . . . but may take a long time to get near the global max.

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 $\min_{i \neq j} \frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2}{\sigma_i^2 + \sigma_j^2}$ is sufficiently large.

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

Ways around the barriers

- **Separation conditions**.

  E.g., assume $\min_{i \neq j} \frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2}{\sigma_i^2 + \sigma_j^2}$ is sufficiently large.

  (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 $\min_{i \neq j} \frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2}{\sigma_i^2 + \sigma_j^2}$ is sufficiently large.

  (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.

Ways around the barriers

- **Separation conditions.**

  E.g., assume $\min_{i \neq j} \frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2}{\sigma_i^2 + \sigma_j^2}$ is sufficiently large.

  (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 talk: 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 third-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.)

Unresolved issues

- Handle model misspecification, increase robustness.
  - Can tolerate some independence assumptions but not others?
- General methodology.
  - At present, *ad hoc* to instantiate; guided by examples.
- Incorporate general prior knowledge.
- Incorporate user feedback interactively.

Outline

1. Warm-up: topic model for single-topic documents.
   - Identifiability.
   - Parameter recovery via decompositions of exact moments.
2. Moment decompositions for other models.
   - Mixtures of Gaussians and linear regressions.
   - Multi-view models.
3. Error-tolerant algorithms for tensor decompositions.

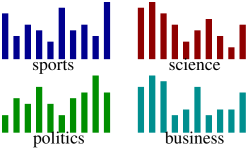
Other models amenable to moment tensor decomposition

- 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. Warm-up: topic model for single-topic documents

Topic model

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



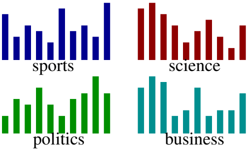
K topics (dists. over words) $\{\boldsymbol{P}_t\}_{t=1}^{K}$.

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

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

Topic model

Topic model for single-topic documents



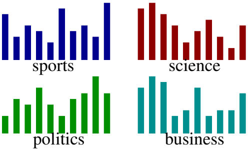
K topics (dists. over words) $\{\boldsymbol{P}_t\}_{t=1}^{K}$.

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

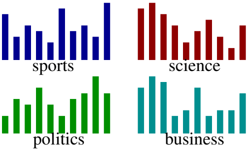


K topics (dists. over words) $\{\boldsymbol{P}_t\}_{t=1}^{K}$.
Pick topic t with prob. w_t (**hidden**).
Word tokens in doc. $\stackrel{\text{iid}}{\sim}$ $\boldsymbol{P}_t$.

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?

Identifiability

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

Identifiability

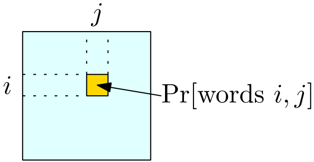
- **Generative process**:
  Pick $t \sim \mathrm{Discrete}(w_1, w_2, \ldots, w_K)$.
  Given t, pick L words from $\boldsymbol{P}_t$.
- $L = 1$: random document $\sim \sum_{t=1}^{K} w_t \boldsymbol{P}_t$

Identifiability

- **Generative process**:
  Pick $t \sim \mathrm{Discrete}(w_1, w_2, \dots, w_K)$.
  Given t, pick L words from $\boldsymbol{P}_t$.
- $L = 1$: random document $\sim \sum_{t=1}^{K} w_t \boldsymbol{P}_t$
  Parameters *not identifiable* from such observations.

Identifiability

- **Generative process**:
  Pick $t \sim \mathrm{Discrete}(w_1, w_2, \ldots, w_K)$.
  Given t, pick L words from $\boldsymbol{P}_t$.
- $L = 1$: random document $\sim \sum_{t=1}^{K} w_t \boldsymbol{P}_t$
  Parameters *not identifiable* from such observations.
- $L = 2$:
  Regard $\boldsymbol{P}_t$ as probability vector.
  Joint distribution of word pairs (for topic t) is given by matrix:

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

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

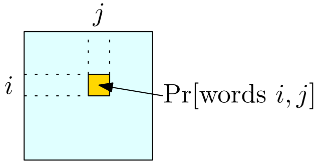
Identifiability

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

- $L = 1$: random document $\sim \sum_{t=1}^{K} w_t \boldsymbol{P}_t$
  Parameters *not identifiable* from such observations.

- $L = 2$:
  Regard $\boldsymbol{P}_t$ as probability vector.
  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?

Identifiability: $L = 2$

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: $L = 2$

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: $L = 2$

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

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

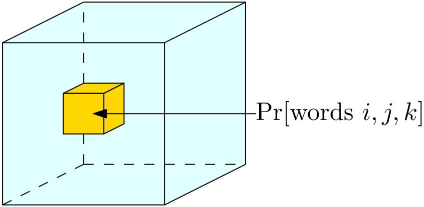
satisfy

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

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 and 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 and 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 and 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*.
- Algorithmic proof via special case of Jennrich's algorithm (Harshman, 1970).

Identifiability from documents of length three

Claim: If $\{\boldsymbol{P}_t\}_{t=1}^K$ are linearly independent and 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*.
- Algorithmic proof via special case of Jennrich's algorithm (Harshman, 1970).

Next: Brief overview of tensors.

Tensors of order two

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

- Think of as *bilinear function* $\boldsymbol{M} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$.

Tensors of order two

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

- Think of as *bilinear function* $\boldsymbol{M} \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$.
- 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): $\boldsymbol{M} \in \mathbb{R}^{d \times d}$.

- Think of as *bilinear function* $\boldsymbol{M} \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$.
- Formula using matrix representation:

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

- Describe $\boldsymbol{M}$ by d^2 values $\boldsymbol{M}(\boldsymbol{e}_i, \boldsymbol{e}_j)$.

Tensors of order two

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

- Think of as *bilinear function* $\boldsymbol{M} \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$.
- Formula using matrix representation:

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

- Describe $\boldsymbol{M}$ by d^2 values $\boldsymbol{M}(\boldsymbol{e}_i, \boldsymbol{e}_j)$.

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

- Describe $\boldsymbol{T}$ by d^p values $\boldsymbol{T}(\boldsymbol{e}_{i_1}, \boldsymbol{e}_{i_2}, \ldots, \boldsymbol{e}_{i_p})$.

Tensors of order p

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

- Describe $\boldsymbol{T}$ by d^p values $\boldsymbol{T}(\boldsymbol{e}_{i_1}, \boldsymbol{e}_{i_2}, \ldots, \boldsymbol{e}_{i_p})$.
- 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**: $\boldsymbol{T} \colon \mathbb{R}^d \times \mathbb{R}^d \times \cdots \times \mathbb{R}^d \to \mathbb{R}$.

- Describe $\boldsymbol{T}$ by d^p values $\boldsymbol{T}(\boldsymbol{e}_{i_1}, \boldsymbol{e}_{i_2}, \ldots, \boldsymbol{e}_{i_p})$.
- Identify $\boldsymbol{T}$ with multi-index array $\boldsymbol{T} \in \mathbb{R}^{d \times d \times \cdots \times d}$.

