Spectral Methods: Latent Variable Models

FOROUGH ARABSHAHI

PGM
Spring 2018
Each document has exactly one single topic

Exchangeability: Joint distribution invariant under permutation

Words are conditionally independent given the hidden state
- Each document has $\ell \geq 3$ words

Generative Process:
- For each document draw a topic based on a discrete distribution $w := (w_1, w_2, \ldots, w_k) \in \Delta^{k-1}$
- Given the topic, draw words independently
Topic Model: Mixture of Unigrams

The topic is modeled with a k-dimensional hidden variable:

$$\Pr[h = j] = w_j, \quad j \in [k].$$

Let the number of words in the dictionary be $d$ \(\Rightarrow\) words can be shown with d-dim vectors

$$x_1, x_2, \ldots, x_\ell \in \mathbb{R}^d$$

Let $e_i$ indicate the standard coordinate basis vector in $\mathbb{R}^d$:

$$x_i = e_i \text{ iff the } t^{th} \text{ word in the document is } i \text{ for } t \in [\ell]$$

Training: For each topic probability what is the word distribution? [topic-word matrix]

- Maximum Likelihood: EM
- Spectral Methods
Some Tensor Notations First

P-way tensor:

\[ [A_{i_1, i_2, \ldots, i_p} : i_1, i_2, \ldots, i_p \in [n]] \]

\( P = 1 \): Vector

\( P = 2 \): Matrix

Alternate notation:

\( P \)-th tensor power:

\( P \)-th tensor of rank \( k \):

\( P \)-th symmetric tensor:

Vector product:

\[ u \otimes v = uv^T \text{ for } u, v \in \mathbb{R}^{n \times 1} \]

\( A \in \bigotimes^p \mathbb{R}^n \)

\( v \otimes^p := v \otimes v \otimes \cdots \otimes v \in \bigotimes^p \mathbb{R}^n \)

\[ T = v_1 \otimes^3 + v_2 \otimes^3 + \cdots, \]

\[ A = \sum_{j=1}^k u_{1,j} \otimes u_{2,j} \otimes \cdots \otimes u_{p,j} \]

\[ A = \sum_{j=1}^k u_j \otimes^p \]

\( u \otimes v = uv^T \) for \( u, v \in \mathbb{R}^{n \times 1} \)
Mixture of Unigrams: properties

Observe: Cross moments of the model corresponds to joint probability table (recall: $x_i = e_i$)

$$
\mathbb{E}[x_1 \otimes x_2] = \sum_{1 \leq i, j \leq d} \Pr[x_1 = e_i, x_2 = e_j] e_i \otimes e_j
$$

$$
= \sum_{1 \leq i, j \leq d} \Pr[\text{1st word} = i, \text{2nd word} = j] e_i \otimes e_j,
$$

Conditional mean:

$$
\mathbb{E}[x_t|h = j] = \sum_{i=1}^{d} \Pr[t\text{-th word} = i|h = j] e_i = \sum_{i=1}^{d} [\mu_j]_i e_i = \mu_j, \quad j \in [k]
$$

Conditional independence:

$$
\mathbb{E}[x_1 \otimes x_2|h = j] = \mathbb{E}[x_1|h = j] \otimes \mathbb{E}[x_2|h = j] = \mu_j \otimes \mu_j, \quad j \in [k]
$$
Spectral Methods: Mixture of Unigrams

Theorem: [Anandkumar et al., 2012c] (cont.)

\[ M_2 := \mathbb{E}[x_1 \otimes x_2] \]

Proof:

\[ M_2 = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \]

\[ \mathbb{E}[x_1 \otimes x_2] = \mathbb{E}[\mathbb{E}[x_1 \otimes x_2|h]] \]
\[ = \mathbb{E}[\mathbb{E}[x_1|h] \otimes \mathbb{E}[x_2|h]] \]
\[ = \sum_{i=1}^{k} \text{Pr}[h = i] \mathbb{E}[x_1|h = i] \otimes \mathbb{E}[x_2|h = i] \]
\[ = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \]
Spectral Methods: Mixture of Unigrams

Training: Estimate $M_2$ from the data and decompose into rank-1 components

Non-convex optimization, but guaranteed global solution!

Fast and scalable

Extensive linear algebra support

But ...
Matrix Decomposition

Matrix decomposition is not unique in general :(  

Matrix decomposition is only unique for **orthogonal** factors  

Guaranteed recovery needs the **separability** assumption: \( \lambda_1 > \lambda_2 > \cdots \)  

Recovery is not guaranteed for **linearly independent** factors  

Not applicable to the **overcomplete** scenario  

This means: If we decompose \( M_2 \), we can only recover the subspace that the factors lie in, and not the factors, unless they are orthogonal
Higher order moments?

