A Method of Moments for Latent Variable Models

Matteo Ruffini   Marta Casanellas   Ricard Gavaldà

Universitat Politècnica de Catalunya

Motivation

Methods of Moments: a promising tool to learn **topic models**: *Single Topic Model* [1, 2], *LDA* [3]...

Existing approaches: strong theory, but poor applicability (poor sample complexity or long running times).

Our Proposal: provide improved MoMs able to compete with most used methods to learn topic models (e.g. Gibbs Sampling).

The Single Topic Model

Generative Process for each text:

- Draw a topic: $Y \in [k]$, with $\mathbb{P}[Y = j] = \omega_j$.
- Sample all the words of the text from a discrete distribution that depends only on the topic:

$$\mathbb{P}[\textit{sampling word } h | \textit{topic} = j] = (\mu_j)_h$$

Model parameters:

- $\mu_1, ..., \mu_k$, the *topics* (each $\mu_i \in \mathbb{R}^d$).
- $\omega = (\omega_1, ..., \omega_k)$, the *weights*.

Latent Dirichlet Allocation

Identical to the Single Topic Model, but each text deals with a multitude of topics.

Generative Process for each text:

- A vector of topic proportions h is sampled from a Dirichlet distribution $h \approx Dir(\omega)$.
- Each word has a unique latent topic, sampled from a discrete distribution with parameter h.

Model parameters:

- $\mu_1, ..., \mu_k$, the *topics*.
- $\omega = (\omega_1, ..., \omega_k)$, the *Dirichlet* parameter.

Method of Moments

Task: to estimate model parameters from data

- Find (model-dependent) estimators of the moments: $\hat{M}_1, \hat{M}_2, \hat{M}_3$

$$\mathbb{E}[\hat{M}_1] = M_1 = \sum_{i=1}^{k} \omega_i \mu_i \in \mathbb{R}^d \quad (1)$$

$$\mathbb{E}[\hat{M}_2] = M_2 = \sum_{i=1}^{k} \omega_i \mu_i \otimes \mu_i \in \mathbb{R}^{d\times d} \quad (2)$$

$$\mathbb{E}[\hat{M}_3] = M_3 = \sum_{i=1}^{k} \omega_i \mu_i \otimes \mu_i \otimes \mu_i \in \mathbb{R}^{d\times d\times d} \quad (3)$$

- Retrieve an estimate of model parameters with tensor decomposition:

$$\mathcal{TD}(\hat{M}_1, \hat{M}_2, \hat{M}_3) \rightarrow (\hat{\mu}_1, ..., \hat{\mu}_k, \hat{\omega})$$

This paper

Improve MoMs for topic models, providing.

- **Moment estimators** with improved sample complexity.

- **SVTD**: a new, fast and robust tensor decomposition algorithm.

Moment Estimators for the Single Topic Model

- $x^{(j)} \in \mathbb{R}^d$: its entry i counts how many times the word i appears in the document j
- c_j: number of words in document j.

Theorem 1: for any $h \le l \le m$

$$(\hat{M}_1)_h := \frac{\sum_{i=1}^{n}(x^{(i)})_h}{\sum_{i=1}^{x} c_i},$$

$$(\hat{M}_2)_{h,l} := \frac{\sum_{i=1}^{n}(x^{(i)})_h(x^{(i)} - \chi_{h=l})_l}{\sum_{i=1}^{n}(c_i - 1)c_i},$$

$$(\hat{M}_3)_{h,l,m} := \frac{\sum_{i=1}^{n}(x^{(i)})_h(x^{(i)} - \chi_{\neg h<l<m})_l(x^{(i)} - 2\chi_{h=l=m})_m}{\sum_{i=1}^{n}(c_i - 2)(c_i - 1)c_i}$$

satisfy Equations (1), (2) and (3).

Comparison with Other Methods:

- [1]: gets the moments averaging the outer product between few words of each document.
- [2]: extends [1] using all the available words, giving the same weight to all documents.
- Our approach: extends [2] giving more weight to longer documents.

Remarks:

- In the paper we provide sample complexity bounds for the proposed estimators.
- We prove theoretically and experimentally that we improve the sample complexity of [1, 2].
- We obtain estimators for LDA applying to theorem 1 an approach similar to that of [3].

Tensor Decomposition

Task: to find $(\mu_1, ..., \mu_k, \omega)$ satisfying Equations (1), (2) and (3), given M_1, M_2, M_3 and k.

Theorem 2: Let $M = [\mu_1|...|\mu_k] \in \mathbb{R}^{d\times k}$. Then SVTD provably returns the model parameters.

