An Introduction to Spectral Learning

Hanxiao Liu

November 8, 2013
Outline

1. Method of Moments
2. Learning topic models using spectral properties
3. Anchor words
\( X_1, \ldots, X_n \sim p(x; \theta), \ \theta = (\theta_1, \ldots \theta_m)^\top \)

\[ \hat{\theta} = \hat{\theta}_n = w(X_1, \ldots, X_n) \]

- **Maximum Likelihood Estimator (MLE)**

\[ \hat{\theta} = \arg \max_{\theta} \log \mathcal{L}(\theta) \]

- **Bayes Estimator (BE)**

\[ \hat{\theta} = \mathbb{E}(\theta | X) = \frac{\int \theta p(x|\theta) \pi(\theta) \, d\theta}{\int p(x|\theta) \pi(\theta) \, d\theta} \]
Preliminaries

**Question**

What makes a good estimator?

- MLE is consistent
- Both the MLE and BE have asymptotic normality

\[
\sqrt{n} \left( \hat{\theta}_n - \theta \right) \rightsquigarrow N \left( 0, \frac{1}{I(\theta)} \right)
\]

under mild (regularity) conditions

Can be computationally expensive
Preliminaries

Example (Gamma Distribution)

\[ p(x_i; \alpha, \theta) = \frac{1}{\Gamma(\alpha) \theta^\alpha} x_i^{\alpha-1} \exp \left(-\frac{x_i}{\theta}\right) \]

\[ \mathcal{L}(\alpha, \theta) = \left(\frac{1}{\Gamma(\alpha) \theta^\alpha}\right)^n \left(\prod_{i=1}^{n} x_i\right)^{\alpha-1} \exp \left(-\frac{\sum_{i=1}^{n} x_i}{\theta}\right) \]

MLE is hard to compute due to the existence of \( \Gamma(\alpha) \)
Method of Moments

\( j \)-th theoretical moment, \( j \in [k] \)

\[
\mu_j (\theta) := \mathbb{E}_\theta \left( X^j \right)
\]

\( j \)-th sample moment, \( j \in [k] \)

\[
M_j := \frac{1}{n} \sum_{i=1}^{n} X_i^j
\]

Plug-in and solve the multivariate polynomial equations

\[
M_j = \mu_j (\theta) \quad j \in [k]
\]

sometimes can be recast as spectral decomposition
Method of Moments

Example (Gamma Distribution)

\[ p(x_i; \alpha, \theta) = \frac{1}{\Gamma(\alpha) \theta^\alpha} x_i^{\alpha-1} \exp \left( -\frac{x_i}{\theta} \right) \]

\[ \bar{X} = \mathbb{E}(X_i) = \alpha \theta \]

\[ \frac{1}{n} \sum_{i=1}^{n} \left( X_i - \bar{X} \right)^2 = \text{Var}(X_i) = \alpha \theta^2 \]

\[ \Rightarrow \hat{\theta} = \frac{1}{n \bar{X}} \sum_{i=1}^{n} \left( X_i - \bar{X} \right)^2, \quad \hat{\alpha} = \frac{\bar{X}}{\hat{\theta}} = \frac{n \bar{X}^2}{\sum_{i=1}^{n} \left( X_i - \bar{X} \right)^2} \]
Method of Moments

- lack guarantee about the solution
- high-order sample moments are hard to estimate

To reach a specified accuracy, the required sample size and computational cost is exponential in $k$ (or $n$)!

**Question**
Could we recover the true $\theta$ from only low-order moments?

**Question**
Could we lower the sample requirement and computational complexity based on some (hopefully mild) assumptions?
Learning the Topic Models

- Papadimitriou et al. (2000)
  - Non-overlapping separation condition (strong)
- Anandkumar et al. (2012), MoM+SD
  - Full rank assumption (weak)
  - Multinomial Mixture, LDA
- Arora et al. (2012), MoM+NMF+LP
  - Anchor words (mild)
  - LDA, Correlated Topic Model
  - A more practical algorithm proposed in 2013
Learning the Topic Models

Suppose there are $n$ documents, $k$ hidden topics, $d$ features

$$M = [\mu_1 | \mu_2 | \ldots | \mu_k] \in R^{d \times k}, \ \mu_j \in \Delta^{d-1} \ \forall j \in [k]$$

$$w = (w_1, \ldots, w_k), \ w \in \Delta^{k-1}$$

$$P(h = j) = w_j, \ j \in [k]$$

For the $v$-th word in a document, $x_v \in \{e_1, \ldots e_d\}$

$$P(x_v = e_i | h = j) = \mu^i_j, \ j \in [k], \ i \in [d]$$

**Goal:** Recover the $M$ using low-order moments
Learning the Topic Models

Construct moment statistics

Pairs\(_{ij}\) := \(P(x_1 = e_i, x_2 = e_j)\)

\[
\text{Pair} = \mathbb{E}[x_1 \otimes x_2] \in \mathbb{R}^{d \times d}
\]

Triples\(_{ij}\) := \(P(x_1 = e_i, x_2 = e_j, x_3 = e_t)\)

\[
\text{Triples} = \mathbb{E}[x_1 \otimes x_2 \otimes x_3] \in \mathbb{R}^{d \times d \times d}
\]