  Formula for function value:

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

Tensors of order p

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

- Describe $\boldsymbol{T}$ by d^p values $\boldsymbol{T}(\boldsymbol{e}_{i_1}, \boldsymbol{e}_{i_2}, \ldots, \boldsymbol{e}_{i_p})$.
- Identify $\boldsymbol{T}$ with multi-index array $\boldsymbol{T} \in \mathbb{R}^{d \times d \times \cdots \times d}$.

  Formula for function value:

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

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

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

Tensors of order p

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

- Describe $\boldsymbol{T}$ by d^p values $\boldsymbol{T}(\boldsymbol{e}_{i_1}, \boldsymbol{e}_{i_2}, \dots, \boldsymbol{e}_{i_p})$.
- 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.

Jennrich's algorithm (simplified)

Task: Given tensor $\boldsymbol{T} = \sum_{t=1}^{K} \boldsymbol{v}_t^{\otimes 3}$ with linearly independent components $\{\boldsymbol{v}_t\}_{t=1}^{K}$, find the components (up to scaling).

Jennrich's algorithm (simplified)

Task: Given tensor $\boldsymbol{T} = \sum_{t=1}^{K} \boldsymbol{v}_t^{\otimes 3}$ with linearly independent components $\{\boldsymbol{v}_t\}_{t=1}^{K}$, find the components (up to scaling).

Jennrich's algorithm: based on "collapsing" the tensor.

Jennrich's algorithm (simplified)

Task: Given tensor $\boldsymbol{T} = \sum_{t=1}^{K} \boldsymbol{v}_t^{\otimes 3}$ with linearly independent components $\{\boldsymbol{v}_t\}_{t=1}^{K}$, find the components (up to scaling).

Jennrich's algorithm: based on "collapsing" the tensor.

- Think of $\boldsymbol{T} \colon \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as $\boldsymbol{T} \colon \mathbb{R}^d \to \mathbb{R}^{d \times d}$:

$$[\boldsymbol{T}(\boldsymbol{x})]_{j,k} = \boldsymbol{T}(\boldsymbol{x}, \boldsymbol{e}_j, \boldsymbol{e}_k)\,.$$

  (Like "currying" in functional programming.)

Jennrich's algorithm (simplified)

Task: Given tensor $\boldsymbol{T} = \sum_{t=1}^{K} \boldsymbol{v}_t^{\otimes 3}$ with linearly independent components $\{\boldsymbol{v}_t\}_{t=1}^{K}$, find the components (up to scaling).

Jennrich's algorithm: based on "collapsing" the tensor.

- Think of $\boldsymbol{T} \colon \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as $\boldsymbol{T} \colon \mathbb{R}^d \to \mathbb{R}^{d \times d}$:

$$[\boldsymbol{T}(\boldsymbol{x})]_{j,k} = \boldsymbol{T}(\boldsymbol{x}, \boldsymbol{e}_j, \boldsymbol{e}_k) \,.$$

(Like "currying" in functional programming.)

input Tensor $\boldsymbol{T} \in \mathbb{R}^{d \times d \times d}$.

1: Pick $\boldsymbol{x}, \boldsymbol{y}$ independently & uniformly at random from S^{d-1}.

2: Compute and return eigenvectors of $\boldsymbol{T}(\boldsymbol{x})\boldsymbol{T}(\boldsymbol{y})^{\dagger}$ (with non-zero eigenvalues).

Analysis of Jennrich's algorithm

For $\boldsymbol{T} = \sum_{t=1}^{K} \boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t$, linearity of "collapsing" implies

$$\boldsymbol{T}(\boldsymbol{x}) \;=\; \sum_{t=1}^{K} (\boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t)(\boldsymbol{x})$$

Analysis of Jennrich's algorithm

For $\boldsymbol{T} = \sum_{t=1}^{K} \boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t$, linearity of "collapsing" implies

$$\boldsymbol{T}(\boldsymbol{x}) \;=\; \sum_{t=1}^{K} (\boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t)(\boldsymbol{x}) \;=\; \sum_{t=1}^{K} \langle \boldsymbol{v}_t, \boldsymbol{x} \rangle \boldsymbol{v}_t \boldsymbol{v}_t^{\top}$$

Analysis of Jennrich's algorithm

For $\boldsymbol{T} = \sum_{t=1}^{K} \boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t$, linearity of "collapsing" implies

$$\boldsymbol{T}(\boldsymbol{x}) \;=\; \sum_{t=1}^{K} (\boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t)(\boldsymbol{x}) \;=\; \sum_{t=1}^{K} \langle \boldsymbol{v}_t, \boldsymbol{x} \rangle \boldsymbol{v}_t \boldsymbol{v}_t^\top \;=\; \boldsymbol{V} \boldsymbol{D}_{\boldsymbol{x}} \boldsymbol{V}^\top$$

where $\boldsymbol{V} = [\boldsymbol{v}_1 | \cdots | \boldsymbol{v}_K]$ and $\boldsymbol{D}_{\boldsymbol{x}} = \operatorname{diag}(\langle \boldsymbol{v}_1, \boldsymbol{x} \rangle, \ldots, \langle \boldsymbol{v}_K, \boldsymbol{x} \rangle)$.

Analysis of Jennrich's algorithm

For $\boldsymbol{T} = \sum_{t=1}^{K} \boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t$, linearity of "collapsing" implies

$$\boldsymbol{T}(\boldsymbol{x}) \;=\; \sum_{t=1}^{K} (\boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t)(\boldsymbol{x}) \;=\; \sum_{t=1}^{K} \langle \boldsymbol{v}_t, \boldsymbol{x} \rangle \boldsymbol{v}_t \boldsymbol{v}_t^\top \;=\; \boldsymbol{V} \boldsymbol{D}_{\boldsymbol{x}} \boldsymbol{V}^\top$$

where $\boldsymbol{V} = [\boldsymbol{v}_1 | \cdots | \boldsymbol{v}_K]$ and $\boldsymbol{D}_{\boldsymbol{x}} = \operatorname{diag}(\langle \boldsymbol{v}_1, \boldsymbol{x} \rangle, \ldots, \langle \boldsymbol{v}_K, \boldsymbol{x} \rangle)$.

By linear independence of $\{\boldsymbol{v}_t\}_{t=1}^{K}$ and random choice of $\boldsymbol{x}$ and $\boldsymbol{y}$:

Analysis of Jennrich's algorithm

For $\boldsymbol{T} = \sum_{t=1}^{K} \boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t$, linearity of "collapsing" implies

$$\boldsymbol{T}(\boldsymbol{x}) \;=\; \sum_{t=1}^{K} (\boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t)(\boldsymbol{x}) \;=\; \sum_{t=1}^{K} \langle \boldsymbol{v}_t, \boldsymbol{x} \rangle \boldsymbol{v}_t \boldsymbol{v}_t^\top \;=\; \boldsymbol{V} \boldsymbol{D}_{\boldsymbol{x}} \boldsymbol{V}^\top$$

where $\boldsymbol{V} = [\boldsymbol{v}_1 | \cdots | \boldsymbol{v}_K]$ and $\boldsymbol{D}_{\boldsymbol{x}} = \mathrm{diag}(\langle \boldsymbol{v}_1, \boldsymbol{x} \rangle, \ldots, \langle \boldsymbol{v}_K, \boldsymbol{x} \rangle)$.

