A New Method of Moments for Latent Variable Models

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Matteo Ruffini, Marta Casanellas, Ricard Gavaldà

Universitat Politècnica de Catalunya,
Barcelona, Spain

Methods of Moments in Statistics and Machine Learning

Methods of Moments in Statistics and Machine Learning

- The method of moments was introduced by Pearson in the 1890's.

Methods of Moments in Statistics and Machine Learning

- The method of moments was introduced by Pearson in the 1890's.
- Estimates the parameters of a model by solving equations that relate the moments of the data with model parameters.

$$\mathcal{X} \sim p_\theta \to \mathbb{E}[f(X)] = g(\theta)$$

Methods of Moments in Statistics and Machine Learning

- The method of moments was introduced by Pearson in the 1890's.
- Estimates the parameters of a model by solving equations that relate the moments of the data with model parameters.

$$\mathcal{X} \sim p_\theta \to \mathbb{E}[f(X)] = g(\theta)$$

- In the last decade has been used in machine learning to obtain PAC-learning algorithms for topic models, hidden Markov models, mixtures of Gaussians, etc.

Methods of Moments in Statistics and Machine Learning

- The method of moments was introduced by Pearson in the 1890's.
- Estimates the parameters of a model by solving equations that relate the moments of the data with model parameters.

$$\mathcal{X} \sim p_\theta \rightarrow \mathbb{E}[f(X)] = g(\theta)$$

- In the last decade has been used in machine learning to obtain PAC-learning algorithms for topic models, hidden Markov models, mixtures of Gaussians, etc.

This Paper

- Introduce improved methods of moments for topic models.
- Experimentally validate their performance against traditional learning methods (e.g. Gibbs Sampling).

Agenda

1. Topic Models and Method of Moments.
2. Our Method.
3. Experiments.

The Single Topic Model

The Single Topic Model

A generative process for texts:

- We have k latent topics.

The Single Topic Model

A generative process for texts:

- We have k latent topics.
- A text only deals with a unique topic i with probability ω_i: $\mathbb{P}[\text{Topic} = i] = \omega_i$.

The Single Topic Model

A generative process for texts:

- We have k latent topics.
- A text only deals with a unique topic i with probability ω_i: $\mathbb{P}[\text{Topic} = i] = \omega_i$.
- Given the latent topic, all the words of a text are sampled from a discrete distribution with parameter $\mu_i \in \mathbb{R}^d$:
$$\mathbb{P}[\text{Sample word } j | \text{topic } = i] = (\mu_i)_j$$

The Single Topic Model

A generative process for texts:

- We have k latent topics.
- A text only deals with a unique topic i with probability ω_i: $\mathbb{P}[\text{Topic} = i] = \omega_i$.
- Given the latent topic, all the words of a text are sampled from a discrete distribution with parameter $\mu_i \in \mathbb{R}^d$:

$$\mathbb{P}[\text{Sample word } j | \text{topic } = i] = (\mu_i)_j$$

Notation:

- d vocabulary size.
- x_j one-hot encoded jth word of a document.

Parameters:

- The topics $M = [\mu_1, ..., \mu_k] \in \mathbb{R}^{d \times k}$.
- Weights $\omega = (\omega_1, ..., \omega_k) \in \mathbb{R}^k$.

Latent Dirichlet Allocation

Latent Dirichlet Allocation

A generative process for texts:

- We have k latent topics.

Latent Dirichlet Allocation

A generative process for texts:

- We have k latent topics.
- A text deals with a multitude of topics, sampled from a Dirichlet distribution.

Latent Dirichlet Allocation

A generative process for texts:

- We have k latent topics.
- A text deals with a multitude of topics, sampled from a Dirichlet distribution.
- First, you sample the topic proportions for the text

$$h \approx Dirichlet(\omega)$$

Latent Dirichlet Allocation

A generative process for texts:

- We have k latent topics.
- A text deals with a multitude of topics, sampled from a Dirichlet distribution.
- First, you sample the topic proportions for the text

$$h \approx Dirichlet(\omega)$$

- Then you sample the latent topic of each word:

$$\mathbb{P}[\text{Topic } i] = (h)_i$$

Latent Dirichlet Allocation

A generative process for texts:

- We have k latent topics.
- A text deals with a multitude of topics, sampled from a Dirichlet distribution.
- First, you sample the topic proportions for the text

$$h \approx Dirichlet(\omega)$$

- Then you sample the latent topic of each word:

$$\mathbb{P}[\text{Topic } i] = (h)_i$$

- Last, you sample the word, depending on its topic:

$$\mathbb{P}[\text{Sample word } j | \text{Topic } = i] = (\mu_i)_j$$

Latent Dirichlet Allocation

A generative process for texts:

- We have k latent topics.
- A text deals with a multitude of topics, sampled from a Dirichlet distribution.
- First, you sample the topic proportions for the text

$$h \approx Dirichlet(\omega)$$

- Then you sample the latent topic of each word:

$$\mathbb{P}[\text{Topic } i] = (h)_i$$

- Last, you sample the word, depending on its topic:

$$\mathbb{P}[\text{Sample word } j | \text{Topic } = i] = (\mu_i)_j$$

Parameters:

- The topics $M = [\mu_1, ..., \mu_k] \in \mathbb{R}^{d \times k}$
- Weights $\omega = (\omega_1, ..., \omega_k) \in \mathbb{R}^k$

Learning a Topic Model

From an iid sample

$$\mathcal{X} = \{x^{(1)}, \dots, x^{(n)}\}, \;\; x^{(i)} = \{x_1^{(i)}, x_2^{(i)}, x_3^{(i)}, \dots\}$$

We want to recover the parameters of the model:

- Single Topic Model:

$$(\mu_1, \dots, \mu_k, \omega)$$

- Latent Dirichlet Allocation:

$$(\mu_1, \dots, \mu_k, \omega)$$

Learning a Topic Model

From an iid sample

$$\mathcal{X} = \{x^{(1)}, ..., x^{(n)}\}, \;\; x^{(i)} = \{x_1^{(i)}, x_2^{(i)}, x_3^{(i)}, ...\}$$

We want to recover the parameters of the model:

- Single Topic Model:

$$(\mu_1, ..., \mu_k, \omega)$$

- Latent Dirichlet Allocation:

$$(\mu_1, ..., \mu_k, \omega)$$

Likelihood-based methods: (EM, sampling, variational methods)

- Either very slow or poor guarantees.

Spectral Method of Moments [Anandkumar et al., (2014)]

Spectral Method of Moments [Anandkumar et al., (2014)]

Applicable to any model admitting a parametrization in terms of centers and weights:

$$M = [\mu_1, ..., \mu_k] \in \mathbb{R}^{d \times k}, \quad \omega = (\omega_1, ..., \omega_k) \in \mathbb{R}^k$$

Spectral Method of Moments [Anandkumar et al., (2014)]

Applicable to any model admitting a parametrization in terms of centers and weights:

$$M = [\mu_1, ..., \mu_k] \in \mathbb{R}^{d \times k}, \quad \omega = (\omega_1, ..., \omega_k) \in \mathbb{R}^k$$

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

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

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

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

Spectral Method of Moments [Anandkumar et al., (2014)]

Applicable to any model admitting a parametrization in terms of centers and weights:

$$M = [\mu_1, \dots, \mu_k] \in \mathbb{R}^{d\times k}, \quad \omega = (\omega_1, \dots, \omega_k) \in \mathbb{R}^k$$

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

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

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

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

2. Retrieve an estimate of model parameters $(\hat{\mu}_1, \dots, \hat{\mu}_k, \hat{\omega})$ with tensor decomposition:

$$\hat{M}_1 \approx \sum_{i=1}^{k} \hat{\omega}_i \hat{\mu}_i, \quad \hat{M}_2 \approx \sum_{i=1}^{k} \hat{\omega}_i \hat{\mu}_i \otimes \hat{\mu}_i, \quad \hat{M}_3 \approx \sum_{i=1}^{k} \hat{\omega}_i \hat{\mu}_i \otimes \hat{\mu}_i \otimes \hat{\mu}_i$$

Pros and Cons

Pros

- Fast – linear in the *sample size*.
- Reduce the model-learning task to a tensor decomposition problem.
- PAC-style guarantees.
- It is the ideal setting for topic models.

Pros and Cons

Pros

- Fast – linear in the *sample size*.
- Reduce the model-learning task to a tensor decomposition problem.
- PAC-style guarantees.
- It is the ideal setting for topic models.

Improvement Points:

- The sample complexity of moment estimators for topic models can be improved.
- Tensor decomposition algorithms either slow or not robust.

Our Paper

- Provide improved moment estimators for the Single Topic Model and LDA.
- Provide a new tensor decomposition algorithm, fast and robust to perturbations.
- Test the proposed method on real data.

Our Paper

- Provide improved moment estimators for the Single Topic Model and LDA.
- Provide a new tensor decomposition algorithm, fast and robust to perturbations.
- Test the proposed method on real data.

Moment Estimators:

From an iid sample $\mathcal{X} = \{x^{(1)}, ..., x^{(n)}\}$, $x^{(i)} = \{x_1^{(i)}, x_2^{(i)}, x_3^{(i)}, ...\}$:

$$\mathbb{E}[\hat{M}_1] = M_1 = \sum_{i=1}^{k} \omega_i \mu_i, \quad \mathbb{E}[\hat{M}_2] = M_2 = \sum_{i=1}^{k} \omega_i \mu_i \otimes \mu_i, \quad \mathbb{E}[\hat{M}_3] = M_3 = \sum_{i=1}^{k} \omega_i \mu_i \otimes \mu_i \otimes \mu_i$$

Moment Estimators:

From an iid sample $\mathcal{X} = \{x^{(1)}, ..., x^{(n)}\}$, $x^{(i)} = \{x_1^{(i)}, x_2^{(i)}, x_3^{(i)}, ...\}$:

$$\mathbb{E}[\hat{M}_1] = M_1 = \sum_{i=1}^{k} \omega_i \mu_i, \quad \mathbb{E}[\hat{M}_2] = M_2 = \sum_{i=1}^{k} \omega_i \mu_i \otimes \mu_i, \quad \mathbb{E}[\hat{M}_3] = M_3 = \sum_{i=1}^{k} \omega_i \mu_i \otimes \mu_i \otimes \mu_i$$

Single Topic Model:

- [Anandkumar et al. (2012a)]

$$\hat{M}_1 = \sum_{i=1}^{n} \frac{x_1^{(i)}}{n}, \quad \hat{M}_2 = \sum_{i=1}^{n} \frac{x_1^{(i)} \otimes x_2^{(i)}}{n}, \quad \hat{M}_3 = \sum_{i=1}^{n} \frac{x_1^{(i)} \otimes x_2^{(i)} \otimes x_3^{(i)}}{n}$$

- [Zou et al. (2013)]: For each document, uses all the possible triples, in closed form.

Moment Estimators for Topic Models

Moment Estimators for Topic Models

Our proposal:

- Start from the estimators of [Anandkumar et al. (2012a)]

$$\hat{M}_1 = \sum_{i=1}^{n} \frac{x_1^{(i)}}{n}, \quad \hat{M}_2 = \sum_{i=1}^{n} \frac{x_1^{(i)} \otimes x_2^{(i)}}{n}, \quad \hat{M}_3 = \sum_{i=1}^{n} \frac{x_1^{(i)} \otimes x_2^{(i)} \otimes x_3^{(i)}}{n}$$

Moment Estimators for Topic Models

Our proposal:

- Start from the estimators of [Anandkumar et al. (2012a)]

$$\hat{M}_1 = \sum_{i=1}^{n} \frac{x_1^{(i)}}{n}, \quad \hat{M}_2 = \sum_{i=1}^{n} \frac{x_1^{(i)} \otimes x_2^{(i)}}{n}, \quad \hat{M}_3 = \sum_{i=1}^{n} \frac{x_1^{(i)} \otimes x_2^{(i)} \otimes x_3^{(i)}}{n}$$

- Extend them to consider all the possible triples, giving more weight to longer documents.

Moment Estimators for Topic Models

Our proposal:

- Start from the estimators of [Anandkumar et al. (2012a)]

$$\hat{M}_1 = \sum_{i=1}^{n} \frac{x_1^{(i)}}{n}, \quad \hat{M}_2 = \sum_{i=1}^{n} \frac{x_1^{(i)} \otimes x_2^{(i)}}{n}, \quad \hat{M}_3 = \sum_{i=1}^{n} \frac{x_1^{(i)} \otimes x_2^{(i)} \otimes x_3^{(i)}}{n}$$

- Extend them to consider all the possible triples, giving more weight to longer documents.
- Intuition: longer documents have a less noisy signal.

Moment Estimators for Topic Models

Our proposal:

- Start from the estimators of [Anandkumar et al. (2012a)]

$$\hat{M}_1 = \sum_{i=1}^{n} \frac{x_1^{(i)}}{n}, \quad \hat{M}_2 = \sum_{i=1}^{n} \frac{x_1^{(i)} \otimes x_2^{(i)}}{n}, \quad \hat{M}_3 = \sum_{i=1}^{n} \frac{x_1^{(i)} \otimes x_2^{(i)} \otimes x_3^{(i)}}{n}$$

- Extend them to consider all the possible triples, giving more weight to longer documents.
- Intuition: longer documents have a less noisy signal.

In the Paper:

- We provide sample complexity bounds for the proposed estimators.
- We show that the proposed estimators have a better sample complexity.
- We provide a variation of these estimators for LDA.

Moment Estimators for Topic Models

Experiment:

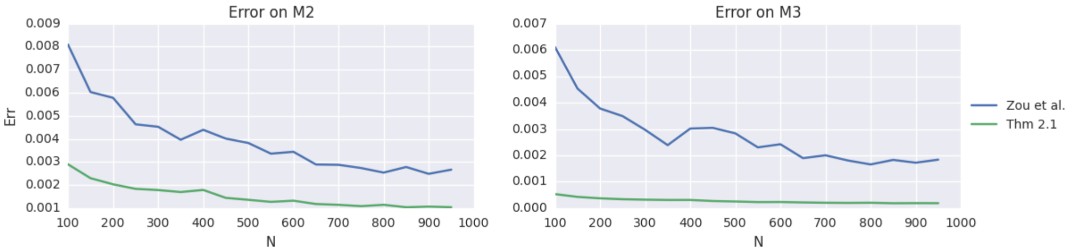
For various sample sizes n:

- Generate a dataset as the Single Topic Model with parameters $(\mu_1, ..., \mu_k, \omega)$.
- Calculate the moments with our estimators and with those from [Zou et al. (2013)].
- For each estimator calculate

$$Err = \| \sum_{i=1}^{k} \omega_i \mu_i \otimes \mu_i - \hat{M}_2 \|$$

$$Err = \| \sum_{i=1}^{k} \omega_i \mu_i \otimes \mu_i \otimes \mu_i - \hat{M}_3 \|$$

Moment Estimators for Topic Models



Our Paper

- Provide improved moment estimators for the Single Topic Model and LDA.
- Provide a new tensor decomposition algorithm, fast and robust to perturbations
- Test the proposed method on real data.

