Identifiability and Learning of Topic Models: Tensor Decompositions under Structural Constraints

Anima Anandkumar

U.C. Irvine

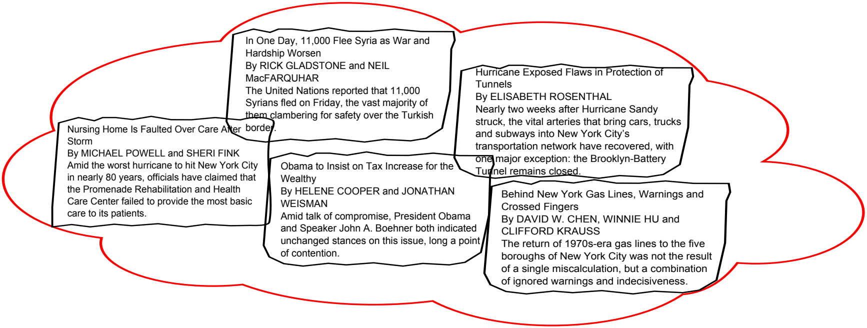
Joint work with Daniel Hsu, Majid Janzamin
Adel Javanmard and Sham Kakade.

Latent Variable Modeling

Goal: Discover hidden effects from observed measurements

Topic Models

- Observations: words. Hidden: topics.



Modeling communities in social networks, modeling gene regulation ...

Challenges in Learning Topic Models

Learning Topic Models Using Word Observations

Challenges in Identifiability

- When can topics be identified?
- Conditions on the model parameter, e.g. on topic-word matrix Φ and on topic proportions distributions (h)?
- Does identifiability also lead to tractable algorithms?

Challenges in Learning Topic Models

Learning Topic Models Using Word Observations

Challenges in Identifiability

- When can topics be identified?
- Conditions on the model parameter, e.g. on topic-word matrix Φ and on topic proportions distributions (h)?
- Does identifiability also lead to tractable algorithms?

Challenges in Design of Learning Algorithms

- Maximum likelihood learning of topic models NP-hard (Arora et. al.)
- In practice, heuristics such as Gibbs sampling, variation Bayes etc.
- Guaranteed learning with minimal assumptions? Efficient methods? Low sample and computational complexities?

Challenges in Learning Topic Models

Learning Topic Models Using Word Observations

Challenges in Identifiability

- When can topics be identified?
- Conditions on the model parameter, e.g. on topic-word matrix Φ and on topic proportions distributions (h)?
- Does identifiability also lead to tractable algorithms?

Challenges in Design of Learning Algorithms

- Maximum likelihood learning of topic models NP-hard (Arora et. al.)
- In practice, heuristics such as Gibbs sampling, variation Bayes etc.
- Guaranteed learning with minimal assumptions? Efficient methods? Low sample and computational complexities?

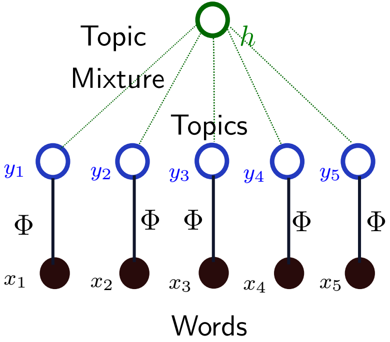
Moment-based approach: learning using low order observed moments

Probabilistic Topic Models

- Useful abstraction for automatic categorization of documents
- Observed: words. Hidden: topics.
- Bag of words: order of words does not matter

Graphical model representation

- l words in a document $x_1, \ldots, x_l$.
- h: proportions of topics in a document.
- Word x_i generated from topic y_i.
- Exchangeability: $\boxed{x_1 \perp\!\!\!\perp x_2 \perp\!\!\!\perp \ldots | h}$
- $\boxed{\Phi(i,j) := \mathbb{P}[x_m = i | y_m = j]}$: topic-word matrix.



Formulation as Linear Models

Distribution of the topic proportions vector h

If there are k topics, distribution over the simplex Δ^{k-1}

$$\boxed{\Delta^{k-1} := \{h \in \mathbb{R}^k, h_i \in [0,1], \sum_i h_i = 1\}.}$$

Distribution of the words $x_1, x_2, \ldots$

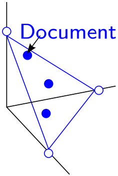
- Order n words in vocabulary. If x_1 is j^{th} word, assign $e_j \in \mathbb{R}^n$
- Distribution of each x_i: supported on vertices of Δ^{n-1}.

Properties

- Linear Model: $\boxed{\mathbb{E}[x_i|h] = \Phi h}$.
- Multiview model: h is fixed and multiple words (x_i) are generated.

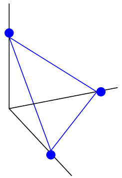
Geometric Picture for Topic Models

Topic proportions vector (h)



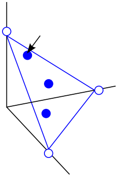
Geometric Picture for Topic Models

Single topic (h)



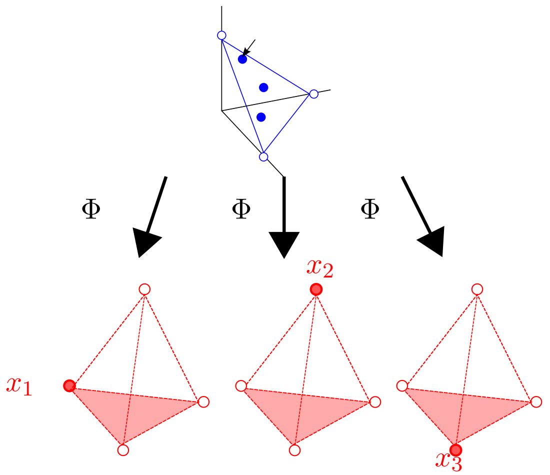
Geometric Picture for Topic Models

Topic proportions vector (h)