By linear independence of $\{\boldsymbol{v}_t\}_{t=1}^{K}$ and random choice of $\boldsymbol{x}$ and $\boldsymbol{y}$:

1. $\boldsymbol{V}$ has rank K;

Analysis of Jennrich's algorithm

For $\boldsymbol{T} = \sum_{t=1}^{K} \boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t$, linearity of "collapsing" implies

$$\boldsymbol{T}(\boldsymbol{x}) \;=\; \sum_{t=1}^{K} (\boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t)(\boldsymbol{x}) \;=\; \sum_{t=1}^{K} \langle \boldsymbol{v}_t, \boldsymbol{x} \rangle \boldsymbol{v}_t \boldsymbol{v}_t^\top \;=\; \boldsymbol{V} \boldsymbol{D}_{\boldsymbol{x}} \boldsymbol{V}^\top$$

where $\boldsymbol{V} = [\boldsymbol{v}_1 | \cdots | \boldsymbol{v}_K]$ and $\boldsymbol{D}_{\boldsymbol{x}} = \operatorname{diag}(\langle \boldsymbol{v}_1, \boldsymbol{x} \rangle, \ldots, \langle \boldsymbol{v}_K, \boldsymbol{x} \rangle)$.

By linear independence of $\{\boldsymbol{v}_t\}_{t=1}^{K}$ and random choice of $\boldsymbol{x}$ and $\boldsymbol{y}$:

1. $\boldsymbol{V}$ has rank K;
2. $\boldsymbol{D}_{\boldsymbol{x}}$ and $\boldsymbol{D}_{\boldsymbol{y}}$ are invertible (a.s.);

Analysis of Jennrich's algorithm

For $\boldsymbol{T} = \sum_{t=1}^{K} \boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t$, linearity of "collapsing" implies

$$\boldsymbol{T}(\boldsymbol{x}) \;=\; \sum_{t=1}^{K} (\boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t)(\boldsymbol{x}) \;=\; \sum_{t=1}^{K} \langle \boldsymbol{v}_t, \boldsymbol{x} \rangle \boldsymbol{v}_t \boldsymbol{v}_t^\top \;=\; \boldsymbol{V} \boldsymbol{D}_{\boldsymbol{x}} \boldsymbol{V}^\top$$

where $\boldsymbol{V} = [\boldsymbol{v}_1 | \cdots | \boldsymbol{v}_K]$ and $\boldsymbol{D}_{\boldsymbol{x}} = \operatorname{diag}(\langle \boldsymbol{v}_1, \boldsymbol{x} \rangle, \dots, \langle \boldsymbol{v}_K, \boldsymbol{x} \rangle)$.

By linear independence of $\{\boldsymbol{v}_t\}_{t=1}^{K}$ and random choice of $\boldsymbol{x}$ and $\boldsymbol{y}$:

1. $\boldsymbol{V}$ has rank K;
2. $\boldsymbol{D}_{\boldsymbol{x}}$ and $\boldsymbol{D}_{\boldsymbol{y}}$ are invertible (a.s.);
3. diagonal entries of $\boldsymbol{D}_{\boldsymbol{x}} \boldsymbol{D}_{\boldsymbol{y}}^{-1}$ are distinct (a.s.);

Analysis of Jennrich's algorithm

For $\boldsymbol{T} = \sum_{t=1}^{K} \boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t$, linearity of "collapsing" implies

$$\boldsymbol{T}(\boldsymbol{x}) \;=\; \sum_{t=1}^{K} (\boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t)(\boldsymbol{x}) \;=\; \sum_{t=1}^{K} \langle \boldsymbol{v}_t, \boldsymbol{x} \rangle \boldsymbol{v}_t \boldsymbol{v}_t^\top \;=\; \boldsymbol{V} \boldsymbol{D}_{\boldsymbol{x}} \boldsymbol{V}^\top$$

where $\boldsymbol{V} = [\boldsymbol{v}_1 | \cdots | \boldsymbol{v}_K]$ and $\boldsymbol{D}_{\boldsymbol{x}} = \operatorname{diag}(\langle \boldsymbol{v}_1, \boldsymbol{x} \rangle, \dots, \langle \boldsymbol{v}_K, \boldsymbol{x} \rangle)$.

By linear independence of $\{\boldsymbol{v}_t\}_{t=1}^{K}$ and random choice of $\boldsymbol{x}$ and $\boldsymbol{y}$:

1. $\boldsymbol{V}$ has rank K;
2. $\boldsymbol{D}_{\boldsymbol{x}}$ and $\boldsymbol{D}_{\boldsymbol{y}}$ are invertible (a.s.);
3. diagonal entries of $\boldsymbol{D}_{\boldsymbol{x}} \boldsymbol{D}_{\boldsymbol{y}}^{-1}$ are distinct (a.s.);
4. $\boldsymbol{T}(\boldsymbol{x}) \boldsymbol{T}(\boldsymbol{y})^\dagger = \boldsymbol{V} (\boldsymbol{D}_{\boldsymbol{x}} \boldsymbol{D}_{\boldsymbol{y}}^{-1}) \boldsymbol{V}^\dagger$ (a.s.).

Analysis of Jennrich's algorithm

For $\boldsymbol{T} = \sum_{t=1}^{K} \boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t$, linearity of "collapsing" implies

$$\boldsymbol{T}(\boldsymbol{x}) \;=\; \sum_{t=1}^{K} (\boldsymbol{v}_t \otimes \boldsymbol{v}_t \otimes \boldsymbol{v}_t)(\boldsymbol{x}) \;=\; \sum_{t=1}^{K} \langle \boldsymbol{v}_t, \boldsymbol{x} \rangle \boldsymbol{v}_t \boldsymbol{v}_t^\top \;=\; \boldsymbol{V} \boldsymbol{D}_{\boldsymbol{x}} \boldsymbol{V}^\top$$

where $\boldsymbol{V} = [\boldsymbol{v}_1 | \cdots | \boldsymbol{v}_K]$ and $\boldsymbol{D}_{\boldsymbol{x}} = \operatorname{diag}(\langle \boldsymbol{v}_1, \boldsymbol{x} \rangle, \dots, \langle \boldsymbol{v}_K, \boldsymbol{x} \rangle)$.

By linear independence of $\{\boldsymbol{v}_t\}_{t=1}^{K}$ and random choice of $\boldsymbol{x}$ and $\boldsymbol{y}$:

1. $\boldsymbol{V}$ has rank K;
2. $\boldsymbol{D}_{\boldsymbol{x}}$ and $\boldsymbol{D}_{\boldsymbol{y}}$ are invertible (a.s.);
3. diagonal entries of $\boldsymbol{D}_{\boldsymbol{x}} \boldsymbol{D}_{\boldsymbol{y}}^{-1}$ are distinct (a.s.);
4. $\boldsymbol{T}(\boldsymbol{x}) \boldsymbol{T}(\boldsymbol{y})^\dagger = \boldsymbol{V} (\boldsymbol{D}_{\boldsymbol{x}} \boldsymbol{D}_{\boldsymbol{y}}^{-1}) \boldsymbol{V}^\dagger$ (a.s.).