Tensor Decomposition for Methods of Moments

Objective You have: M_1, M_2, M_3.

- You want to obtain: $M = [\mu_1, ..., \mu_k]$ and ω such that:

$$M_1 = \sum_{i=1}^{k} \omega_i \mu_i, \quad M_2 = \sum_{i=1}^{k} \omega_i \mu_i \otimes \mu_i, \quad M_3 = \sum_{i=1}^{k} \omega_i \mu_i \otimes \mu_i \otimes \mu_i$$

- If the moments are empirical (perturbed), returns $(\hat{M}, \hat{\omega})$ close to (M, ω).

Tensor Decomposition for Methods of Moments

Objective You have: M_1, M_2, M_3.

- You want to obtain: $M = [\mu_1, ..., \mu_k]$ and ω such that:

$$M_1 = \sum_{i=1}^{k} \omega_i \mu_i, \quad M_2 = \sum_{i=1}^{k} \omega_i \mu_i \otimes \mu_i, \quad M_3 = \sum_{i=1}^{k} \omega_i \mu_i \otimes \mu_i \otimes \mu_i$$

- If the moments are empirical (perturbed), returns $(\hat{M}, \hat{\omega})$ close to (M, ω).

A Scan of the Literature..

- Most used methods have no guarantees on the decomposition – ALS [Kolda et al.(2009)].
- Fast methods are sensitive to perturbations – SVD method [Anandkumar et al. (2012a)].
- Robust methods are slow – TPM is $O(k^5)$ [Anandkumar et al., (2014)].

We need something fast and robust.

A Tensor Decomposition Algorithm: SVTD

A Tensor Decomposition Algorithm: SVTD

- You have: M_1, M_2, M_3, k.
- You want to obtain: $M = [\mu_1, ..., \mu_k]$ and ω such that:

$$M_1 = \sum_{i=1}^{k} \omega_i \mu_i, \quad M_2 = \sum_{i=1}^{k} \omega_i \mu_i \otimes \mu_i, \quad M_3 = \sum_{i=1}^{k} \omega_i \mu_i \otimes \mu_i \otimes \mu_i$$

A Tensor Decomposition Algorithm: SVTD

- You have: M_1, M_2, M_3, k.
- You want to obtain: $M = [\mu_1, ..., \mu_k]$ and ω such that:

$$M_1 = \sum_{i=1}^{k} \omega_i \mu_i, \quad M_2 = \sum_{i=1}^{k} \omega_i \mu_i \otimes \mu_i, \quad M_3 = \sum_{i=1}^{k} \omega_i \mu_i \otimes \mu_i \otimes \mu_i$$

Theorem

- Let $M_{3,r} \in \mathbb{R}^{d \times d}$ be the r-th slice of M_3 and m_r the r-th row of M.
- There exists a projection of $M_{3,r}$ to a matrix $H_r \in \mathbb{R}^{k \times k}$ whose singular values are m_r.

A Tensor Decomposition Algorithm: SVTD

- You have: M_1, M_2, M_3, k.
- You want to obtain: $M = [\mu_1, ..., \mu_k]$ and ω such that:

$$M_1 = \sum_{i=1}^{k} \omega_i \mu_i, \quad M_2 = \sum_{i=1}^{k} \omega_i \mu_i \otimes \mu_i, \quad M_3 = \sum_{i=1}^{k} \omega_i \mu_i \otimes \mu_i \otimes \mu_i$$

Theorem

- Let $M_{3,r} \in \mathbb{R}^{d \times d}$ be the r-th slice of M_3 and m_r the r-th row of M.
- There exists a projection of $M_{3,r}$ to a matrix $H_r \in \mathbb{R}^{k \times k}$ whose singular values are m_r.

Algorithm

- Loop $r : 1 \to d$
  - Find H_r properly projecting $M_{3,r}$ to $\mathbb{R}^{k \times k}$ (*whitening* step).
  - Find the r-th row of M as the singular values of H_r.

SVTD - Considerations

Remarks:

- With no perturbations on the moments, we get the exact model parameters.
- The row i of M is the singular values of H_i, which are robust to perturbations.
- Time complexity: $O(d^2k + k^3 + d^3k)$ – can get to $O(dk^2n)$ with optimized implementations.

SVTD - Considerations

Remarks:

- With no perturbations on the moments, we get the exact model parameters.
- The row i of M is the singular values of H_i, which are robust to perturbations.
- Time complexity: $O(d^2k + k^3 + d^3k)$ – can get to $O(dk^2n)$ with optimized implementations.