Geometric Picture for Topic Models

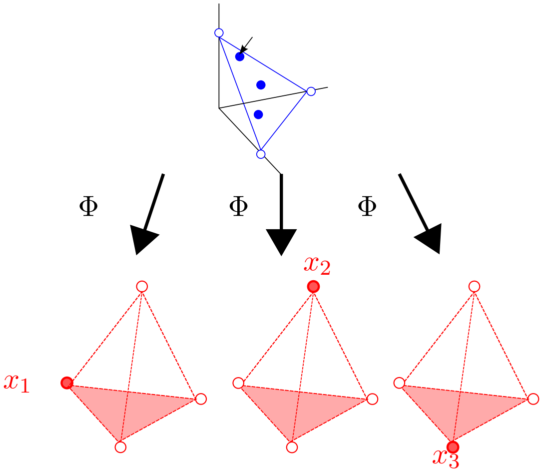
Topic proportions vector (h)



Word generation $(x_1, x_2, \ldots)$

Geometric Picture for Topic Models

Topic proportions vector (h)



Word generation $(x_1, x_2, \ldots)$

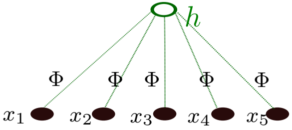
Moment-based estimation: co-occurrences of words in documents

Outline

Recap of CP Decomposition

Recall form of moments for single topic/Dirichlet model.

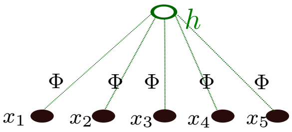
- $\boxed{\mathbb{E}[x_i|h] = \Phi h.}$ $\boxed{\vec{\lambda} := [\mathbb{E}[h]]_i.}$
- Learn topic-word matrix Φ, vector $\vec{\lambda}$



Recap of CP Decomposition

Recall form of moments for single topic/Dirichlet model.

- $\boxed{\mathbb{E}[x_i|h] = \Phi h.}$ $\boxed{\vec{\lambda} := [\mathbb{E}[h]]_i.}$
- Learn topic-word matrix Φ, vector $\vec{\lambda}$



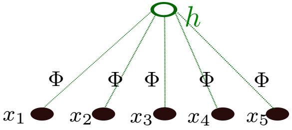
Pairs Matrix M_2

$$M_2 := \mathbb{E}[x_1 x_2^\top] = \mathbb{E}[\mathbb{E}[x_1 x_2^\top|h]] = \Phi \mathbb{E}[hh^\top]\Phi^\top = \sum_{r=1}^{k} \lambda_r \phi_r \phi_r^\top$$

Recap of CP Decomposition

Recall form of moments for single topic/Dirichlet model.

- $\boxed{\mathbb{E}[x_i|h] = \Phi h.}$ $\boxed{\vec{\lambda} := [\mathbb{E}[h]]_i.}$
- Learn topic-word matrix Φ, vector $\vec{\lambda}$



Pairs Matrix M_2

$$M_2 := \mathbb{E}[x_1 x_2^\top] = \mathbb{E}[\mathbb{E}[x_1 x_2^\top | h]] = \Phi \mathbb{E}[hh^\top]\Phi^\top = \sum_{r=1}^{k} \lambda_r \phi_r \phi_r^\top$$

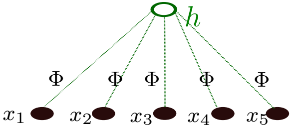
Similarly Triples Tensor M_3

$$M_3 := \mathbb{E}(x_1 \otimes x_2 \otimes x_3) = \sum_r \lambda_r \phi_r \otimes \phi_r \otimes \phi_r$$

Recap of CP Decomposition

Recall form of moments for single topic/Dirichlet model.

- $\boxed{\mathbb{E}[x_i|h] = \Phi h.}$ $\boxed{\vec{\lambda} := [\mathbb{E}[h]]_i.}$
- Learn topic-word matrix Φ, vector $\vec{\lambda}$



Pairs Matrix M_2

$$M_2 := \mathbb{E}[x_1 x_2^\top] = \mathbb{E}[\mathbb{E}[x_1 x_2^\top|h]] = \Phi \mathbb{E}[hh^\top]\Phi^\top = \sum_{r=1}^{k} \lambda_r \phi_r \phi_r^\top$$

Similarly Triples Tensor M_3

$$M_3 := \mathbb{E}(x_1 \otimes x_2 \otimes x_3) = \sum_r \lambda_r \phi_r \otimes \phi_r \otimes \phi_r$$

Matrix and Tensor Forms: $\phi_r := r^{\text{th}}$ column of Φ.

$$\boxed{M_2 = \sum_{r=1}^{k} \lambda_r \phi_r \otimes \phi_r. \qquad M_3 = \sum_{r=1}^{k} \lambda_r \phi_r \otimes \phi_r \otimes \phi_r}$$

Multi-linear Transformation

- For a tensor T, define (for matrices V_i of appropriate dimensions)

$$\boxed{[T(V_1, V_2, V_3)]_{i_1,i_2,i_3} := \sum_{j_1,j_2,j_3} (T)_{j_1,j_2,j_3} \prod_{m\in[3]} V_1(j_m, i_m)}$$

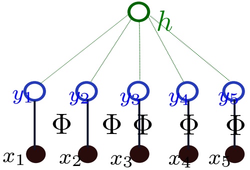
- For a matrix M_2, $\boxed{M(V_1, V_2) := V_1^\top M_2 V_2}$.

$$\boxed{T = \sum_{r=1}^{k} \lambda_r \phi_r \otimes \phi_r \otimes \phi_r}$$

$$\boxed{\begin{aligned} T(W,W,W) &= \sum_{r\in[k]} \lambda_r (W^\top \phi_r)^{\otimes 3} \\ T(I,v,v) &= \sum_{r\in[k]} \lambda_r \langle v, \phi_r\rangle^2 \phi_r. \\ T(I,I,v) &= \sum_{r\in[k]} \lambda_r \langle v, \phi_r\rangle \phi_r \phi_r^\top. \end{aligned}}$$

Form of Moments for a general Topic Model

- $\boxed{\mathbb{E}[x_i|h] = \Phi h.}$
- Learn Φ, distribution of h
- Form of moments?

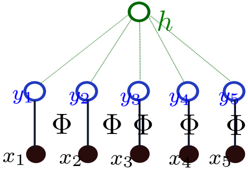


Pairs Matrix M_2

$$M_2 := \mathbb{E}[x_1 x_2^\top] = \mathbb{E}[\mathbb{E}[x_1 x_2^\top|h]] = \boxed{\Phi \mathbb{E}[hh^\top]\Phi^\top}$$

- $\mathbb{E}[hh^\top]$ NOT be a diagonal matrix.