So $\{\boldsymbol{v}_t\}_{t=1}^{K}$ are the eigenvectors of $\boldsymbol{T}(\boldsymbol{x}) \boldsymbol{T}(\boldsymbol{y})^\dagger$ with distinct non-zero eigenvalues. □

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 pre-condition for Jennrich's algorithm:

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

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 pre-condition for Jennrich's algorithm:

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

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

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 pre-condition for Jennrich's algorithm:

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

- 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 $\boldsymbol{T}$.

□

Recap

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

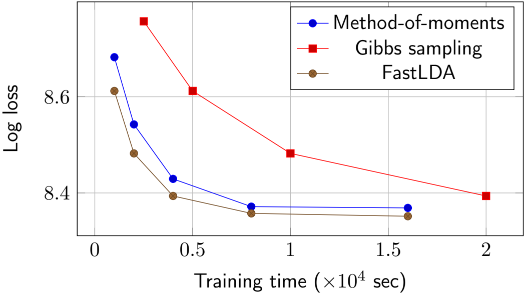
Recap

- 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.

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.

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*?

     Instead of Jennrich's algorithm, use more error-tolerant decomposition algorithm (also computationally efficient).

  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*?

     Instead of Jennrich's algorithm, use more error-tolerant decomposition algorithm (also computationally efficient).

  3. *Generality* beyond simple topic models?

Next: Moment decompositions for other models.

2. Moment decompositions for other models

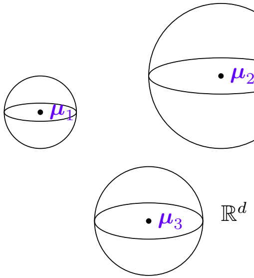
Moment decompositions

Some examples of usable moment decompositions.

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

Mixtures of spherical Gaussians

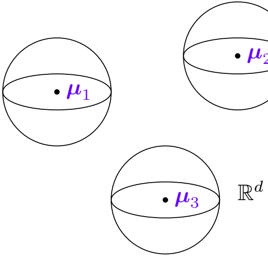
$$\begin{aligned} H &\sim \text{Discrete}(\pi_1, \pi_2, \ldots, \pi_K) \quad \text{(hidden)}\,; \\ \boldsymbol{X} \mid H = t &\sim \text{Normal}(\boldsymbol{\mu}_t, \sigma_t^2 \boldsymbol{I}_d)\,, \quad t \in [K]\,. \end{aligned}$$



Mixtures of spherical Gaussians

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

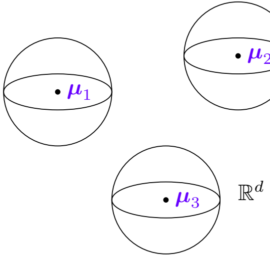
(For simplicity, restrict $\sigma_1 = \sigma_2 = \cdots = \sigma_K = \sigma$.)



Mixtures of spherical Gaussians

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

(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 $\boldsymbol{Y}$).

Moments for spherical Gaussian mixtures

First- and second-order moments:

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

Moments for spherical Gaussian mixtures

First- and second-order moments:

$$\mathbb{E}(\boldsymbol{X}) = \sum_{t=1}^{K} \pi_t \cdot \boldsymbol{\mu}_t \,,$$

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

(Vempala & Wang, 2002):
Span of top K eigenvectors of $\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X})$ contains $\{\boldsymbol{\mu}_t\}_{t=1}^{K}$.
$\rightarrow$ Principal component analysis (PCA).

Use of moments for mixtures of spherical Gaussians

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

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

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, 2007):
  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

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

- (Dasgupta & Schulman, 2000, 2007):
  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, 2007):
  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, 2007):
  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 $\Omega(K)$th-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 $\boldsymbol{Z} \sim \mathrm{Normal}(\boldsymbol{0}, \boldsymbol{I}_d)$ (indep. of $\boldsymbol{Y}$).

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 $\boldsymbol{Z} \sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{I}_d)$ (indep. of $\boldsymbol{Y}$).

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 $\boldsymbol{Z} \sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{I}_d)$ (indep. of $\boldsymbol{Y}$).

Third-order moment tensor:

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

Tensor decomposition for spherical Gaussian mixtures

(H. & Kakade, 2013)

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

Tensor decomposition for spherical Gaussian mixtures

(H. & Kakade, 2013)

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

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

Tensor decomposition for spherical Gaussian mixtures

(H. & Kakade, 2013)

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

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

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

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

Tensor decomposition for spherical Gaussian mixtures

(H. & Kakade, 2013)

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

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

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

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

Can use, e.g., Jennrich's algorithm 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 $K \leq d$.

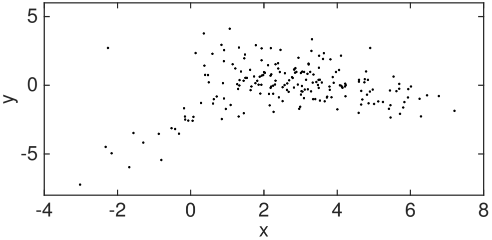
- (Anderson, Belkin, Goyal, Rademacher, & Voss, 2014), (Bhaskara, Charikar, Moitra, & Vijayaraghavan, 2014)
  Mixtures of $d^{O(1)}$ Gaussians (w/ simple or known covariance) via **smoothed analysis** and $O(1)$-order moments.
- (Ge, Huang, & Kakade, 2015)
  Also with **arbitrary unknown covariances**.

Mixtures of linear regressions

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

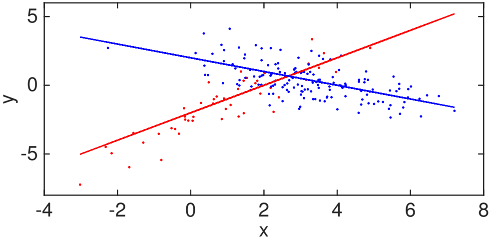
Mixtures of linear regressions

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



Mixtures of linear regressions

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



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 $\boldsymbol{X} \sim \mathrm{Normal}(\boldsymbol{0}, \boldsymbol{I}_d)$):

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

- 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).

Simpler setting: mixed random linear equations

(Yi, Caramanis, & Sanghavi, 2016)

$$\begin{aligned} H &\sim \mathrm{Discrete}(\pi_1, \pi_2, \ldots, \pi_K) \quad \text{(hidden)}\,; \\ \boldsymbol{X} &\sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{I}_d)\,; \\ Y &= \langle \boldsymbol{\beta}_H, \boldsymbol{X} \rangle\,. \end{aligned}$$

Simpler setting: mixed random linear equations

(Yi, Caramanis, & Sanghavi, 2016)

$$\begin{aligned} H &\sim \mathrm{Discrete}(\pi_1, \pi_2, \ldots, \pi_K) \quad \text{(hidden)}\,; \\ \boldsymbol{X} &\sim \mathrm{Normal}(\boldsymbol{0}, \boldsymbol{I}_d)\,; \\ Y &= \langle \boldsymbol{\beta}_H, \boldsymbol{X} \rangle\,. \end{aligned}$$

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

$$\boldsymbol{T} \; := \; \mathbb{E}\left(Y^3 \boldsymbol{X}^{\otimes 3}\right) \; = \; 6 \sum_{t=1}^K \pi_t \cdot \boldsymbol{\beta}_t^{\otimes 3} + \underbrace{\tau(\mathbb{E}\, Y^3 \boldsymbol{X})}_{\text{some tensor}}\,.$$

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.