Comparison with Other Methods

- *SVD method* [Anandkumar et al. (2012a)]: similar to SVTD but based on singular vectors. We expect it to be less robust to perturbations.
- *TPM* [Anandkumar et al., (2014)] has a worse dependence on k: It should be slower for high number of topics.
- *ALS* [Kolda et al.(2009)]: no whitening – i.e. should be slower. No guarantees on the decomposition.

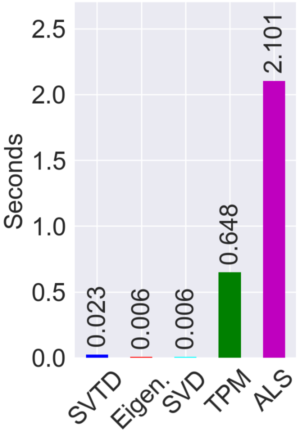
Experiments

For various sample sizes n:

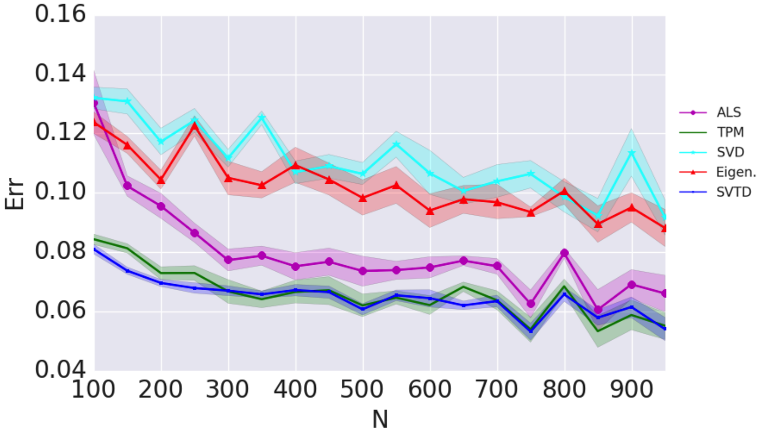
- Generate a dataset as the Single Topic Model with parameters $(\mu_1, \dots, \mu_k, \omega)$.
- Calculate the moments with the proposed estimators.
- Perform tensor decomposition with various methods.
- Calculate the average running time for each method.
- Calculate the decomposition error:

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

Experiments



(a) Running Time



(b) Decomposition Error

Our Paper

- Provide improved moment estimators for the Single Topic Model and LDA.
- Provide a new tensor decomposition algorithm, fast and robust to perturbations
- Test the proposed method on real data.

Objective

We have:

- An end-to-end algorithm $\mathcal{A}$ to learn from data topic models:

$$\mathcal{A} : \mathcal{X} \to (\mu_1, ..., \mu_k, \omega)$$

- Good performance in comparison with other methods of moments.

Objective

We have:

- An end-to-end algorithm $\mathcal{A}$ to learn from data topic models:

$$\mathcal{A} : \mathcal{X} \to (\mu_1, ..., \mu_k, \omega)$$

- Good performance in comparison with other methods of moments.

We want to:

- Test our approach on real data.
- Compare it with state-of-the-art methods, i.e. Sampling methods.

Objective

We have:

- An end-to-end algorithm $\mathcal{A}$ to learn from data topic models:

$$\mathcal{A} : \mathcal{X} \to (\mu_1, \ldots, \mu_k, \omega)$$

- Good performance in comparison with other methods of moments.

We want to:

- Test our approach on real data.
- Compare it with state-of-the-art methods, i.e. Sampling methods.

Data:

- US presidents' *State of the Union Addresses*.
- $n = 65$ speeches, $d = 1184$ words.

Evaluation Method

For various values of the number of latent topics k:

- Learn a Single Topic Model and an LDA with k topics with the proposed approach.
- Learn an LDA with k topics with Gibbs Sampling [Griffiths and Steyvers (2004)].

Evaluation Method

For various values of the number of latent topics k:

- Learn a Single Topic Model and an LDA with k topics with the proposed approach.
- Learn an LDA with k topics with Gibbs Sampling [Griffiths and Steyvers (2004)].

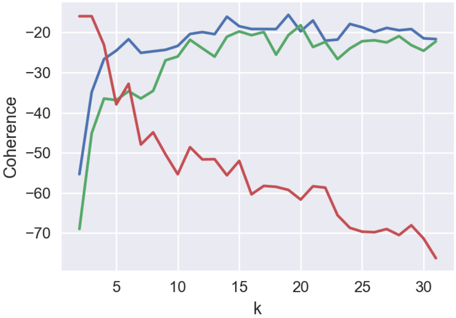
For each learned model:

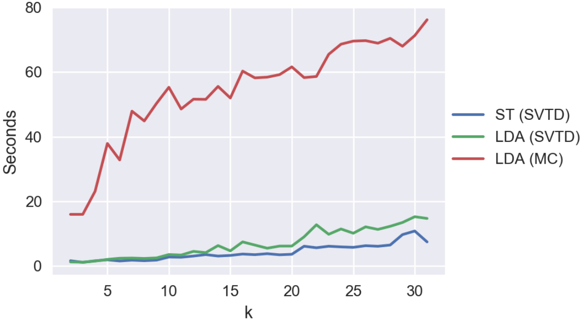
- Calculate the coherence of the retrieved topics:

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

- Calculate the running time needed to learn the model.