Form of Moments for a general Topic Model



- $\boxed{\mathbb{E}[x_i|h] = \Phi h.}$
- Learn Φ, distribution of h
- Form of moments?

Pairs Matrix M_2

$$M_2 := \mathbb{E}[x_1 x_2^\top] = \mathbb{E}[\mathbb{E}[x_1 x_2^\top|h]] = \boxed{\Phi \mathbb{E}[hh^\top]\Phi^\top}$$

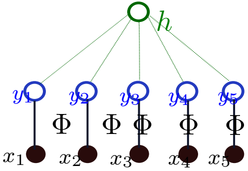
- $\mathbb{E}[hh^\top]$ NOT be a diagonal matrix.

Similarly Triples Tensor M_3

$$M_3 := \mathbb{E}(x_1 \otimes x_2 \otimes x_3) = \sum_{i,j,k} \mathbb{E}[h^{\otimes 3}]_{i,j,k} \phi_i \otimes \phi_j \otimes \phi_k = \boxed{\mathbb{E}[h^{\otimes 3}](\Phi, \Phi, \Phi)}$$

Form of Moments for a general Topic Model

- $\boxed{\mathbb{E}[x_i|h] = \Phi h.}$
- Learn Φ, distribution of h
- Form of moments?



Pairs Matrix M_2

$$M_2 := \mathbb{E}[x_1 x_2^\top] = \mathbb{E}[\mathbb{E}[x_1 x_2^\top|h]] = \boxed{\Phi \mathbb{E}[hh^\top]\Phi^\top}$$

- $\mathbb{E}[hh^\top]$ NOT be a diagonal matrix.

Similarly Triples Tensor M_3

$$M_3 := \mathbb{E}(x_1 \otimes x_2 \otimes x_3) = \sum_{i,j,k} \mathbb{E}[h^{\otimes 3}]_{i,j,k} \phi_i \otimes \phi_j \otimes \phi_k = \boxed{\mathbb{E}[h^{\otimes 3}](\Phi, \Phi, \Phi)}$$

Tucker Tensor Decomposition

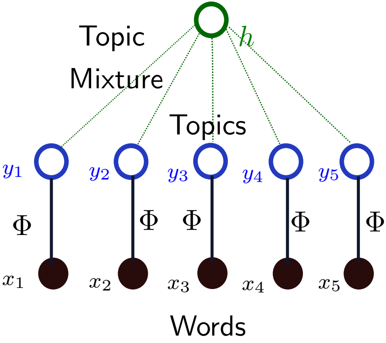
- Find decomposition $M_3 = \mathbb{E}[h^{\otimes 3}](\Phi, \Phi, \Phi)$
- Key difference from CP: $\mathbb{E}[h^{\otimes 3}]$ NOT a diagonal tensor
- Lot more parameters to estimate

Guaranteed Learning of Topic Models

Two Learning approaches

- CP Tensor decomposition: Parametric topic distributions (constraints on h) but general topic-word matrix Φ

- Tucker Tensor decomposition: Constrain topic-word matrix Φ but general (non-degenerate) distributions on h



Outline

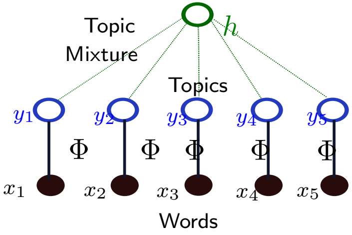
Outline

1. Introduction
2. Form of Moments
3. Matrix Case: Learning using Pairwise Moments
   - Identifiability and Learning of Topic-Word Matrix
   - Learning Latent Space Parameters of the Topic Model
4. Tensor Case: Learning From Higher Order Moments
   - Overcomplete Representations
5. Conclusion

Pairwise moments for learning

Learning using pairwise moments: minimal information.

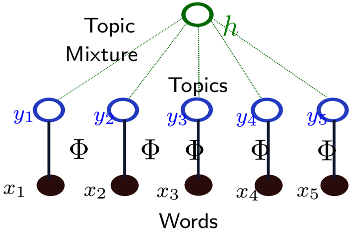
Exchangeable Topic Model



Pairwise moments for learning

Learning using pairwise moments: minimal information.

Exchangeable Topic Model

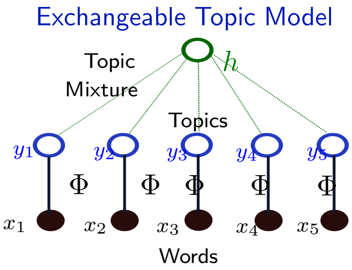


So far..

- Parametric h: Dirichlet, single topic, independent components, ...
- No restrictions on Φ (other than non-degeneracy).
- Learning using third order moment through tensor decompositions

Pairwise moments for learning

Learning using pairwise moments: minimal information.



So far..

- Parametric h: Dirichlet, single topic, independent components, ...
- No restrictions on Φ (other than non-degeneracy).
- Learning using third order moment through tensor decompositions

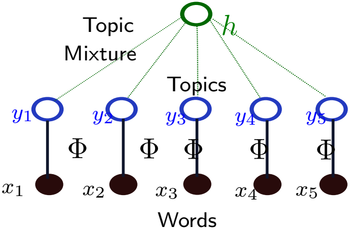
What if..

- Allow for general h: model arbitrary topic correlations
- Constrain topic-word matrix Φ:

Pairwise moments for learning

Learning using pairwise moments: minimal information.

Exchangeable Topic Model



So far..

- Parametric h: Dirichlet, single topic, independent components, ...
- No restrictions on Φ (other than non-degeneracy).
- Learning using third order moment through tensor decompositions

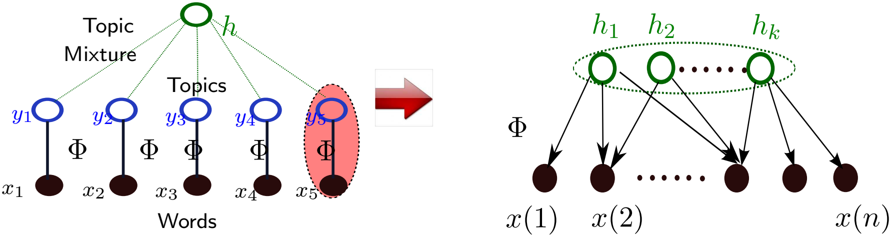
What if..

