Estimation of Latent Variable Models via Tensor Decompositions

Sham M. Kakade
Microsoft Research, New England
Two canonical examples

Latent variable models are handy...

Two canonical examples:

- Mixture of Gaussians
  - each point generated by (unknown) cluster
- Topic models
  - “bag of words” model for documents
  - documents have one (or more) topics

What is the statistical efficiency of the estimator we find?
practical heuristics: $k$-Means, EM, Gibbs sampling?
What are the limits of learning?

- **computational and statistically efficient estimation:**
  - **stat. lower bound:** exponential($k$), overlapping clusters. [Moitra & Valiant, 2010]
  - **comp. lower bound:**
    ML estimation is NP-hard (for LDA). [Arora, Ge Moitra, 2012]

Are there computationally and statistically estimation methods?
- Under what assumptions and models?
- How general?
This talk: Efficient, closed form estimation procedures for (spherical) mixture of Gaussians and topic models.

- simple (linear algebra) approach
  - for a non-convex problem
- extensions to richer settings:
  - latent Dirichlet allocation, HMMs...

Are there fundamental limitations for learning general mixture models? NEW: in high dimensions, they are efficiently learnable.
Related Work

- **Mixture of Gaussians:**
  - with “separation” assumptions:
  - with no “separation” assumptions:

- **Topic models:**
  - with separation conditions:
    - Papadimitriou, Raghavan, Tamaki & Vempala (2000),
  - algebraic methods for phylogeny trees:
  - with multiple topics + “separation conditions”:
    - Arora, Ge & Moitra (2012)...

S. M. Kakade (MSR)
Mixture Models

(spherical) Mixture of Gaussian:

- $k$ means: $\mu_1, \ldots, \mu_k$
- sample cluster $i$ with prob. $w_i$
- observe $x$, with spherical noise,
  \[ x = \mu_i + \eta, \quad \eta \sim \mathcal{N}(0, \sigma_i^2 I) \]

(single) Topic Models

- $k$ topics: $\mu_1, \ldots, \mu_k$
- sample topic $i$ with prob. $w_i$
- observe $m$ (exchangeable) words
  \[ x_1, x_2, \ldots, x_m \text{ sampled i.i.d. from } \mu_i \]

- dataset: multiple points / $m$-word documents
- how to learn the params? $\mu_1, \ldots, \mu_k$, $w_1, \ldots, w_k$ (and $\sigma_i$'s)
The Method of Moments

- (Pearson, 1894): find params consistent with observed moments
- MOGs moments:
  \[ \mathbb{E}[x], \mathbb{E}[xx^\top], \mathbb{E}[x \otimes x \otimes x], \ldots \]
- Topic model moments:
  \[ \text{Pr}[x_1], \text{Pr}[x_1, x_2], \text{Pr}[x_1, x_2, x_3], \ldots \]
- **Identifiability:** with exact moments, what order moment suffices?
  - how many words per document suffice?
  - efficient algorithms?
vector notation and multinomials!

- \( k \) clusters, \( d \) dimensions/words, \( d \geq k \)
- for MOGs:
  - the conditional expectations are:
    \[
    \mathbb{E}[x|\text{cluster } i] = \mu_i
    \]
- topic models:
  - binary word encoding: \( x_1 = [0, 1, 0, \ldots]^\top \)
  - the \( \mu_i \)'s are probability vectors
  - for each word, the conditional probabilities are:
    \[
    \Pr[x_1|\text{topic } i] = \mathbb{E}[x_1|\text{topic } i] = \mu_i
    \]
With the first moment?

MOGs:

- have:
  
  \[ \mathbb{E}[x] = \sum_{i=1}^{k} w_i \mu_i \]

Single Topics:

- with 1 word per document:
  
  \[ \Pr[x_1] = \sum_{i=1}^{k} w_i \mu_i \]

Not identifiable: only \( d \) nums.
With the second moment?

MOGs:  
- additive noise

\[
\mathbb{E}[x \otimes x] = \mathbb{E}[(\mu_i + \eta) \otimes (\mu_i + \eta)] = \sum_{i=1}^{k} \mu_i \otimes \mu_i + \sigma^2 I
\]

- have a full rank matrix

Single Topics:  
- by exchangeability:

\[
\Pr[x_1, x_2] = \mathbb{E}[\mathbb{E}[x_1|\text{topic}] \otimes \mathbb{E}[x_2|\text{topic}]] = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i
\]

- have a low rank matrix!

Still not identifiable!
With three words per document?