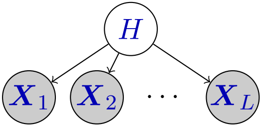
Recap: mixtures of Gaussians and linear regressions

- 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



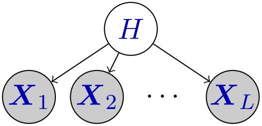
K topics (dists. over words) $\{\boldsymbol{P}_t\}_{t=1}^{K}$.

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**).

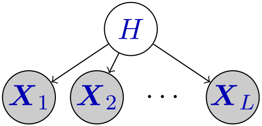
Word tokens $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_L \overset{\text{iid}}{\sim} \boldsymbol{P}_H$.

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

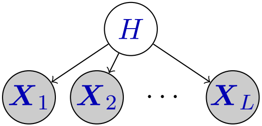
Key property:

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

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

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:

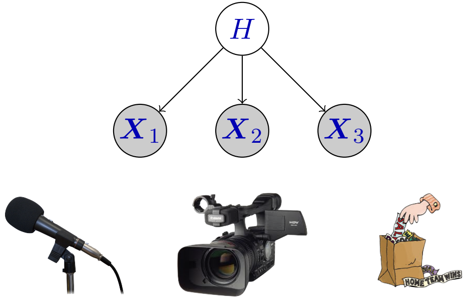
$\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_L$ conditionally independent given H.

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

Some previous theoretical analysis:

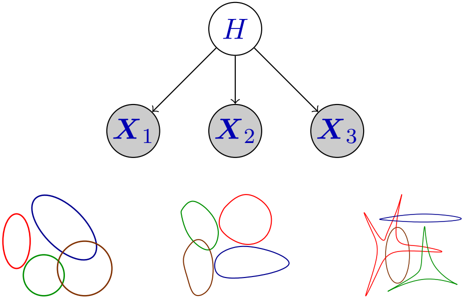
- (Blum & Mitchell, 1998)
  Co-training in semi-supervised learning.
- (Chaudhuri, Kakade, Livescu, & Sridharan, 2009)
  Multi-view Gaussian mixture models.

Multi-view mixture model



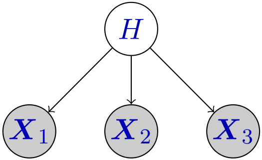
View 1: $\boldsymbol{X}_1$ View 2: $\boldsymbol{X}_2$ View 3: $\boldsymbol{X}_3$

Multi-view mixture model



View 1: $\boldsymbol{X}_1$ View 2: $\boldsymbol{X}_2$ View 3: $\boldsymbol{X}_3$

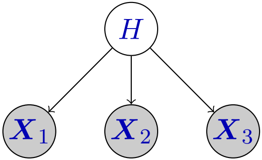
Multi-view mixture model



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

$$\begin{aligned} \text{where } \boldsymbol{\mu}_t^{(i)} &= \mathbb{E}[\boldsymbol{X}_i \mid H = t], \\ \pi_t &= \Pr(H = t). \end{aligned}$$

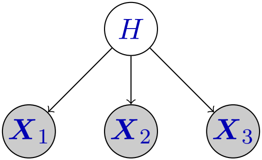
Multi-view mixture model



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

Jennrich's algorithm works in this asymmetric case provided $\{\boldsymbol{\mu}_t^{(j)}\}_{t=1}^{K}$ are linearly independent for each j, and all $\pi_t > 0$.

Multi-view mixture model



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

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

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

Jennrich's algorithm works in this asymmetric case provided $\{\boldsymbol{\mu}_t^{(j)}\}_{t=1}^{K}$ are linearly independent for each j, and all $\pi_t > 0$.

(Also possible to "symmetrize" using second-order moments.)

Examples of multi-view mixture models

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

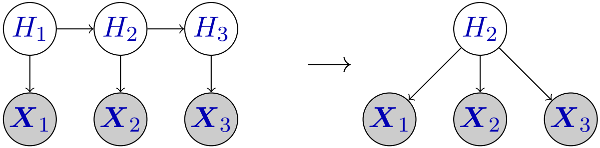
1. Mixtures of high-dimensional product distributions.
   (E.g., mixtures of axis-aligned Gaussians.)

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.)

2. Hidden Markov models.

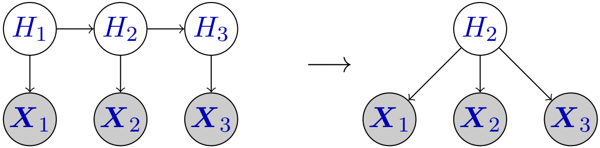


Examples of multi-view mixture models

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

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

2. Hidden Markov models.



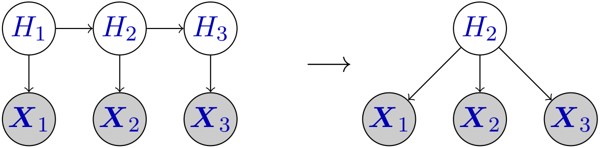
3. Phylogenetic trees.
   - $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$: genes of three extant species.
   - H: LCA of most closely related pair of species.

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.)

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. ...

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 $\rightarrow$ empirical moment tensor $\widehat{\boldsymbol{T}}$.
  2. *Approximately decompose* $\widehat{\boldsymbol{T}}$ $\rightarrow$ parameter estimate $\hat{\boldsymbol{\theta}}$.

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 $\rightarrow$ empirical moment tensor $\widehat{\boldsymbol{T}}$.
  2. *Approximately decompose* $\widehat{\boldsymbol{T}}$ $\rightarrow$ parameter estimate $\hat{\boldsymbol{\theta}}$.

Next: Error-tolerant *(approximate) tensor decomposition*.

3. Error-tolerant algorithms for tensor decompositions

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.,

$$\|\boldsymbol{T}\| \;:=\; \sup_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}\in S^{d-1}} \boldsymbol{T}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) \quad \text{(operator norm)}\,,$$

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

Using Jennrich's algorithm

Recall: Jennrich's algorithm (simplified)

Goal: Given tensor $\boldsymbol{T} = \sum_{t=1}^{K} \boldsymbol{v}_t^{\otimes 3}$, find components $\{\boldsymbol{v}_t\}_{t=1}^{K}$.

---

input Tensor $\boldsymbol{T} \in \mathbb{R}^{d \times d \times d}$.

1: Pick $\boldsymbol{x}, \boldsymbol{y}$ independently & uniformly at random from S^{d-1}.

2: Compute and return eigenvectors of $\boldsymbol{T}(\boldsymbol{x})\boldsymbol{T}(\boldsymbol{y})^{\dagger}$ (with non-zero eigenvalues).

Using Jennrich's algorithm

Recall: Jennrich's algorithm (simplified)

<u>Goal</u>: Given tensor $\boldsymbol{T} = \sum_{t=1}^{K} \boldsymbol{v}_t^{\otimes 3}$, find components $\{\boldsymbol{v}_t\}_{t=1}^{K}$.