- Allow for general h: model arbitrary topic correlations
- Constrain topic-word matrix Φ: Sparsity constraints

Pairwise moments for learning

Learning using pairwise moments: minimal information.

Exchangeable Topic Model | Topic-word matrix



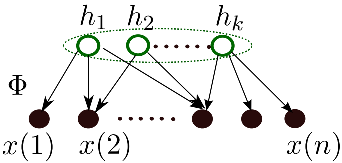
So far..

- Parametric h: Dirichlet, single topic, independent components, ...
- No restrictions on Φ (other than non-degeneracy).
- Learning using third order moment through tensor decompositions

What if..

- Allow for general h: model arbitrary topic correlations
- Constrain topic-word matrix Φ: Sparsity constraints

Some Intuitions..

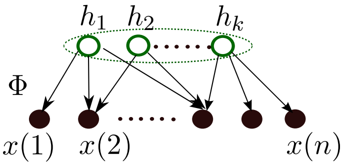


Learning using second-order moments

- Linear model: $\boxed{\mathbb{E}[x_i|h] = \Phi\, h.}$ and $\boxed{\mathbb{E}[x_1 x_2^\top] = \Phi \mathbb{E}[hh^\top]\Phi^\top}$
- Learning: recover Φ from $\Phi\mathbb{E}[hh^\top]\Phi^\top$.

Ill-posed without further restrictions

Some Intuitions..



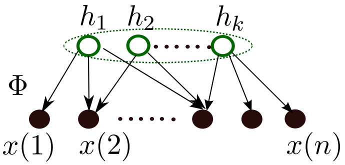
Learning using second-order moments

- Linear model: $\boxed{\mathbb{E}[x_i|h] = \Phi\, h.}$ and $\boxed{\mathbb{E}[x_1 x_2^\top] = \Phi \mathbb{E}[hh^\top]\Phi^\top}$
- Learning: recover Φ from $\Phi\mathbb{E}[hh^\top]\Phi^\top$.

Ill-posed without further restrictions

- When h is not degenerate: $\boxed{\text{recover } \Phi \text{ from } \mathrm{Col}(\Phi)}$
- No other restrictions on h: arbitrary dependencies

Some Intuitions..



Learning using second-order moments

- Linear model: $\boxed{\mathbb{E}[x_i|h] = \Phi\, h.}$ and $\boxed{\mathbb{E}[x_1 x_2^\top] = \Phi\mathbb{E}[hh^\top]\Phi^\top}$
- Learning: recover Φ from $\Phi\mathbb{E}[hh^\top]\Phi^\top$.

Ill-posed without further restrictions

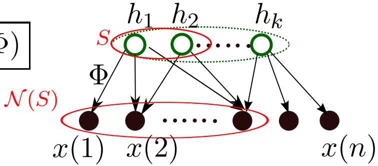
- When h is not degenerate: $\boxed{\text{recover } \Phi \text{ from } \mathrm{Col}(\Phi)}$
- No other restrictions on h: arbitrary dependencies

Sparsity constraints on topic-word matrix Φ

- Main constraint: $\boxed{\text{columns of } \Phi \text{ are sparsest vectors in } \mathrm{Col}(\Phi)}$

Sufficient Conditions for Identifiability

$\boxed{\text{columns of } \Phi \text{ are } \textcolor{red}{\text{sparsest}} \text{ vectors in } \mathrm{Col}(\Phi)}$

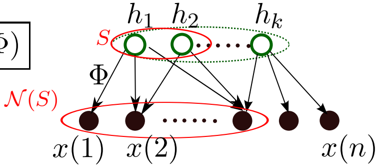


- Sufficient conditions?

Sufficient Conditions for Identifiability

$\boxed{\text{columns of } \Phi \text{ are sparsest vectors in } \mathrm{Col}(\Phi)}$

- Sufficient conditions?



Structural Condition: (Additive) Graph Expansion

$$\boxed{|\mathcal{N}(S)| > |S| + d_{\max}, \text{ for all } S \subset [k]}$$

Parametric Conditions: Generic Parameters

$$\boxed{\|\Phi v\|_0 > |\mathcal{N}_\Phi(\mathrm{supp}(v))| - |\,\mathrm{supp}(v)|}$$

A. Anandkumar, D. Hsu, A. Javanmard, and, S. M. Kakade. Learning Bayesian Networks with Latent Variables. In Proc. of Intl. Conf. on Machine Learning, June 2013.

Brief Proof Sketch

Structural Condition: (Additive) Graph Expansion

$\boxed{|\mathcal{N}(S)| > |S| + d_{\max}, \text{ for all } S \subset [k]}$

Parametric Conditions: Generic Parameters

$\boxed{\|\Phi v\|_0 > |\mathcal{N}_\Phi(\operatorname{supp}(v))| - |\operatorname{supp}(v)|}$

Structural and Parametric Conditions Imply:

When $|\operatorname{supp}(v)| > 1$, $\boxed{\|\Phi v\|_0 > |\mathcal{N}_\Phi(\operatorname{supp}(v))| - |\operatorname{supp}(v)| > d_{\max}}$

$\boxed{\text{Thus, } |\operatorname{supp}(v)| = 1, \text{ for } \Phi v \text{ to be one of } k \text{ sparsest vectors in Col}(\Phi)}$

Brief Proof Sketch

Structural Condition: (Additive) Graph Expansion

$\boxed{|\mathcal{N}(S)| > |S| + d_{\max}, \text{ for all } S \subset [k]}$

Parametric Conditions: Generic Parameters

$\boxed{\|\Phi v\|_0 > |\mathcal{N}_\Phi(\mathrm{supp}(v))| - |\mathrm{supp}(v)|}$

Structural and Parametric Conditions Imply:

When $|\mathrm{supp}(v)| > 1$, $\boxed{\|\Phi v\|_0 > |\mathcal{N}_\Phi(\mathrm{supp}(v))| - |\mathrm{supp}(v)| > d_{\max}}$

$\boxed{\text{Thus, } |\mathrm{supp}(v)| = 1, \text{ for } \Phi v \text{ to be one of } k \text{ sparsest vectors in } \mathrm{Col}(\Phi)}$

Claim: Parametric conditions are satisfied for generic parameters

Tractable Learning Algorithm

Learning Task

Recover topic-word matrix Φ from $\boxed{M_2 = \Phi \mathbb{E}[hh^\top]\Phi^\top}$.