Theorem: [Anandkumar et al., 2012c]

Proof: similar to $M_2$

Observe:

$$\mathbb{E}[x_1 \otimes x_2 \otimes x_3] = \sum_{1 \leq i, j, l \leq k} \mathbb{P}[x_1 = e_i, x_2 = e_j, x_3 = e_l] e_i \otimes e_j \otimes e_l$$

$$\mathbb{E}[x_1 \otimes x_2 \otimes x_3] = \mathbb{E}[\mathbb{E}[x_1 \otimes x_2 \otimes x_3|h]]$$

$$= \mathbb{E}[\mathbb{E}[x_1|h] \otimes \mathbb{E}[x_2|h] \otimes \mathbb{E}[x_3|h]]$$

$$= \sum_{i=1}^{k} \mathbb{P}[h = i] \mathbb{E}[x_1|h = i] \otimes \mathbb{E}[x_2|h = i] \otimes \mathbb{E}[x_3|h = i]$$

$$= \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i$$
Spectral Methods: Tensor Decomposition

Training: Estimate $M_3$ from the data and decompose into rank-1 components

$T = v_1 \otimes^3 + v_2 \otimes^3 + \cdots$,

The decomposition is unique for \textit{linearly independent} factors.
Training Algorithm

Condition [non-degeneracy]: vectors \( \mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^d \) are linearly independent and scalars \( w_1, w_2, \ldots, w_k > 0 \) are strictly positive. \( \Rightarrow \) \( M_2 \) is positive semi-definite and has rank \( k \).

\[
M_2(W, W) = W^T M_2 W = I
\]

\[
M_2(W, W) = \sum_{i=1}^{k} W^T (\sqrt{w_i} \mu_i) (\sqrt{w_i} \mu_i)^T W = \sum_{i=1}^{k} \tilde{\mu}_i \tilde{\mu}_i^T = I
\]

\[
\tilde{\mu}_i := \sqrt{w_i} W^T \mu_i
\]

\[
\tilde{M}_3 = \sum_{i=1}^{k} w_i (W^T \mu_i)^\otimes 3 = \sum_{i=1}^{k} \frac{1}{\sqrt{w_i}} \tilde{\mu}_i^\otimes 3
\]
Mixture of Unigrams

Raw cross moments of the observations directly yield a symmetric tensor structure

Assumes that each document has a single topic: limiting assumption [Blei et al., 2003]

More realistic topic models?
Latent Dirichlet Allocation: LDA
Latent Dirichlet Allocation: LDA
LDA: an Admixture Model

For each document, draw the topic proportions, given these proportions draw a topic and then draw the word

\[ p_\alpha(h) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} h_i^{\alpha_i - 1}, \quad h \in \Delta^{k-1} \]

\[ \alpha_0 := \alpha_1 + \alpha_2 + \cdots + \alpha_k. \]
Spectral Methods: LDA (Cont.)

Theorem: [Anandkumar et al., 2012a] (Cont.)

Raw moments are not directly diagonalizable.

Trick: Find the right off diagonals that compensate for it
Spectral Methods: LDA (Cont.)

Proof:

\[ M_1 = \mathbb{E}[\lambda_i] = \mathbb{E} \left[ \mathbb{E}[x_1 | h] \right] = \mathbb{E}[h] = 0 \]

\[ \mathbb{E}[x_1 \otimes x_2] = \mathbb{E}[\mathbb{E}[x_1 \otimes x_2 | h]] = \mathbb{E}[h \otimes h] O^T \]

\[ \mathbb{E}[h \otimes h] = \frac{1}{(\alpha_0 + 1)\alpha_0} \left( \text{diag}(\alpha) + \alpha \alpha^T \right) \]

\[ \mathbb{E}[h] \otimes \mathbb{E}[h] = \frac{1}{\alpha_0^2} \alpha \otimes \alpha \]

\[ M_2 = O \mathbb{E}[h \otimes h] O^T - \frac{\alpha_0}{\alpha_0 + 1} \mathbb{E}[h] \otimes \mathbb{E}[h] O^T \]

\[ = O \left[ \frac{1}{(\alpha_0 + 1)\alpha_0} \left( \text{diag}(\alpha) + \alpha \otimes \alpha \right) - \frac{1}{(\alpha_0 + 1)\alpha_0} \alpha \otimes \alpha \right] O^T \]

\[ = \sum \frac{\alpha_i}{(\alpha_0 + 1)\alpha_0} O_i \otimes O_i \]

\[ \mathbb{E}[h_i] = \frac{\alpha_i}{\alpha_0} \]

\[ \mathbb{E}[h_i^2] = \frac{(\alpha_i + 1)\alpha_i}{(\alpha_0 + 1)\alpha_0} \]

\[ \mathbb{E}[h_i h_j] = \frac{\alpha_i \alpha_j}{(\alpha_0 + 1)\alpha_0} \]

\[ \mathbb{E}[x_1 \otimes x_2] = \mathbb{E}_h [\mathbb{E}[x_1 \otimes x_2 | h]] \]

\[ M_2 = \mathbb{E}[M_1 \otimes M_2] - \frac{\alpha_0}{\alpha_0 + 1} \]

\[ M_2 = \sum M_i \otimes M_i \]
Spectral Methods: LDA (Cont.)