Singular **V**alue based **T**ensor **D**ecomposition

Require: M_1, M_2, M_3, k

1. k components SVD: $M_2 = U_k S_k U_k^\top$
2. Get the whitening matrix $E = U_k S_k^{1/2}$.
3. Select a feature r and let $M_{3,r}$ be the r-th slice of M_3
4. Define $H_r := E^\dagger M_{3,r} E^{\dagger\top}$
5. Let O be the Singular Vectors of H_r.
6. **for** $i = 1 \rightarrow d$ **do**
7. Compute $H_i := E^\dagger M_{3,i} E^{\dagger\top}$
8. Get the i-th row of M from the diagonal of $O^\top H_i O$.
9. **end for**
10. Obtain ω solving Eq. (1).
11. Return (M, ω)

Remarks

- The rows of M are the singular values of H_i, which are robust to perturbations.
- Time complexity: $O(d^2k + k^3 + d^3k)$ ($O(dk^2n)$ with optimized implementations).

Comparison with Other Methods

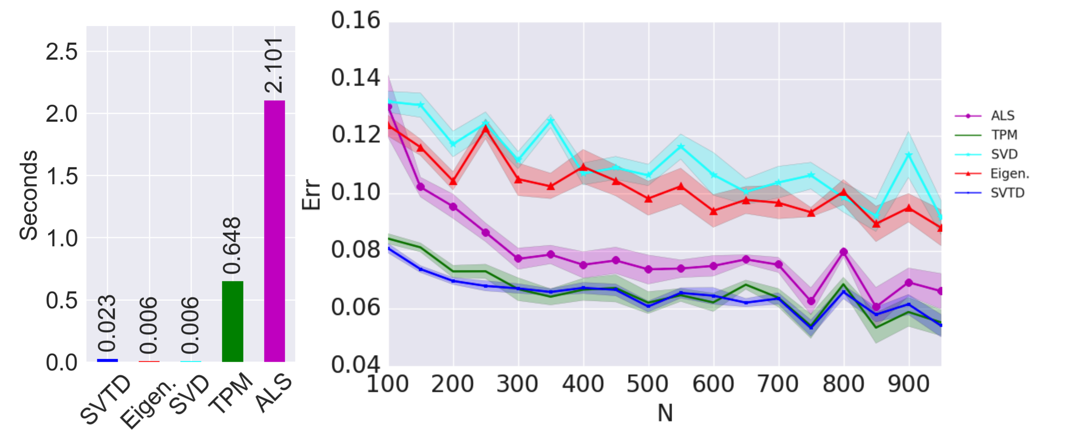
- *SVD method* [1]: similar to SVTD but based on singular vectors.
- *TPM* [4]; worse dependence on k (i.e. slower for high number of topics).
- *ALS* [5]: widely used heuristic; no guarantees on the decomposition, long running times.

Experiments

Synthetic Data

Generate synthetic data from a single topic model with parameters $(\mu_1, ..., \mu_k, \omega)$; estimate from data the model parameters with various methods. Compare run times and reconstruction error:

$$Err = \sum_{i=1}^{k} \|\mu_i - \hat{\mu}_i\|^2$$



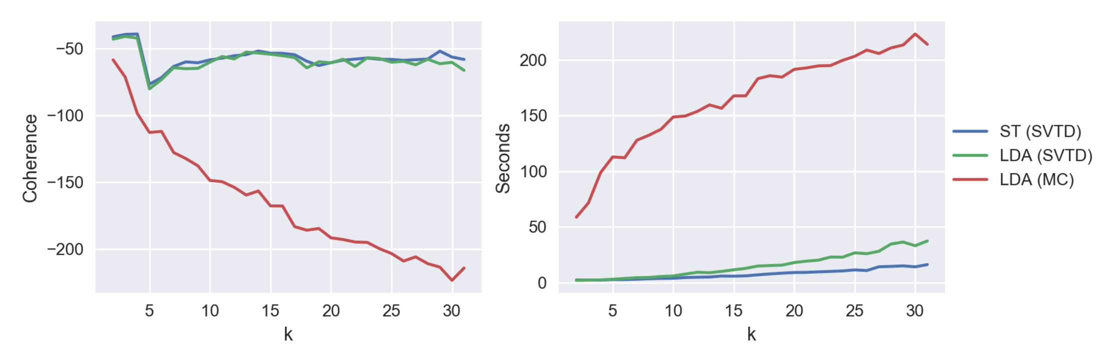
Real Data

Plot running time and *Topic Coherence* for Single Topic Model and LDA learned with **SVTD** and **Gibbs Sampling**.

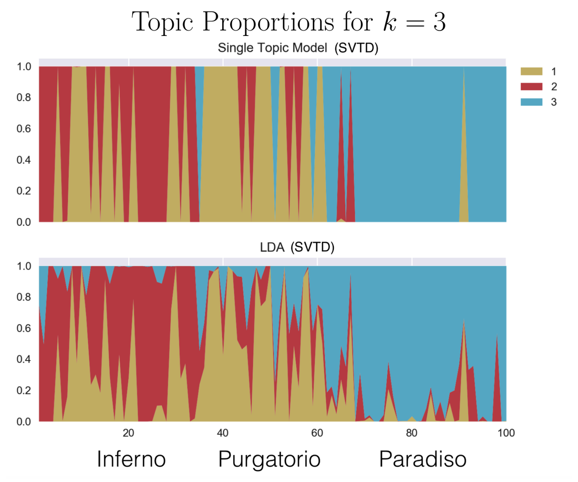
$$Coherence(\mu) = \sum_{j=2}^{L} \sum_{i=1}^{j-1} \log \frac{D(w_i, w_j) + 1}{D(w_i)}$$

Data: Dante's *Divina Commedia* and *State of The Union Addresses*.