Exhaustive search

$$\boxed{\min_{z \neq 0} \|\Phi z\|_0}$$

Convex relaxation

$$\boxed{\min_z \|\Phi z\|_1, \quad b^\top z = 1,}$$

where b is a row in Φ.

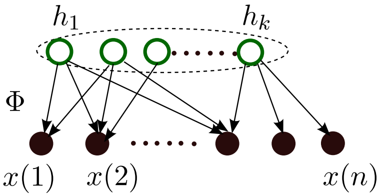
Change of Variables

$$\boxed{\min_w \|M_2^{1/2} w\|_1, \quad e_i^\top M_2^{1/2} w = 1.}$$

Under "reasonable" conditions, the above program exactly recovers Φ

Outline

Latent General Topic Models



So far: recover topic-word matrix Φ from $\boxed{\Phi\mathbb{E}[hh^\top]\Phi^\top}$.

Learning topic proportion distribution

- $\mathbb{E}[hh^\top]$ not enough to recover general distributions
- Need higher order moments to learn distribution of h
- Any models where low order moments suffice? e.g. Dirichlet/single topic require only third order moments. What about any other distributions?

Are there other topic distributions which can be learned efficiently?

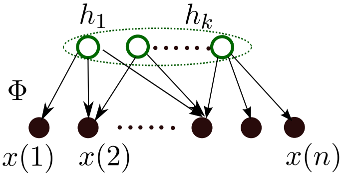
Learning Latent Bayesian Networks

BN: Markov relationships on DAG

Pa_i: parents of node i. $\mathbb{P}(h) = \prod_{i=1}^{n} \mathbb{P}(h_i | h_{\mathrm{Pa}_i})$

Linear Bayesian Network: $h_j = \sum_{i \in \mathrm{Pa}_j} \lambda_{ji} h_i + \eta_j$

$\boxed{h = \Lambda h + \eta}$ $\boxed{\mathbb{E}[x_i | \eta] = \Phi(I - \Lambda)^{-1} \eta = \Phi' \eta}$ and η_i uncorrelated



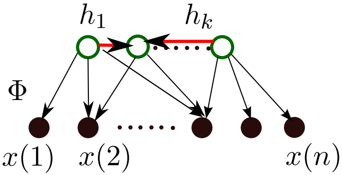
Learning Latent Bayesian Networks

BN: Markov relationships on DAG

Pa_i: parents of node i. $\mathbb{P}(h) = \prod_{i=1}^{n} \mathbb{P}(h_i | h_{\mathrm{Pa}_i})$

Linear Bayesian Network: $h_j = \sum_{i \in \mathrm{Pa}_j} \lambda_{ji} h_i + \eta_j$

$\boxed{h = \Lambda h + \eta}$ $\boxed{\mathbb{E}[x_i|\eta] = \Phi(I - \Lambda)^{-1}\eta = \Phi'\eta}$ and η_i uncorrelated



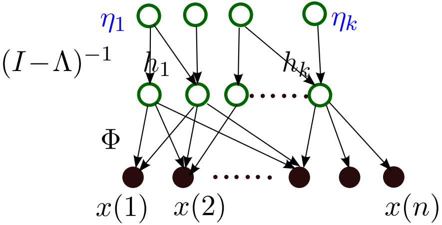
Learning Latent Bayesian Networks

BN: Markov relationships on DAG

Pa_i: parents of node i. $\mathbb{P}(h) = \prod_{i=1}^{n} \mathbb{P}(h_i | h_{\mathrm{Pa}_i})$

Linear Bayesian Network: $h_j = \sum_{i \in \mathrm{Pa}_j} \lambda_{ji} h_i + \eta_j$

$\boxed{h = \Lambda h + \eta}$ $\boxed{\mathbb{E}[x_i|\eta] = \Phi(I-\Lambda)^{-1}\eta = \Phi'\eta}$ and η_i uncorrelated



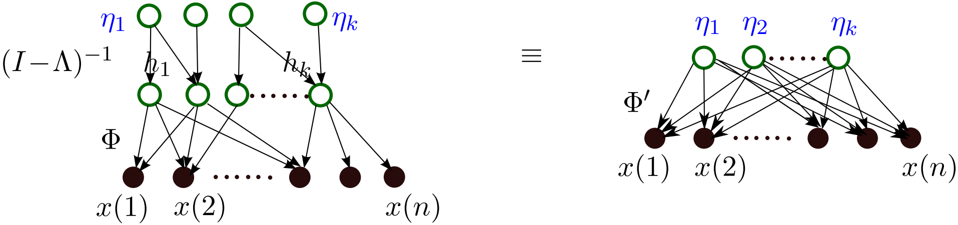
Learning Latent Bayesian Networks

BN: Markov relationships on DAG

Pa_i: parents of node i. $\mathbb{P}(h) = \prod_{i=1}^{n} \mathbb{P}(h_i | h_{\mathrm{Pa}_i})$

Linear Bayesian Network: $h_j = \sum_{i \in \mathrm{Pa}_j} \lambda_{ji} h_i + \eta_j$

$\boxed{h = \Lambda h + \eta}$ $\boxed{\mathbb{E}[x_i|\eta] = \Phi(I-\Lambda)^{-1}\eta = \Phi'\eta}$ and η_i uncorrelated