---

input Tensor $\boldsymbol{T} \in \mathbb{R}^{d \times d \times d}$.

1: Pick $\boldsymbol{x}, \boldsymbol{y}$ independently & uniformly at random from S^{d-1}.

2: Compute and return eigenvectors of $\boldsymbol{T}(\boldsymbol{x})\boldsymbol{T}(\boldsymbol{y})^{\dagger}$ (with non-zero eigenvalues).

But we only have $\widehat{\boldsymbol{T}}$, an estimate of $\boldsymbol{T} = \sum_{t=1}^{K} \boldsymbol{v}_t^{\otimes 3}$ with (say)

$$\|\widehat{\boldsymbol{T}} - \boldsymbol{T}\| \lesssim n^{-1/2}.$$

Stability of Jennrich's algorithm

Stability of eigenvectors requires eigenvalue gaps.

Stability of Jennrich's algorithm

Stability of eigenvectors requires eigenvalue gaps.

- Eigenvalue gaps for $\boldsymbol{T}(\boldsymbol{x})\boldsymbol{T}(\boldsymbol{y})^\dagger$:

$$\Delta \ := \ \min_{i \neq j} \left| \frac{\langle \boldsymbol{v}_i, \boldsymbol{x} \rangle}{\langle \boldsymbol{v}_i, \boldsymbol{y} \rangle} - \frac{\langle \boldsymbol{v}_j, \boldsymbol{x} \rangle}{\langle \boldsymbol{v}_j, \boldsymbol{y} \rangle} \right| .$$

Stability of Jennrich's algorithm

Stability of eigenvectors requires eigenvalue gaps.

- Eigenvalue gaps for $\boldsymbol{T}(\boldsymbol{x})\boldsymbol{T}(\boldsymbol{y})^\dagger$:

$$\Delta \; := \; \min_{i \neq j} \left| \frac{\langle \boldsymbol{v}_i, \boldsymbol{x} \rangle}{\langle \boldsymbol{v}_i, \boldsymbol{y} \rangle} - \frac{\langle \boldsymbol{v}_j, \boldsymbol{x} \rangle}{\langle \boldsymbol{v}_j, \boldsymbol{y} \rangle} \right| .$$

- Need $\|\widehat{\boldsymbol{T}}(\boldsymbol{x})\widehat{\boldsymbol{T}}(\boldsymbol{y})^\dagger - \boldsymbol{T}(\boldsymbol{x})\boldsymbol{T}(\boldsymbol{y})^\dagger\| \ll \Delta$ so that $\widehat{\boldsymbol{T}}(\boldsymbol{x})\widehat{\boldsymbol{T}}(\boldsymbol{y})^\dagger$ also has sufficient eigenvalue gaps.

Stability of Jennrich's algorithm

Stability of eigenvectors requires eigenvalue gaps.

- Eigenvalue gaps for $\boldsymbol{T}(\boldsymbol{x})\boldsymbol{T}(\boldsymbol{y})^\dagger$:

$$\Delta \;:=\; \min_{i \neq j} \left| \frac{\langle \boldsymbol{v}_i, \boldsymbol{x} \rangle}{\langle \boldsymbol{v}_i, \boldsymbol{y} \rangle} - \frac{\langle \boldsymbol{v}_j, \boldsymbol{x} \rangle}{\langle \boldsymbol{v}_j, \boldsymbol{y} \rangle} \right| .$$

- Need $\|\widehat{\boldsymbol{T}}(\boldsymbol{x})\widehat{\boldsymbol{T}}(\boldsymbol{y})^\dagger - \boldsymbol{T}(\boldsymbol{x})\boldsymbol{T}(\boldsymbol{y})^\dagger\| \ll \Delta$ so that $\widehat{\boldsymbol{T}}(\boldsymbol{x})\widehat{\boldsymbol{T}}(\boldsymbol{y})^\dagger$ also has sufficient eigenvalue gaps.

- Ultimately, appears to need $\|\widehat{\boldsymbol{T}} - \boldsymbol{T}\|_F \ll \frac{1}{\mathrm{poly}(d)}$.

Stability of Jennrich's algorithm

Stability of eigenvectors requires eigenvalue gaps.

- Eigenvalue gaps for $\boldsymbol{T}(\boldsymbol{x})\boldsymbol{T}(\boldsymbol{y})^\dagger$:

$$\Delta := \min_{i \neq j} \left| \frac{\langle \boldsymbol{v}_i, \boldsymbol{x} \rangle}{\langle \boldsymbol{v}_i, \boldsymbol{y} \rangle} - \frac{\langle \boldsymbol{v}_j, \boldsymbol{x} \rangle}{\langle \boldsymbol{v}_j, \boldsymbol{y} \rangle} \right| .$$

- Need $\|\widehat{\boldsymbol{T}}(\boldsymbol{x})\widehat{\boldsymbol{T}}(\boldsymbol{y})^\dagger - \boldsymbol{T}(\boldsymbol{x})\boldsymbol{T}(\boldsymbol{y})^\dagger\| \ll \Delta$ so that $\widehat{\boldsymbol{T}}(\boldsymbol{x})\widehat{\boldsymbol{T}}(\boldsymbol{y})^\dagger$ also has sufficient eigenvalue gaps.
- Ultimately, appears to need $\|\widehat{\boldsymbol{T}} - \boldsymbol{T}\|_F \ll \frac{1}{\mathrm{poly}(d)}$.

Next: A different approach.

Reduction to orthonormal case

In many (all?) applications, we can estimate 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

In many (all?) applications, we can estimate 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.

Reduction to orthonormal case

In many (all?) applications, we can estimate 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**.

(Nearly) orthogonally decomposable tensors $(d = K)$

<u>Goal</u>: Given tensor $\widehat{\boldsymbol{T}} \in \mathbb{R}^{d \times d \times d}$ such that $\|\widehat{\boldsymbol{T}} - \boldsymbol{T}\| \leq \varepsilon$ for some $\boldsymbol{T} = \sum_{t=1}^{d} \lambda_t \cdot \boldsymbol{v}_t^{\otimes 3}$ where $\{\boldsymbol{v}_t\}_{t=1}^{d}$ are orthonormal and all $\lambda_t > 0$, approximately recover $\{(\boldsymbol{v}_t, \lambda_t)\}_{t=1}^{d}$.

(Nearly) orthogonally decomposable tensors ($d = K$)

<u>Goal</u>: Given tensor $\widehat{\boldsymbol{T}} \in \mathbb{R}^{d \times d \times d}$ such that $\|\widehat{\boldsymbol{T}} - \boldsymbol{T}\| \leq \varepsilon$ for some $\boldsymbol{T} = \sum_{t=1}^{d} \lambda_t \cdot \boldsymbol{v}_t^{\otimes 3}$ where $\{\boldsymbol{v}_t\}_{t=1}^{d}$ are orthonormal and all $\lambda_t > 0$, approximately recover $\{(\boldsymbol{v}_t, \lambda_t)\}_{t=1}^{d}$.

Analogous matrix problems:

- $\varepsilon = 0$: eigendecomposition.
  ("Promised" decomposition always exists by symmetry.)

