Estimation of Latent Variable Models via Tensor Decompositions

Sham M. Kakade

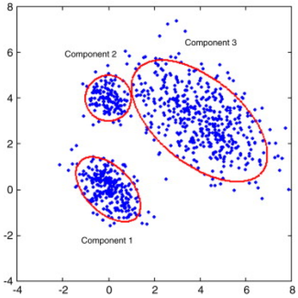
Microsoft Research, New England

Two canonical examples

Latent variable models are handy...

Two canonical examples:

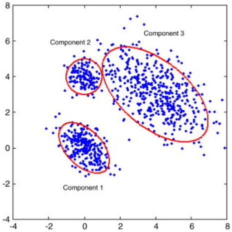
- Mixture of Gaussians
  - each point generated by (unknown) cluster
- Topic models
  - “bag of words” model for documents
  - documents have one (or more) topics



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?

A Different Approach

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:
    Dasgupta (1999), Arora & Kannan (2001), Vempala & Wang (2002) Achlioptas & McSherry (2005), Brubaker & Vempala (2008), Belkin & Sinha (2010), Dasgupta & Schulman (2007), ...
  - with no “separation” assumptions:
    Belkin & Sinha (2010), Kalai, Moitra, & Valiant (2010), Moitra & Valiant (2010), Feldman, Servedio, and O’Donnell (2006), Lindsay & Basak (1993)
- Topic models:
  - with separation conditions:
    Papadimitriou, Raghavan, Tamaki & Vempala (2000),
  - algebraic methods for phylogeny trees:
    J. T. Chang (1996), E. Mossel & S. Roch (2006),
  - with multiple topics + “separation conditions”:
    Arora, Ge &Moitra (2012)...

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 σ_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:

$$\Pr[x_1], \Pr[x_1, x_2],\ \Pr[x_1, x_2, x_3], \ldots$$

- 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 μ_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

$$
\begin{aligned}
& \mathbb{E}[x \otimes x] \\
= \; & \mathbb{E}[(\mu_i + \eta) \otimes (\mu_i + \eta)] \\
= \; & \sum_{i=1}^{k} w_i \, \mu_i \otimes \mu_i + \sigma^2 I
\end{aligned}
$$

- have a full rank matrix

Single Topics:

- by exchangeability:

$$
\begin{aligned}
& \Pr[x_1, x_2] \\
= \; & \mathbb{E}[\, \mathbb{E}[x_1 | topic] \otimes \mathbb{E}[x_2 | topic] \,] \\
= \; & \sum_{i=1}^{k} w_i \, \mu_i \otimes \mu_i
\end{aligned}
$$

- 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:

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

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

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

$$\tilde{M}_2 = I$$

$$\tilde{M}_3 = \sum_{i=1}^{k} \tilde{w}_i \, \tilde{\mu}_i \otimes \tilde{\mu}_i \otimes \tilde{\mu}_i$$

Tensors and Linear Algebra

- as bilinear and trilinear operators:

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

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

- matrix eigenvectors:

$$M_2(\cdot, v) = \lambda v$$

- define tensor eigenvectors:

$$M_3(\cdot, v, v) = \lambda v$$

Tensor eigenvectors

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 v so that:

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

Theorem

Assume the μ_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:

$$\mathrm{argmax}_{\|v\|=1} \hat{M}_3(v, v, v)$$

Theorem

1. *computational efficiency: in poly time, obtain estimates $\hat{\mu}_i$'s.*
2. *statistical efficiency: relevant parameters (e.g. min. singular value of μ_i's)*

$$\|\hat{\mu}_i - \mu_i\| \leq \frac{\textit{poly(relevant params)}}{\sqrt{\textit{sample size}}}$$

- related algo's from independent component analysis

Mixtures of spherical Gaussians

Theorem

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

$$\begin{aligned} M_2 &:= \mathbb{E}[x \otimes x] - \sigma^2 I \\ M_3 &:= \mathbb{E}[x \otimes x \otimes x] \\ &\qquad - \sigma^2 \sum_{i=1}^{d} \left(\mathbb{E}[x] \otimes e_i \otimes e_i + e_i \otimes \mathbb{E}[x] \otimes e_i + e_i \otimes e_i \otimes \mathbb{E}[x] \right), \end{aligned}$$

then

$$\begin{aligned} M_2 &= \sum w_i \, \mu_i \otimes \mu_i \\ M_3 &= \sum w_i \, \mu_i \otimes \mu_i \otimes \mu_i . \end{aligned}$$

Differing σ_i case now solved.

MV '11 lower bound has k means on a line.

Latent Dirichlet Allocation

prior for topic mixture π:

$$p_\alpha(\pi) = \frac{1}{Z} \prod_{i=1}^{k} \pi_i^{\alpha_i - 1}, \quad \alpha_0 := \alpha_1 + \alpha_2 + \cdots + \alpha_k$$

Theorem

Again, three words per doc suffice. Define

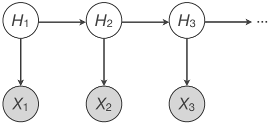
$$\begin{aligned} M_2 &:= \mathbb{E}[x_1 \otimes x_2] - \frac{\alpha_0}{\alpha_0 + 1} \mathbb{E}[x_1] \otimes \mathbb{E}[x_1] \\ M_3 &:= \mathbb{E}[x_1 \otimes x_2 \otimes x_3] - \frac{\alpha_0}{\alpha_0 + 2} \mathbb{E}[x_1 \otimes x_2 \otimes \mathbb{E}[x_1]] - \textit{more stuff...} \end{aligned}$$

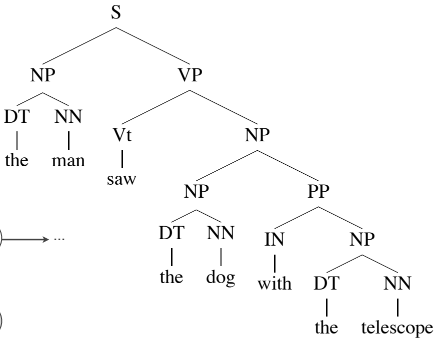
Then

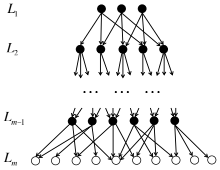
$$\begin{aligned} M_2 &= \sum \tilde{w}_i \, \mu_i \otimes \mu_i \\ M_3 &= \sum \tilde{w}_i \, \mu_i \otimes \mu_i \otimes \mu_i. \end{aligned}$$

Learning without inference!

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

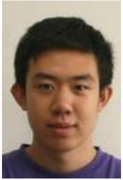
Thanks!

- Tensor decompositions provide simple/efficient learning algorithms.
  - see website for papers

Collaborators:


A. Anandkumar

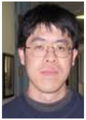

D. Foster


R. Ge


Q. Huang


D. Hsu


P. Liang


Y. Liu


A. Javanmard


M. Telgarsky


T. Zhang