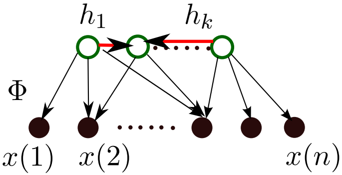
Learning Latent Bayesian Networks

BN: Markov relationships on DAG

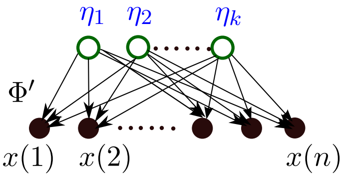
Pa_i: parents of node i. $\mathbb{P}(h) = \prod_{i=1}^{n} \mathbb{P}(h_i | h_{\mathrm{Pa}_i})$

Linear Bayesian Network: $h_j = \sum_{i \in \mathrm{Pa}_j} \lambda_{ji} h_i + \eta_j$

$\boxed{h = \Lambda h + \eta}$ $\boxed{\mathbb{E}[x_i|\eta] = \Phi(I - \Lambda)^{-1}\eta = \Phi'\eta}$ and η_i uncorrelated



$\equiv$



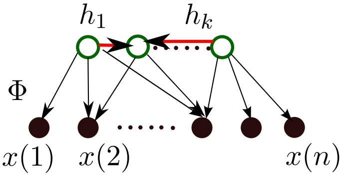
Learning Latent Bayesian Networks

BN: Markov relationships on DAG

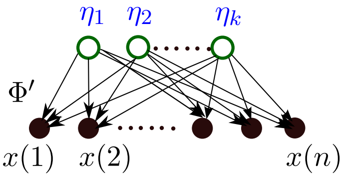
Pa_i: parents of node i. $\mathbb{P}(h) = \prod_{i=1}^{n} \mathbb{P}(h_i | h_{\mathrm{Pa}_i})$

Linear Bayesian Network: $h_j = \sum_{i \in \mathrm{Pa}_j} \lambda_{ji} h_i + \eta_j$

$\boxed{h = \Lambda h + \eta}$ $\boxed{\mathbb{E}[x_i|\eta] = \Phi(I - \Lambda)^{-1}\eta = \Phi'\eta}$ and η_i uncorrelated



$\equiv$



- Φ: structured and sparse while Φ' is dense
- h: correlated topics while η are uncorrelated

Learning Latent Bayesian Networks

$$\boxed{\mathbb{E}[x_i|\eta] = \Phi(I-\Lambda)^{-1}\eta = \Phi'\eta} \quad \boxed{\mathbb{E}[\eta] = \lambda}$$

$$\boxed{\mathbb{E}[x_1 \otimes x_2 \otimes x_3] = \mathbb{E}[\eta^{\otimes 3}](\Phi', \Phi', \Phi') = \sum_i \lambda_i (\phi_i')^{\otimes 3}}$$

Learning Latent Bayesian Networks

$$\boxed{\mathbb{E}[x_i|\eta] = \Phi(I-\Lambda)^{-1}\eta = \Phi'\eta} \quad \boxed{\mathbb{E}[\eta] = \lambda}$$

$$\boxed{\mathbb{E}[x_1 \otimes x_2 \otimes x_3] = \mathbb{E}[\eta^{\otimes 3}](\Phi', \Phi', \Phi') = \sum_i \lambda_i (\phi'_i)^{\otimes 3}}$$

Solving CP decomposition through Tensor Power Method

- Recall η_i are uncorrelated: $\mathbb{E}[\eta^{\otimes}]$ is diagonal.
- Reduction to CP decomposition: can be efficient solved via tensor power method

Learning Latent Bayesian Networks

$$\boxed{\mathbb{E}[x_i|\eta] = \Phi(I-\Lambda)^{-1}\eta = \Phi'\eta} \quad \boxed{\mathbb{E}[\eta] = \lambda}$$

$$\boxed{\mathbb{E}[x_1 \otimes x_2 \otimes x_3] = \mathbb{E}[\eta^{\otimes 3}](\Phi', \Phi', \Phi') = \sum_i \lambda_i (\phi_i')^{\otimes 3}}$$

Solving CP decomposition through Tensor Power Method

- Recall η_i are uncorrelated: $\mathbb{E}[\eta^{\otimes}]$ is diagonal.
- Reduction to CP decomposition: can be efficient solved via tensor power method

Sparse Tucker Decomposition: Unmixing via Convex Optimization

Un-mix Φ from $\Phi' = \Phi(I-\Lambda)^{-1}$ through ℓ_1 optimization.

Learning Latent Bayesian Networks

$$\boxed{\mathbb{E}[x_i|\eta] = \Phi(I-\Lambda)^{-1}\eta = \Phi'\eta} \quad \boxed{\mathbb{E}[\eta] = \lambda}$$

$$\boxed{\mathbb{E}[x_1 \otimes x_2 \otimes x_3] = \mathbb{E}[\eta^{\otimes 3}](\Phi', \Phi', \Phi') = \sum_i \lambda_i (\phi_i')^{\otimes 3}}$$

Solving CP decomposition through Tensor Power Method

- Recall η_i are uncorrelated: $\mathbb{E}[\eta^{\otimes}]$ is diagonal.
- Reduction to CP decomposition: can be efficient solved via tensor power method

Sparse Tucker Decomposition: Unmixing via Convex Optimization

Un-mix Φ from $\Phi' = \Phi(I-\Lambda)^{-1}$ through ℓ_1 optimization.

Learning both structure and parameters of Φ and distribution of h
Combine non-convex and convex methods for learning!

Outline

Outline

Extension to learning overcomplete representations

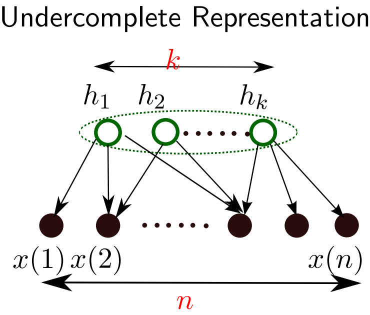
So far..

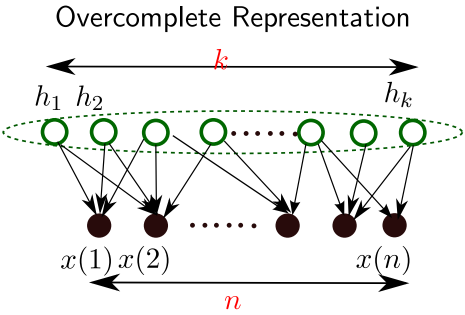
- Pairwise moments for learning structured topic-word matrices
- Third order moments for learning latent Bayesian network models
- Number of topics k, n is vocabulary size and $k < n$.