Theorem: [Anandkumar et al., 2012a]

\[ M_3 := E[x_1 \otimes x_2 \otimes x_3] \]
\[ - \frac{\alpha_0}{\alpha_0 + 2} \left( E[x_1 \otimes x_2 \otimes M_1] + E[x_1 \otimes M_1 \otimes x_2] + E[M_1 \otimes x_1 \otimes x_2] \right) \]
\[ + \frac{2\alpha_0^2}{(\alpha_0 + 2)(\alpha_0 + 1)} M_1 \otimes M_1 \otimes M_1. \]

\[ M_3 = \sum_{i=1}^{k} \frac{2\alpha_i}{(\alpha_0 + 2)(\alpha_0 + 1)\alpha_0} \mu_i \otimes \mu_i \otimes \mu_i. \]

Raw moments are not directly diagonalizable.

Trick: Find the right off diagonals that compensate for it
Spectral Methods: Training LDA

Goal: recover the topic word matrix:

Estimate word triplets from the data and decompose:
Practical Notes (Cont.)

Estimating $M_3$ : Although only 3 words are needed per document, one must use all word triplets.

Should we average over all $\binom{\ell}{3} \cdot 3!$ Ordered triplets in each document? Computationally expensive.

As shown by [Zou et al., 2013], this averaging can be done implicitly, in an efficient manner.

\[ \mathbb{E}[\mathbf{n}_1 \mathbf{n}_2 \mathbf{n}_3] \]
Let $c \in \mathbb{R}^d$ be the word count vector for the document, then its contribution to the moment is:

$$\frac{1}{(\ell^3)} \cdot \frac{1}{3!} \cdot \left( c \otimes c \otimes c + 2 \sum_{i=1}^{d} c_i (e_i \otimes e_i \otimes e_i) \right)$$

$$- \sum_{i=1}^{d} \sum_{j=1}^{d} c_i c_j (e_i \otimes e_i \otimes e_j) - \sum_{i=1}^{d} \sum_{j=1}^{d} c_i c_j (e_i \otimes e_j \otimes e_i) - \sum_{i=1}^{d} \sum_{j=1}^{d} c_i c_j (e_i \otimes e_j \otimes e_j)$$

Which is equal to:

$$\frac{1}{(\ell^3)} \cdot \frac{1}{3!} \cdot \sum_{\text{ordered word triple } (x, y, z)} e_x \otimes e_y \otimes e_z$$
Practical Notes

The same can be observed for the second moment.

\[
\frac{1}{|2|} \cdot \frac{1}{2!} \left( c \otimes c - \text{diag}(c) \right)
\]

Why?

\[
\mathbb{E} \left[ c_n(i)^2 - c_n(i) \right] = \mathbb{E} \left[ \left( \sum_{p=1}^{\ell_n} x_{n,p}(i) \right)^2 - \sum_{p=1}^{\ell_n} x_{n,p}(i) \right]
\]

\[
= \mathbb{E} \left[ \sum_{p=1}^{\ell_n} x_{n,p}(i)^2 + 2 \sum_{p<q} x_{n,p}(i)x_{n,q}(i) - \sum_{p=1}^{\ell_n} x_{n,p}(i) \right]
\]

\[
= 2 \sum_{p<q} \mathbb{E} \left[ x_{n,p}(i)x_{n,q}(i) \right] \quad \text{(since } x_{n,p}(i)^2 = x_{n,p}(i)\text{)}
\]

\[
= \ell_n(\ell_n - 1)[M_2]_{i,i}
\]
Some practical results

Learning topics from the PubMed dataset with 8M documents

Perplexity = $\exp[-\text{likelihood}]$
Generalizability?

Applicable to:
- Spherical Gaussians
- Independent Component Analysis (ICA)
- Multi-view Models (e.g. Hidden Markov Models)
- Correlated Topic Models
- Hierarchical Topic Models

In general should be re-derived for a new model.

Not known if it can be derived for arbitrary models.
Latent Variable Models

Training:
- Maximum Likelihood: EM
- Spectral Methods: Replace objective with the best tensor decomposition
- Preserves global optima
EM vs. Spectral Methods

**EM**
- Aims to find MLE, so more “statistically efficient”
- Can get stuck in local optima
- Lack of theoretical Guarantees
- Easily derived for new models

**Spectral**
- Does not aim to find MLE so less “statistically efficient”
- Local-optima free
- Provably consistent
- Very fast
- Challenging to derive for new models (not known whether it can generalize to arbitrary models)