Dante's Divina Commedia



Topic Proportions for $k = 3$

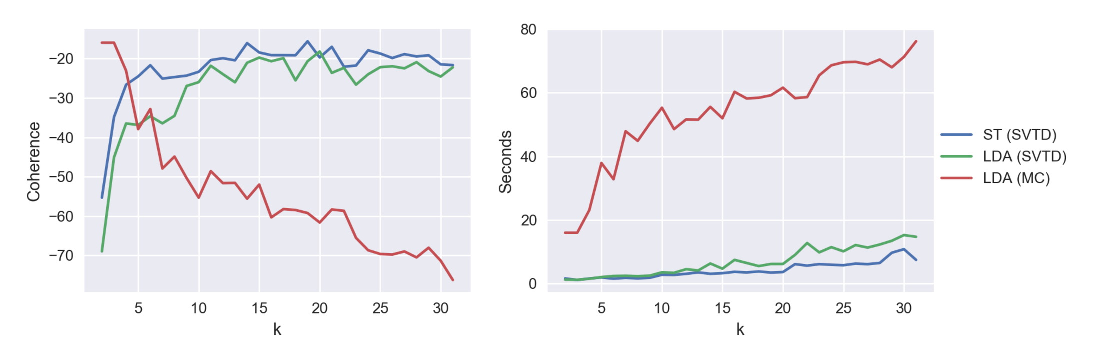


Most relevant words per topic for $k = 3$

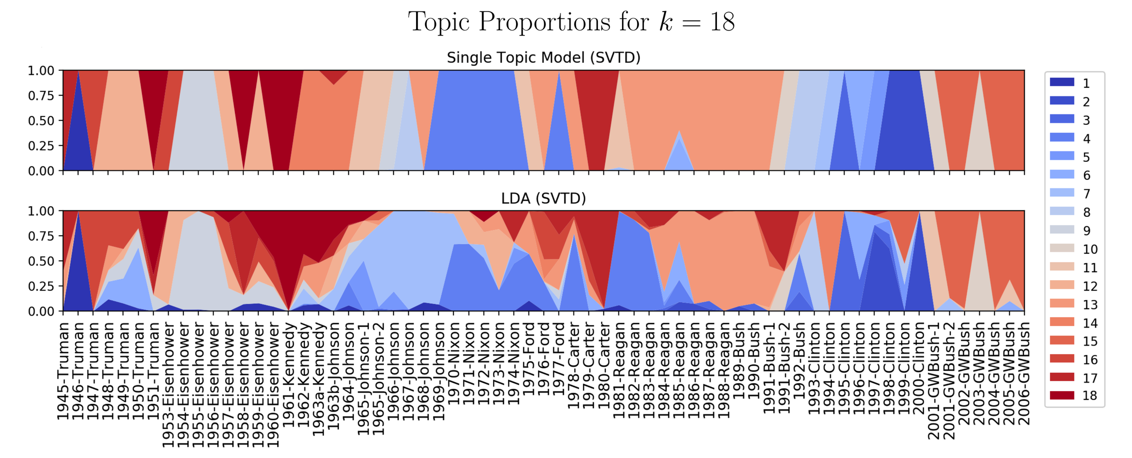
single-topic model (SVTD)
Topic 1: know, dead, tower, master
Topic 2: snake, heavy, master, above
Topic 3: light, Christ, created, lighted

LDA (SVTD)
Topic 1: dead, tower, wise, said
Topic 2: I saw, heavy, snake, master
Topic 3: light, created, near, Christ

State of The Union Addresses



Topic Proportions for $k = 18$



Some Interesting Topics for $k = 18$

single-topic model (SVTD)	LDA (SVTD)
Topic 13: federal, Hussein, intelligence	**Topic 2:** college, affordable, children, child
Topic 14: alliance, Atlantic, free, Europe	**Topic 7:** Vietnam, south, tonight, north, conflict
Topic 15: terrorists, Iraq, terror, Iraqi	**Topic 17:** soviet, military, peace, disarmament
Topic 18: disarmament, space, defense, soviets	**Topic 18:** space, civil, defense, Latin

References

[1] A. Anandkumar et al., (2012), A method of moments for mixture models and HMM.
[2] J.Y Zou et al., (2013), Contrastive learning using spectral methods.
[3] A. Anandkumar et al., (2012), A Spectral Method for LDA.
[4] A. Anandkumar et al., (2014), Tensor decomposition for learning latent variable models.
[5] T. Kolda et al., (2009), Tensor decompositions and applications.