- for topics: \(d \times d\) matrix, a \(d \times d \times d\) tensor:

\[
M_2 := \text{Pr}[x_1, x_2] = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i
\]

\[
M_3 := \text{Pr}[x_1, x_2, x_3] = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i
\]

- Whiten: project to \(k\) dimensions; make the \(\tilde{\mu}_i\)'s orthogonal

\[
\tilde{M}_2 = I
\]

\[
\tilde{M}_3 = \sum_{i=1}^{k} \tilde{w}_i \tilde{\mu}_i \otimes \tilde{\mu}_i \otimes \tilde{\mu}_i
\]
as bilinear and trilinear operators:

\[ a^\top M_2 b = M_2(a, b) = \sum_{i,j} [M_2]_{i,j} a_i b_j \]

\[ M_3(a, b, c) = \sum_{i,j,k} [M_3]_{i,j,k} a_i b_j c_k \]

matrix eigenvectors:

\[ M_2(\cdot, v) = \lambda v \]

define tensor eigenvectors:

\[ M_3(\cdot, v, v) = \lambda v \]
Recall, whitening makes $\tilde{\mu}_1, \tilde{\mu}_2, \ldots, \tilde{\mu}_k$ orthogonal.

What are the eigenvectors of $\tilde{M}_3$?

$$\tilde{M}_3(\cdot, v, v) = \sum_i \tilde{w}_i (v \cdot \tilde{\mu}_i)^2 \tilde{\mu}_i = \lambda v$$
Estimation

- find \( \nu \) so that:

\[
\tilde{M}_3(\cdot, \nu, \nu) = \sum_i \tilde{w}_i (\nu \cdot \tilde{\mu}_i)^2 \tilde{\mu}_i = \lambda \nu
\]

**Theorem**

*Assume the \( \mu_i \)'s are linearly independent.*

*The (robust) tensor eigenvectors of \( \tilde{M}_3 \) are the (projected) topics, up to permutation and scale.*

- this decomposition is easy; NP-Hard in general
- minor issues: un-projecting, un-scaling, no multiplicity issues
Algorithm: Tensor Power Iteration

- "plug-in" estimation: $\hat{M}_2, \hat{M}_3$
- power iteration:
  \[ v \leftarrow \hat{M}_3(\cdot, v, v) \]
  then deflate
- alternative: find local "skewness" maximizers:
  \[ \operatorname{argmax}_{\|v\|=1} \hat{M}_3(v, v, v) \]

Theorem

1. **computational efficiency**: in poly time, obtain estimates $\hat{\mu}_i$'s.
2. **statistical efficiency**: relevant parameters (e.g. min. singular value of $\mu_i$'s)
   \[ \|\hat{\mu}_i - \mu_i\| \leq \frac{\text{poly(relevant params)}}{\sqrt{\text{sample size}}} \]

- related algo's from independent component analysis
Mixtures of spherical Gaussians

Theorem

The variance \( \sigma^2 \) is the smallest eigenvalue of the observed covariance matrix \( \mathbb{E}[x \otimes x] - \mathbb{E}[x] \otimes \mathbb{E}[x] \). Furthermore, if

\[
M_2 := \mathbb{E}[x \otimes x] - \sigma^2 I
\]
\[
M_3 := \mathbb{E}[x \otimes x \otimes x]
\]

\[- \sigma^2 \sum_{i=1}^{d} (\mathbb{E}[x] \otimes e_i \otimes e_i + e_i \otimes \mathbb{E}[x] \otimes e_i + e_i \otimes e_i \otimes \mathbb{E}[x]),\]

then

\[
M_2 = \sum w_i \mu_i \otimes \mu_i
\]
\[
M_3 = \sum w_i \mu_i \otimes \mu_i \otimes \mu_i.\]

Differing \( \sigma_i \) case now solved.

MV '11 lower bound has \( k \) means on a line.
Latent Dirichlet Allocation

prior for topic mixture $\pi$:

$$p_\alpha(\pi) = \frac{1}{Z} \prod_{i=1}^{k} \pi_{\alpha_{i}}^{\alpha_{i}-1}, \quad \alpha_{0} := \alpha_{1} + \alpha_{2} + \cdots + \alpha_{k}$$

**Theorem**

Again, *three words per doc suffice*. Define

$$M_2 := \mathbb{E}[x_1 \otimes x_2] - \frac{\alpha_0}{\alpha_0 + 1} \mathbb{E}[x_1] \otimes \mathbb{E}[x_1]$$

$$M_3 := \mathbb{E}[x_1 \otimes x_2 \otimes x_3] - \frac{\alpha_0}{\alpha_0 + 2} \mathbb{E}[x_1 \otimes x_2 \otimes \mathbb{E}[x_1]] - more \ stuff...$$

Then

$$M_2 = \sum \tilde{w}_i \mu_i \otimes \mu_i$$

$$M_3 = \sum \tilde{w}_i \mu_i \otimes \mu_i \otimes \mu_i.$$
Richer probabilistic models

Hidden Markov models
- 3 length chains suffice

Probabilistic Context Free Grammars
- not-identifiable in general
- learning (under restrictions)

(latent) Bayesian networks
- give identifiability conditions
- new techniques/algos

Figure 3: Two parse trees (derivations) for the sentence "the man saw the dog with the telescope, under the C F G in figure 1."
Tensor decompositions provide simple/efficient learning algorithms.

see website for papers

Collaborators:

A. Anandkumar  D. Foster  R. Ge  Q. Huang  D. Hsu

P. Liang  Y. Liu  A. Javanmard  M. Telgarsky  T. Zhang