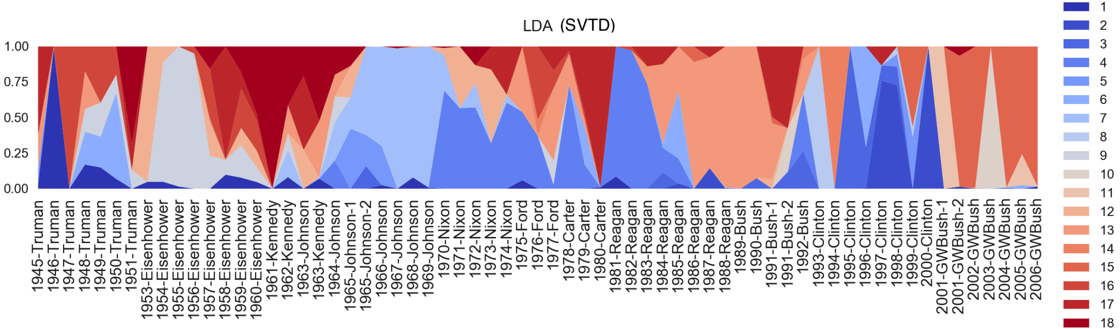
Results I: quantitative analysis



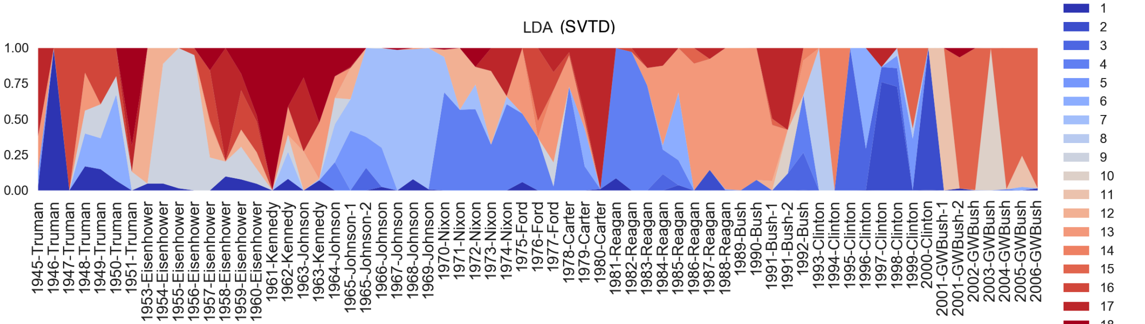


Results II: qualitative analysis

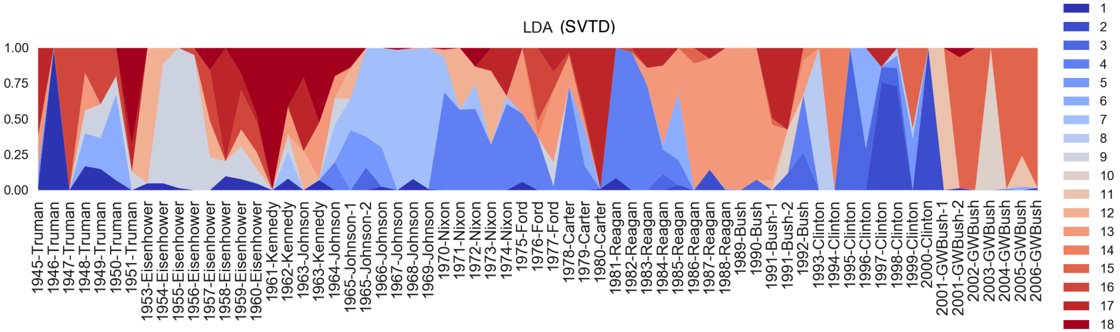
- Keep the LDA model.
- Select a value of k for which we have a high coherence.
- For each speech, visualize how much it deals with the various topics.