Extension to learning overcomplete representations

So far..

- Pairwise moments for learning structured topic-word matrices
- Third order moments for learning latent Bayesian network models
- Number of topics k, n is vocabulary size and $k < n$.





What about overcomplete models: $k > n$? Do higher-order moments help?

Learning Overcomplete Representations

Why Overcomplete Representations?

- Flexible modeling, robust to noise
- Huge gains in many applications, e.g. speech and computer vision.

Learning Overcomplete Representations

Why Overcomplete Representations?

- Flexible modeling, robust to noise
- Huge gains in many applications, e.g. speech and computer vision.

Recall Tucker Form of Moments for Topic Models

$$M_2 := \mathbb{E}(x_1 \otimes x_2) = \boxed{\mathbb{E}[h^{\otimes 2}](\Phi, \Phi) \equiv \Phi \mathbb{E}[hh^\top]\Phi^\top}$$

$$M_3 := \mathbb{E}(x_1 \otimes x_2 \otimes x_3) = \boxed{\mathbb{E}[h^{\otimes 3}](\Phi, \Phi, \Phi)}$$

- $k > n$: Tucker decomposition not unique: model non-identifiable.

Learning Overcomplete Representations

Why Overcomplete Representations?

- Flexible modeling, robust to noise
- Huge gains in many applications, e.g. speech and computer vision.

Recall Tucker Form of Moments for Topic Models

$$M_2 := \mathbb{E}(x_1 \otimes x_2) = \boxed{\mathbb{E}[h^{\otimes 2}](\Phi, \Phi) \equiv \Phi \mathbb{E}[hh^\top]\Phi^\top}$$

$$M_3 := \mathbb{E}(x_1 \otimes x_2 \otimes x_3) = \boxed{\mathbb{E}[h^{\otimes 3}](\Phi, \Phi, \Phi)}$$

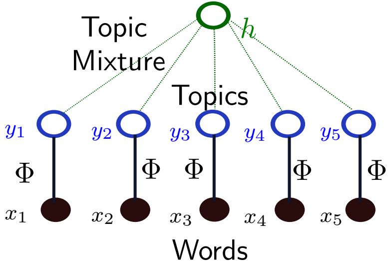
- $k > n$: Tucker decomposition not unique: model non-identifiable.

Identifiability of Overcomplete Models

- Possible under the notion of topic persistence
- Includes single topic model as a special case.

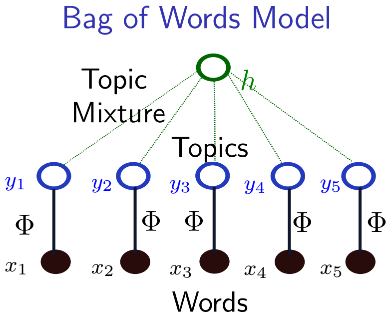
Persistent Topic Models

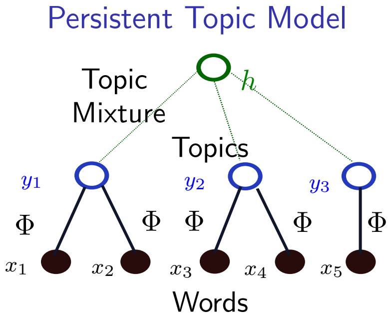
Bag of Words Model



A. Anandkumar, D. Hsu, M. Janzamin, and S. M. Kakade. When are Overcomplete Representations Identifiable? Uniqueness of Tensor Decompositions Under Expansion Constraints, Preprint, June 2013.

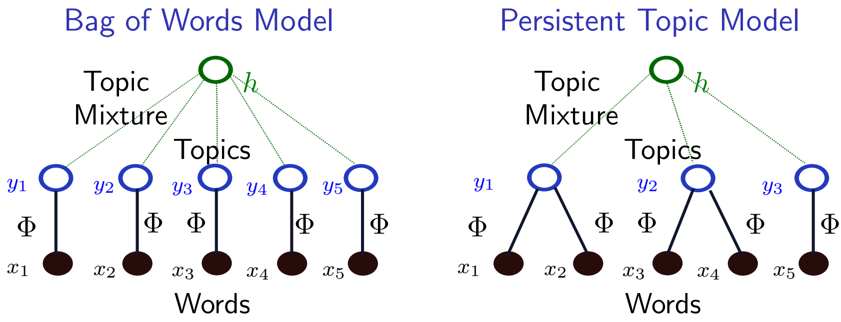
Persistent Topic Models





A. Anandkumar, D. Hsu, M. Janzamin, and S. M. Kakade. When are Overcomplete Representations Identifiable? Uniqueness of Tensor Decompositions Under Expansion Constraints, Preprint, June 2013.

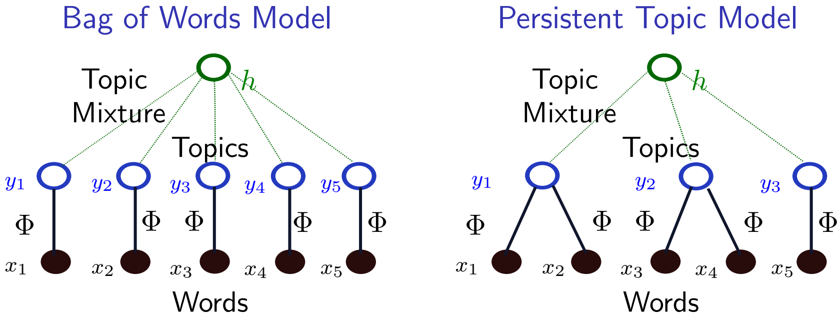
Persistent Topic Models



- Single-topic model is a special case.
- Persistence: incorporates locality or order of words.

A. Anandkumar, D. Hsu, M. Janzamin, and S. M. Kakade. When are Overcomplete Representations Identifiable? Uniqueness of Tensor Decompositions Under Expansion Constraints, Preprint, June 2013.

Persistent Topic Models



- Single-topic model is a special case.
- Persistence: incorporates locality or order of words.