- Empirical plug-ins i.e. \(\hat{\text{Pairs}}\) and \(\hat{\text{Triples}}\) could be obtained from data through a straightforward manner.
- We want to establish some equivalence between the empirical moments and parameters of interest.
Learning the Topic Models

\[
\text{Triples} (\eta) := \mathbb{E}[x_1 \otimes x_2 \otimes \langle x_3, \eta \rangle] \in \mathbb{R}^{d \times d}
\]

\[
\text{Triples} (\eta) : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}
\]

\begin{center}
Lemma
\end{center}

\[
\text{Pairs} = M \text{diag} (w) M^\top
\]

\[
\text{Triples} (\eta) = M \left( \text{diag} \left( M^\top \eta \right) \text{diag} (w) \right) M^\top
\]

The unknown \( M \) and \( w \) are twisted.
Learning the Topic Models

**Assumption (Non-degeneracy)**

$M$ has full column rank $k$

1. Find $U, V \in \mathbb{R}^{d \times k}$ s.t. $(U^\top M)^{-1}$ and $(V^\top M)^{-1}$ exist.
2. $\forall \eta \in \mathbb{R}^d$, define $B(\eta) \in \mathbb{R}^{k \times k}$

$$B(\eta) := (U^\top \text{Triples}(\eta) V)(U^\top \text{Pairs} V)^{-1}$$

**Lemma (Observable Operator)**

$$B(\eta) = (U^\top M) \text{diag}(M^\top \eta)(U^\top M)^{-1}$$
Learning the Topic Models

**Input:** Pairs and Triples

**Output:** topic-word distributions $\hat{M}$

$\hat{U}, \hat{V} \leftarrow$ top $k$ left, right eigenvectors of Pairs $^a$

$\eta \leftarrow$ random sample from range($\hat{U}$)

$(\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_k) \leftarrow$ right eigenvectors of $B (\eta) \ b$

**for** $j \leftarrow 1$ **to** $k$ **do**

$\hat{\mu}_j \leftarrow \hat{U} \hat{\xi}_j / \langle 1, \hat{U} \hat{\xi}_j \rangle$

**end**

**return** $\hat{M} = [\hat{\mu}_1 | \hat{\mu}_2 | \ldots | \hat{\mu}_k]$

---

$^a$ Pairs $= M \text{diag} (w) M^\top$

$^b$ $B (\eta) = (U^\top M) \text{diag} (M^\top \eta) (U^\top M)^{-1}$
Learning the Topic Models

Lemma (Observable Operator)

\[ B(\eta) = \left(U^\top M\right) \text{diag} \left(M^\top \eta\right) \left(U^\top M\right)^{-1} \]

We hope \( M^\top \eta \) has distinct entries. How to pick \( \eta \)?

\[ \eta \leftarrow e_i \Rightarrow M^\top \eta \quad \text{i-th word's distribution over topics} \]

Prior knowledge required!
Otherwise, \( \eta \leftarrow U\theta, \theta \sim \text{Uniform}(S^{k-1}) \)
Learning the Topic Models

- SVD is carried out on $\mathbb{R}^{k \times k}, k \ll d$
- Only involves trigram statistics i.e. low-order moments
- Guaranteed to recover the parameters
- Parameters of more complicated models like LDA can be recovered in the same manner
Tensor Decomposition

Recall

Pairs = \( M \text{diag} (w) M^\top \)

Triples (\( \eta \)) = \( M \left( \text{diag} \left( M^\top \eta \right) \text{diag} (w) \right) M^\top \)

Pairs = \( \sum_{j}^{k} w_j \cdot \mu_j \otimes \mu_j \)

Triples = \( \sum_{j}^{k} w_j \cdot \mu_j \otimes \mu_j \otimes \mu_j \)

Symmetric tensor decomposition? \( \mu_j \) need to be orthogonal
Tensor Decomposition

Whiten Pairs

\[ W := UD^{\frac{1}{2}} \Rightarrow W^\top \text{Pairs} W = I \]

\[ \mu'_j := \sqrt{w_j} W^\top \mu_j \]

We can check that \( \mu'_j, j \in [k] \) are orthonormal vectors.

Do orthogonal tensor decomposition on

\[ \text{Triples} (W, W, W) = \sum_{j=1}^{k} w_j \left( W^\top \mu_j \right)^\otimes 3 = \sum_{j=1}^{k} \frac{1}{\sqrt{w_j}} \mu'_j \otimes 3 \]

Then recover \( \mu_j \) from \( \mu'_j \).
An Introduction to Spectral Learning

Anchor Words

Drawbacks of previous algorithms

- topics cannot be correlated
- the bound is weak (comparatively speaking)
- empirical runtime performance is not satisfactory

Alternatively assumptions?
**Anchor Words**

**Definition (p-separable)**

\[ M \text{ is } p\text{-separable if } \forall j, \exists i \text{ s.t. } M_{ij} \geq p \text{ and } M_{ij'} = 0 \text{ for } j' \neq j \]

- Documents do not necessarily contain anchor words
- Two-fold algorithm
  1. Selection: find the anchor word for each topic
  2. Recover: recover \( M \) based on anchor words
- Good theoretical guarantees and empirical results
Anchor Words

1

The illustration is taken from Ankur Moitra’s slides, http://people.csail.mit.edu/moitra/docs/IASM.pdf
Discussion

Summary

- A brief introduction to MoM
- Learning topic models by spectral decomposition
- Anchor words assumption

Connections with our work?