Results II: qualitative analysis

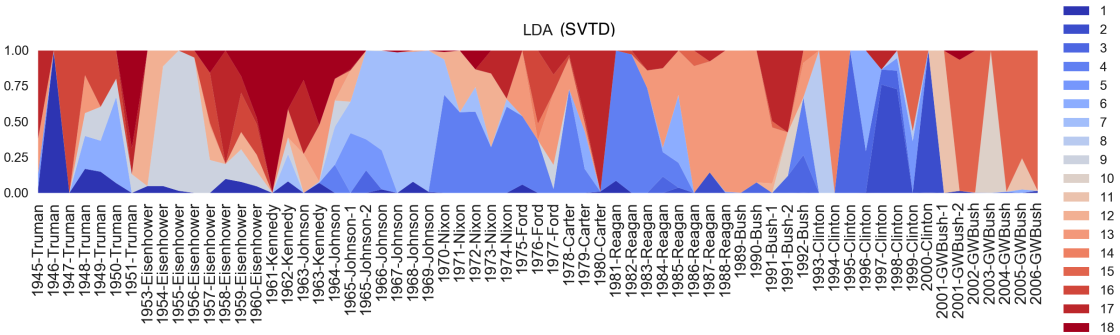


Results II: qualitative analysis



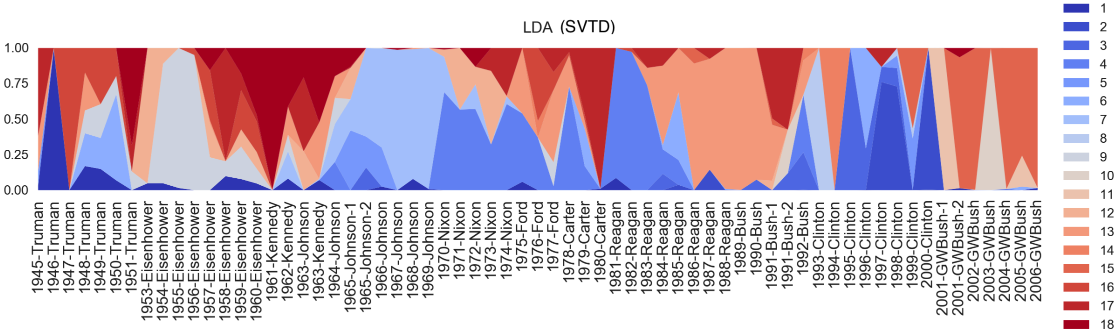
- **Topic 2:** college, affordable, children, child.

Results II: qualitative analysis



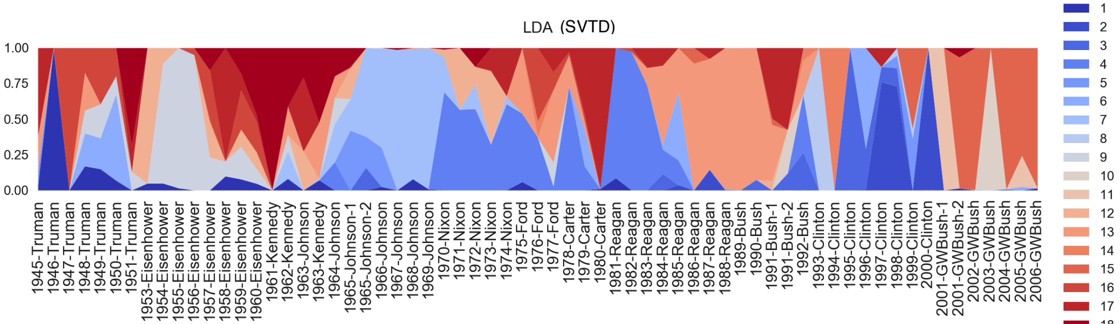
- **Topic 2:** college, affordable, children, child.
- **Topic 7:** Vietnam, south, tonight, north, conflict.

Results II: qualitative analysis



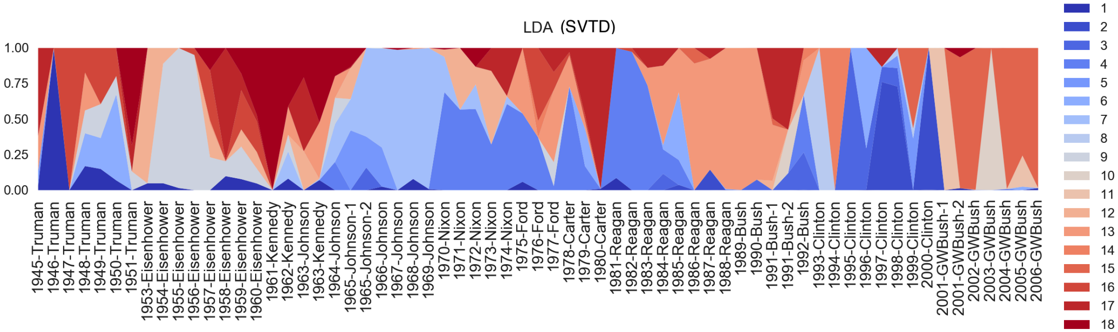
- **Topic 2:** college, affordable, children, child.
- **Topic 7:** Vietnam, south, tonight, north, conflict.
- **Topic 15:** Iraq, terrorists, terrorist, seniors.

Results II: qualitative analysis



- **Topic 2:** college, affordable, children, child.
- **Topic 7:** Vietnam, south, tonight, north, conflict.
- **Topic 15:** Iraq, terrorists, terrorist, seniors.
- **Topic 17:** soviet, military, peace, disarmament.