Identifiability conditions for overcomplete models?

A. Anandkumar, D. Hsu, M. Janzamin, and S. M. Kakade. When are Overcomplete Representations Identifiable? Uniqueness of Tensor Decompositions Under Expansion Constraints, Preprint, June 2013.

Identifiability of Overcomplete Models

Recall Tucker Form of Moments for Bag-of-Words Model

- Tensor form: $\mathbb{E}(x_1 \otimes x_2 \otimes x_3 \otimes x_4) = \mathbb{E}[h^{\otimes 4}](\Phi, \Phi, \Phi, \Phi)$
- Matricized form:
  $\mathbb{E}((x_1 \otimes x_2)(x_3 \otimes x_4)^\top) = \boxed{(\Phi \otimes \Phi)\mathbb{E}[(h \otimes h)(h \otimes h)^\top](\Phi \otimes \Phi)^\top}$

Identifiability of Overcomplete Models

Recall Tucker Form of Moments for Bag-of-Words Model

- Tensor form: $\mathbb{E}(x_1 \otimes x_2 \otimes x_3 \otimes x_4) = \mathbb{E}[h^{\otimes 4}](\Phi, \Phi, \Phi, \Phi)$
- Matricized form:
  $\mathbb{E}((x_1 \otimes x_2)(x_3 \otimes x_4)^\top) = \boxed{(\Phi \otimes \Phi)\mathbb{E}[(h \otimes h)(h \otimes h)^\top](\Phi \otimes \Phi)^\top}$

For Persistent Topic Model

- Tensor form: $\mathbb{E}(x_1 \otimes x_2 \otimes x_3 \otimes x_4) = \boxed{\mathbb{E}[hh^\top](\Phi \odot \Phi, \Phi \odot \Phi)}$
- Matricized form:
  $\mathbb{E}((x_1 \otimes x_2)(x_3 \otimes x_4)^\top) = \boxed{(\Phi \odot \Phi)\mathbb{E}[hh^\top](\Phi \odot \Phi)^\top}$

Identifiability of Overcomplete Models

Recall Tucker Form of Moments for Bag-of-Words Model

- Tensor form: $\mathbb{E}(x_1 \otimes x_2 \otimes x_3 \otimes x_4) = \mathbb{E}[h^{\otimes 4}](\Phi, \Phi, \Phi, \Phi)$
- Matricized form:
  $\mathbb{E}((x_1 \otimes x_2)(x_3 \otimes x_4)^\top) = \boxed{(\Phi \otimes \Phi)\mathbb{E}[(h \otimes h)(h \otimes h)^\top](\Phi \otimes \Phi)^\top}$

For Persistent Topic Model

- Tensor form: $\mathbb{E}(x_1 \otimes x_2 \otimes x_3 \otimes x_4) = \boxed{\mathbb{E}[hh^\top](\Phi \odot \Phi, \Phi \odot \Phi)}$
- Matricized form:
  $\mathbb{E}((x_1 \otimes x_2)(x_3 \otimes x_4)^\top) = \boxed{(\Phi \odot \Phi)\mathbb{E}[hh^\top](\Phi \odot \Phi)^\top}$

Kronecker vs. Khatri-Rao Products

- Φ: Topic-word matrix, is $n \times k$.
- $(\Phi \otimes \Phi)$: Kronecker product, is $n^2 \times k^2$ matrix.
- $(\Phi \odot \Phi)$: Khatri-Rao product, is $n^2 \times k$ matrix.

Some Intuitions

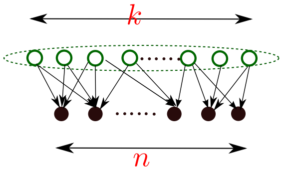
- Bag-of-words Model:

$$\boxed{(\Phi \otimes \Phi)\mathbb{E}[(h \otimes h)(h \otimes h)^\top](\Phi \otimes \Phi)^\top}$$

- Persistent Model:

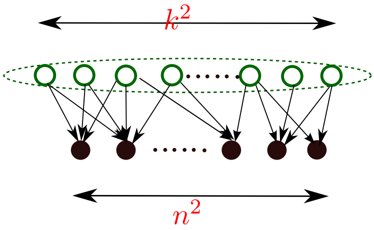
$$\boxed{(\Phi \odot \Phi)\mathbb{E}[hh^\top](\Phi \odot \Phi)^\top}$$

Topic-Word Matrix Φ



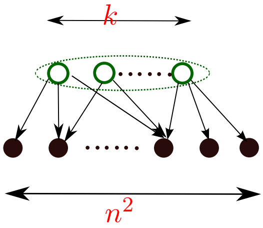
Effective Topic-Word Matrix Given Fourth-Order Moments:

Bag of Words Model:
Kronecker Product $\Phi \otimes \Phi$



Not Identifiable

Persistent Model:
Khatri-Rao Product $\Phi \odot \Phi$



Identifiable

Outline

Conclusion

Moment-based Estimation of Latent Variable Models

- Moments are easy to estimate.
- Low-order moments have good concentration properties

Conclusion

Moment-based Estimation of Latent Variable Models

- Moments are easy to estimate.
- Low-order moments have good concentration properties

Tensor Decomposition Methods

- Moment tensors have tractable forms for many models, e.g. Topic models, HMMs, Gaussian mixtures, ICA.
- Efficient CP tensor decomposition through power iterations.
- Efficient structured Tucker decomposition through ℓ_1.
- Structured topic-word matrices: Expansion conditions
- Can be extended to overcomplete representations

Conclusion

Moment-based Estimation of Latent Variable Models

- Moments are easy to estimate.
- Low-order moments have good concentration properties

Tensor Decomposition Methods

- Moment tensors have tractable forms for many models, e.g. Topic models, HMMs, Gaussian mixtures, ICA.
- Efficient CP tensor decomposition through power iterations.
- Efficient structured Tucker decomposition through ℓ_1.
- Structured topic-word matrices: Expansion conditions
- Can be extended to overcomplete representations

Practical Considerations for Tensor Methods

- Not covered in detail in this tutorial.
- Matrix algebra and iterative methods.
- Scalable: Parallel implementation on GPUs