(Nearly) orthogonally decomposable tensors ($d = K$)

<u>Goal</u>: Given tensor $\widehat{\boldsymbol{T}} \in \mathbb{R}^{d \times d \times d}$ such that $\|\widehat{\boldsymbol{T}} - \boldsymbol{T}\| \leq \varepsilon$ for some $\boldsymbol{T} = \sum_{t=1}^{d} \lambda_t \cdot \boldsymbol{v}_t^{\otimes 3}$ where $\{\boldsymbol{v}_t\}_{t=1}^{d}$ are orthonormal and all $\lambda_t > 0$, approximately recover $\{(\boldsymbol{v}_t, \lambda_t)\}_{t=1}^{d}$.

Analogous matrix problems:

- $\varepsilon = 0$: eigendecomposition.
  ("Promised" decomposition always exists by symmetry.)

  Decomposition is unique iff the $\{\lambda_t\}_{t=1}^{d}$ are distinct.

(Nearly) orthogonally decomposable tensors ($d = K$)

<u>Goal</u>: Given tensor $\widehat{\boldsymbol{T}} \in \mathbb{R}^{d \times d \times d}$ such that $\|\widehat{\boldsymbol{T}} - \boldsymbol{T}\| \leq \varepsilon$ for some $\boldsymbol{T} = \sum_{t=1}^{d} \lambda_t \cdot \boldsymbol{v}_t^{\otimes 3}$ where $\{\boldsymbol{v}_t\}_{t=1}^{d}$ are orthonormal and all $\lambda_t > 0$, approximately recover $\{(\boldsymbol{v}_t, \lambda_t)\}_{t=1}^{d}$.

Analogous matrix problems:

- $\varepsilon = 0$: eigendecomposition.
  ("Promised" decomposition always exists by symmetry.)

  Decomposition is unique iff the $\{\lambda_t\}_{t=1}^{d}$ are distinct.

- $\varepsilon > 0$: perturbation theory for eigenvalues (Weyl) and eigenvectors (Davis & Kahan).

Exact orthogonally decomposable tensor

(Zhang & Golub, 2001)

For now assume $\varepsilon = 0$, so $\widehat{\boldsymbol{T}} = \boldsymbol{T}$.

Matching moments:

$$\{(\hat{\boldsymbol{v}}_t, \hat{\lambda}_t)\}_{t=1}^{d} \quad := \quad \operatorname*{arg\,min}_{\{(\boldsymbol{x}_t, \sigma_t)\}_{t=1}^{d}} \left\| \boldsymbol{T} - \sum_{t=1}^{d} \sigma_t \cdot \boldsymbol{x}_t^{\otimes 3} \right\|_F^2 .$$

Exact orthogonally decomposable tensor

(Zhang & Golub, 2001)

For now assume $\varepsilon = 0$, so $\widehat{\boldsymbol{T}} = \boldsymbol{T}$.

Matching moments:

$$\{(\hat{\boldsymbol{v}}_t, \hat{\lambda}_t)\}_{t=1}^{d} \ := \ \operatorname*{arg\,min}_{\{(\boldsymbol{x}_t, \sigma_t)\}_{t=1}^{d}} \left\| \boldsymbol{T} - \sum_{t=1}^{d} \sigma_t \cdot \boldsymbol{x}_t^{\otimes 3} \right\|_F^2 .$$

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

$$(\hat{\boldsymbol{v}}, \hat{\lambda}) \ := \ \operatorname*{arg\,min}_{(\boldsymbol{x}, \sigma) \in S^{d-1} \times \mathbb{R}_+} \left\| \boldsymbol{T} - \sigma \cdot \boldsymbol{x}^{\otimes 3} \right\|_F^2 .$$

  - "Deflate" $\boldsymbol{T} := \boldsymbol{T} - \hat{\lambda} \cdot \hat{\boldsymbol{v}}^{\otimes 3}$ and repeat.

Exact orthogonally decomposable tensor

(Zhang & Golub, 2001)

For now assume $\varepsilon = 0$, so $\widehat{\boldsymbol{T}} = \boldsymbol{T}$.

Matching moments:

$$\{(\hat{\boldsymbol{v}}_t, \hat{\lambda}_t)\}_{t=1}^d \;:=\; \operatorname*{arg\,min}_{\{(\boldsymbol{x}_t,\sigma_t)\}_{t=1}^d} \left\| \boldsymbol{T} - \sum_{t=1}^d \sigma_t \cdot \boldsymbol{x}_t^{\otimes 3} \right\|_F^2 .$$

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

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

  - "Deflate" $\boldsymbol{T} := \boldsymbol{T} - \hat{\lambda} \cdot \hat{\boldsymbol{v}}^{\otimes 3}$ 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

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

Algorithm: use gradient ascent to find each component $\boldsymbol{v}_t$.

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

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

Algorithm: use gradient ascent to find each component $\boldsymbol{v}_t$.

Next: "Parameter-free" fixed-point algorithm.

Fixed-point algorithm

(De Lathauwer, De Moore, & Vandewalle, 2000)

First-order (necessary but not sufficient) optimality condition:

$$\nabla_{\boldsymbol{x}} \boldsymbol{T}(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{x}) \; = \; \lambda \boldsymbol{x} \, .$$

Fixed-point algorithm

(De Lathauwer, De Moore, & Vandewalle, 2000)

First-order (necessary but not sufficient) optimality condition:

$$\nabla_{\boldsymbol{x}} \boldsymbol{T}(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{x}) \;=\; \lambda \boldsymbol{x}\,.$$

Gradient is "partial evaluation" of $\boldsymbol{T}$:

$$\nabla_{\boldsymbol{x}} \boldsymbol{T}(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{x}) \;=\; 3\sum_{i,j} T_{i,j,k} \cdot x_i x_j \boldsymbol{e}_k \;=\; 3\boldsymbol{T}(\boldsymbol{x}, \boldsymbol{x}, \cdot)\,.$$

Fixed-point algorithm

(De Lathauwer, De Moore, & Vandewalle, 2000)

First-order (necessary but not sufficient) optimality condition:

$$\nabla_{\boldsymbol{x}} \boldsymbol{T}(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{x}) \; = \; \lambda \boldsymbol{x} \, .$$

Gradient is "partial evaluation" of $\boldsymbol{T}$:

$$\nabla_{\boldsymbol{x}} \boldsymbol{T}(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{x}) \; = \; 3 \sum_{i,j} T_{i,j,k} \cdot x_i x_j \boldsymbol{e}_k \; = \; 3\boldsymbol{T}(\boldsymbol{x}, \boldsymbol{x}, \cdot) \, .$$

(Third-order) **tensor power iteration**:

$$\text{For } i = 1, 2, \ldots : \quad \boldsymbol{x}^{(i+1)} \; := \; \frac{\boldsymbol{T}(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(i)}, \cdot)}{\|\boldsymbol{T}(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(i)}, \cdot)\|} \, .$$

Comparison to matrix power iteration

Matrix power iteration $\boldsymbol{x}^{(i+1)} = \frac{\boldsymbol{M}\boldsymbol{x}^{(i)}}{\|\boldsymbol{M}\boldsymbol{x}^{(i)}\|}$ for $\boldsymbol{M} = \sum_t \lambda_t \boldsymbol{v}_t \boldsymbol{v}_t^\top$.