Results II: qualitative analysis



- **Topic 2:** college, affordable, children, child.
- **Topic 7:** Vietnam, south, tonight, north, conflict.
- **Topic 15:** Iraq, terrorists, terrorist, seniors.
- **Topic 17:** soviet, military, peace, disarmament.
- **Topic 18:** space, civil, defense, Latin.

A New Method of Moments for Latent Variable Models

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Matteo Ruffini, Marta Casanellas, Ricard Gavaldà

Universitat Politècnica de Catalunya,
Barcelona, Spain

References I

Anandkumar, A., Chaudhuri, K., Hsu, D. J., Kakade, S. M., Song, L., and Zhang, T. (2011).
Spectral methods for learning multivariate latent tree structure.
In Advances in Neural Information Processing Systems.

Anandkumar, A., Foster, D. P., Hsu, D. J., Kakade, S. M., and Liu, Y. K. (2012b).
A spectral algorithm for latent dirichlet allocation.
In Advances in Neural Information Processing Systems.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014).
Tensor decompositions for learning latent variable models.
The Journal of Machine Learning Research.

Anandkumar, A., Hsu, D., and Kakade, S. M. (2012a).
A method of moments for mixture models and hidden Markov models.
In Conference on Learning Theory.

Chaganty, A. T., and Liang, P. (2013).
Spectral experts for estimating mixtures of linear regressions.
In International Conference on Machine Learning.

Arabshahi, F., and Anandkumar, A. (2017).
Spectral Methods for Correlated Topic Models.
In Artificial Intelligence and Statistics.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977).
Maximum likelihood from incomplete data via the EM algorithm.
Journal of the royal statistical society.

References II

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S. and Bengio, Y. (2014).
Generative adversarial nets.
In Advances in neural information processing systems.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B. and Smola, A. (2012).
A kernel two-sample test.
Journal of Machine Learning Research.

Griffiths, T. L. and Steyvers, M. (2004).
Finding scientific topics.
Proceedings of the National academy of Sciences.

Hsu, D., Kakade, S. M. and Zhang, T. (2012).
A spectral algorithm for learning hidden Markov models.
Journal of Computer and System Sciences.

Hsu, D. and Kakade, S. M. (2013).
Learning mixtures of spherical Gaussians: moment methods and spectral decompositions.
In Proceedings of the 4-th conference on Innovations in Theoretical Computer Science.

Jain, P. and Oh, S. (2014).
Learning mixtures of discrete product distributions using spectral decompositions.
In Conference on Learning Theory.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. and Saul, L. K. (1999).
An introduction to variational methods for graphical models.
Machine learning Journal.

References III

Kolda, T. G. and Bader, B. W., (2009).
Tensor decompositions and applications.
SIAM review.

Kuleshov, V., Chaganty, A. and Liang, P., (2015).
Tensor factorization via matrix factorization.
In Artificial Intelligence and Statistics.

Pepke, P. and Ver Steeg, G. (2016).
Comprehensive discovery of subsample gene expression components by information explanation: therapeutic implications in cancer.
BMC Medical Genomics.

Roqueiro, D, Witteveen, M. J, Anttila, V., Terwindt, G., van den Maagdenberg, A. and Borgwardt, K., (2015).
In silico phenotyping via co-training for improved phenotype prediction from genotype.
Bioinformatics.

Ruffini, M., Casanellas, M., and Gavaldà, R. (2018)
A New Spectral Method for Latent Variable Models.
Machine Learning Journal.

Ruffini, M., Rabusseau, G. and Balle, B. (2017).
Hierarchical Methods of Moments.
In Advances in Neural Information Processing Systems.

Ruffini, M. and Gavaldà, R. (2018).
Hierarchical Methods of Moments for Clustering High-Dimensional Binary Data.
Submitted.

References IV

Ruffini, M., Gavalda, R. and Limon, E. (2017).
Clustering Patients with Tensor Decomposition.
In Machine Learning for Healthcare Conference.

He, X., Li, L., Roqueiro, D. and Borgwardt, K. (2017).
Multi-view Spectral Clustering on Conflicting Views
Joint European Conference on Machine Learning and Knowledge Discovery in Databases

Wang, Y. and Anandkumar, A. (2016).
Online and differentially-private tensor decomposition.
In Advances in Neural Information Processing Systems.

Wang, Y., Cai, Z., Stothard, P., Moore, S., Goebel, R., Wang, L. and Lin, G (2016).
Fast accurate missing SNP genotype local imputation.
BMC research notes.

Warde-Farley, D., Brudno, M., Morris, Q. and Goldenberg, A. (2012).
Mixture model for sub-phenotyping in GWAS.
Biocomputing.

Zou, J. Y., Hsu, D. J., Parkes, D. C., and Adams, R. P.
Contrastive learning using spectral methods.
Advances in Neural Information Processing Systems.