Tensor Decompositions: Exploiting Structure in Observed Correlations

Anima Anandkumar, Daniel Hsu, Sham M. Kakade

Microsoft Research, New England

Learning Hidden Structure

- With unlabeled data, how do we discover hidden structure?
  - topics in documents?
  - clusters? hidden communities in social networks?
  - hidden interactions?

Learning is easy with cluster labels. Learning without cluster labels?

Using Observed Correlations

Two step approach:

1. Under *modeling assumptions*, what correlations arise?
   topic models, HMMs, LDA, mixture of Gaussians models, parsing (e.g. PCFGs), Bayesian networks
2. Can we "invert"/reverse engineer the model from these correlations?

This Tutorial

How to utilize observed correlations?

- part 1: the correlational structure
  - When are the correlations sufficient for learning?
- part 2: “invert” (CP decomposition)
  - generalizations of simple (linear algebra) approach
  - aren’t these problems hard/non-convex?
- part 3: “invert... differently” (Tucker)
  - exploit different structural conditions

Two Extremes

- Single hidden state active
  - mixture of Gaussians, single topic per document
- Independent Component Analysis
  - Blind source separation
    audio signal has different speakers talking
  - independent factors

What about the middle ground?

Mixture Models

(spherical) Mixture of Gaussian:

- k means: $\mu_1, \ldots \mu_k$
- sample cluster $H = i$ 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 $H = i$ 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)

vector notation!

- 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:

$$\Pr[x_1|\text{topic i}] = \mathbb{E}[x_1|\text{topic i}] = \mu_i$$

ICA

- k mixing directions: $\mu_1, \ldots \mu_k$
- each hidden (scalar) factor, $H_1, H_2, \ldots H_k$, is independently distributed
- observe mixture x, with Gaussian noise,

$$x = \sum_i \mu_i H_i + \eta, \quad \eta \sim \mathcal{N}(0, \sigma^2)$$

- in MOG's, only one $H_i = 1$
- how to learn the params? $\mu_1, \ldots \mu_k$

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?

(some) Related Work

- Kruskal's Theorem
  Kruskal (1977), Bhaskara, Charikar, & Vijayaraghavan (2013), ...
- Algebraic Work
  - ICA literature: Cardoso&Common, '96, ...
  - for phylogeny trees: J. T. Chang (1996), E. Mossel & S. Roch (2006),
- Tensor Decomposition Algorithms
  Lathauwer, Moor, & Vandewalle (2000), Zhang & Golub (2001), Anandkumar et. al. (2012), ...
- Structural assumptions/Dictionary learning
  Spielman, Wang & Right (2012), Arora, Ge, & Moitra (2012)

With the first moment?

MOGs:

- have:

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

Single Topics:

- with 1 word per document:

$$\Pr[x_1] = \sum_{i=1}^{k} w_i \mu_i$$

ICA:

- define $\mathbb{E}[H_i] := w_i$

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

Not identifiable: only d nums.

With the second moment?

MOGs/ICA:

- 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:

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

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

Whitening

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

$$\begin{aligned} \tilde{M}_2 &= I \\ \tilde{M}_3 &= \sum_{i=1}^{k} \tilde{w}_i \, \tilde{\mu}_i \otimes \tilde{\mu}_i \otimes \tilde{\mu}_i \end{aligned}$$

  (for a $k \times k \times k$ tensor)
- Is there a unique solution? parameter counting?
  - yes: $k < d$ +generic params (Kruskal (1977))
  - what about $k > d$? (Lathauwer, Castaing, & Cardoso (2007))
- How is this different form an SVD?
- Can we solve this efficiently?

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

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

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 also solved.

Independent Component Analysis

Theorem

Different higher order moments from MOGs. Use cumulants:

$$\begin{aligned} M_4 \;&:=\; \mathbb{E}[x \otimes x \otimes x \otimes x] \\ &\qquad - \left(\mathbb{E}[x \otimes x] \otimes \mathbb{E}[x \otimes x] \;+\; \textit{more stuff...}\right), \end{aligned}$$

then

$$M_4 \;=\; \sum w_i\, \mu_i \otimes \mu_i \otimes \mu_i \otimes \mu_i .$$

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

- approaches richer probabilistic models:
- setting 1: have a "diagonalization" problem (like the SVD)
  topic models/LDA, HMMs, mixture of Gaussians models, parsing (e.g. PCFGs),
  - rely on correlational structure/prior of hidden variables
- setting 2: have a "sparse" problem:
  Bayesian networks, Dictionary learning, topic modeling
  - suppose only a few "topics" are on. no other prior assumptions.
  - rely on sparsity + incoherence

Thanks!

- The structure of the correlations gives rise to certain decomposition problems.
- Identifiability: This is the first step.
- Stay Tuned:
  How do we estimate efficiently?