Comparison to matrix power iteration

Matrix power iteration $\boldsymbol{x}^{(i+1)} = \frac{\boldsymbol{M}\boldsymbol{x}^{(i)}}{\|\boldsymbol{M}\boldsymbol{x}^{(i)}\|}$ for $\boldsymbol{M} = \sum_t \lambda_t \boldsymbol{v}_t \boldsymbol{v}_t^\top$.

- Requires gap $\min_{i \neq 1} 1 - \lambda_i/\lambda_1 > 0$ to converge to $\boldsymbol{v}_1$.

Comparison to matrix power iteration

Matrix power iteration $\boldsymbol{x}^{(i+1)} = \frac{\boldsymbol{M}\boldsymbol{x}^{(i)}}{\|\boldsymbol{M}\boldsymbol{x}^{(i)}\|}$ for $\boldsymbol{M} = \sum_t \lambda_t \boldsymbol{v}_t \boldsymbol{v}_t^\top$.

- Requires gap $\min_{i \neq 1} 1 - \lambda_i/\lambda_1 > 0$ to converge to $\boldsymbol{v}_1$.

  Tensor power iteration:
  No gap required.

Comparison to matrix power iteration

Matrix power iteration $\boldsymbol{x}^{(i+1)} = \frac{\boldsymbol{M}\boldsymbol{x}^{(i)}}{\|\boldsymbol{M}\boldsymbol{x}^{(i)}\|}$ for $\boldsymbol{M} = \sum_t \lambda_t \boldsymbol{v}_t \boldsymbol{v}_t^\top$.

- Requires gap $\min_{i \neq 1} 1 - \lambda_i / \lambda_1 > 0$ to converge to $\boldsymbol{v}_1$.

  Tensor power iteration:
  No gap required.

- If $\langle \boldsymbol{v}_1, \boldsymbol{x}^{(0)} \rangle \neq 0$ (and gap > 0), converges to $\boldsymbol{v}_1$.

Comparison to matrix power iteration

Matrix power iteration $\boldsymbol{x}^{(i+1)} = \frac{\boldsymbol{M}\boldsymbol{x}^{(i)}}{\|\boldsymbol{M}\boldsymbol{x}^{(i)}\|}$ for $\boldsymbol{M} = \sum_t \lambda_t \boldsymbol{v}_t \boldsymbol{v}_t^\top$.

- Requires gap $\min_{i \neq 1} 1 - \lambda_i/\lambda_1 > 0$ to converge to $\boldsymbol{v}_1$.

  Tensor power iteration:
  No gap required.

- If $\langle \boldsymbol{v}_1, \boldsymbol{x}^{(0)} \rangle \neq 0$ (and gap > 0), converges to $\boldsymbol{v}_1$.

  Tensor power iteration:
  If $t := \arg\max_{t'} \lambda_{t'} |\langle \boldsymbol{v}_{t'}, \boldsymbol{x}^{(0)} \rangle|$, converges to $\boldsymbol{v}_t$.

Comparison to matrix power iteration

Matrix power iteration $\boldsymbol{x}^{(i+1)} = \frac{\boldsymbol{M}\boldsymbol{x}^{(i)}}{\|\boldsymbol{M}\boldsymbol{x}^{(i)}\|}$ for $\boldsymbol{M} = \sum_t \lambda_t \boldsymbol{v}_t \boldsymbol{v}_t^\top$.

- Requires gap $\min_{i \neq 1} 1 - \lambda_i / \lambda_1 > 0$ to converge to $\boldsymbol{v}_1$.

  Tensor power iteration:
  No gap required.

- If $\langle \boldsymbol{v}_1, \boldsymbol{x}^{(0)} \rangle \neq 0$ (and gap > 0), converges to $\boldsymbol{v}_1$.

  Tensor power iteration:
  If $t := \arg\max_{t'} \lambda_{t'} |\langle \boldsymbol{v}_{t'}, \boldsymbol{x}^{(0)} \rangle|$, converges to $\boldsymbol{v}_t$.

- Converges at linear rate.

Comparison to matrix power iteration

Matrix power iteration $\boldsymbol{x}^{(i+1)} = \frac{\boldsymbol{M}\boldsymbol{x}^{(i)}}{\|\boldsymbol{M}\boldsymbol{x}^{(i)}\|}$ for $\boldsymbol{M} = \sum_t \lambda_t \boldsymbol{v}_t \boldsymbol{v}_t^\top$.

- Requires gap $\min_{i \neq 1} 1 - \lambda_i / \lambda_1 > 0$ to converge to $\boldsymbol{v}_1$.

  Tensor power iteration:
  No gap required.

- If $\langle \boldsymbol{v}_1, \boldsymbol{x}^{(0)} \rangle \neq 0$ (and gap > 0), converges to $\boldsymbol{v}_1$.

  Tensor power iteration:
  If $t := \arg\max_{t'} \lambda_{t'} |\langle \boldsymbol{v}_{t'}, \boldsymbol{x}^{(0)} \rangle|$, converges to $\boldsymbol{v}_t$.

- Converges at linear rate.

  Tensor power iteration:
  Converges at quadratic rate.

Nearly orthogonally decomposable tensor

(Mu, H., & Goldfarb, 2015)

Now allow $\varepsilon = \|\boldsymbol{E}\| > 0$, for $\boldsymbol{E} := \widehat{\boldsymbol{T}} - \boldsymbol{T}$.

Claim: Let $\hat{\boldsymbol{v}} := \arg\max_{\boldsymbol{x} \in S^{d-1}} \widehat{\boldsymbol{T}}(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{x})$ 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)

Now allow $\varepsilon = \|\boldsymbol{E}\| > 0$, for $\boldsymbol{E} := \widehat{\boldsymbol{T}} - \boldsymbol{T}$.

Claim: Let $\hat{\boldsymbol{v}} := \arg\max_{\boldsymbol{x} \in S^{d-1}} \widehat{\boldsymbol{T}}(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{x})$ 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).

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 $\boldsymbol{v}_1$:

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

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.)

Recap

- Reduction to (nearly) orthogonally decomposable tensor permits simple and error-tolerant algorithms.

  Lots of on-going work on **non-orthogonal / over-complete tensor decompositions** (e.g., Goyal, Vempala, & Xiao, 2014; Ge & Ma, 2015; Barak, Kelner, & Steurer, 2015; Ma, Shi, & Steurer, 2016).

- Many similarities to matrix decompositions and algorithms, but **differences due to non-linearity are crucial**.

Summary

- Using method-of-moments with $O(1)$-order moments, can efficiently estimate parameters for many latent variable 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 $O(1)$-order moments, can efficiently estimate parameters for many latent variable 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.
  - Incorporate user feedback interactively.

Acknowledgements

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

Funding: NSF (DMR-1534910, IIS-1563785), Sloan Foundation

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`
- Moitra. **Algorithmic aspects of machine learning**. 2014.
  `http://people.csail.mit.edu/moitra/docs/bookex.pdf` (Chapter 3)

Thank you