

# Spectral Methods: Latent Variable Models

---

FOROUGH ARABSHAHI

PGM  
Spring 2018



# Topic Model: Mixture of Unigrams (Cont)

Each document has exactly one single topic

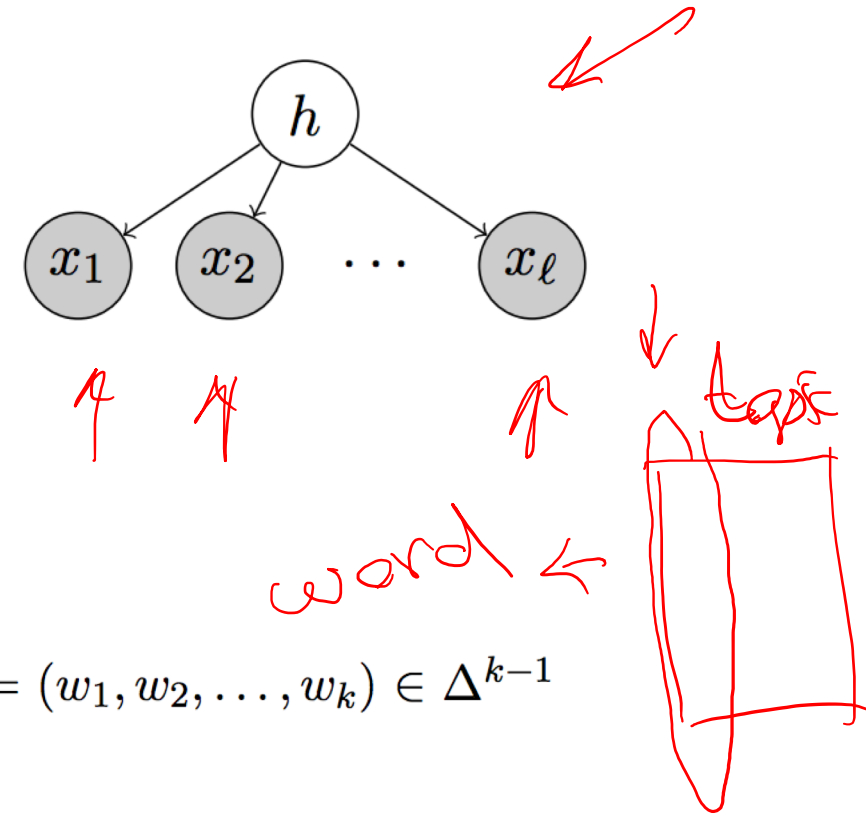
Exchangeability: Joint distribution invariant under permutation

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
- 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

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

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

$$A \in \bigotimes^3 \mathbb{R}^n$$

P-way tensor:

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

$P = 1$  : Vector

$P = 2$  : Matrix

Alternate notation:

P-th tensor power:

$$A \in \bigotimes^p \mathbb{R}^n$$

$$\underline{v}^{\otimes p} := v \otimes v \otimes \dots \otimes v \in \bigotimes^p \mathbb{R}^n$$

P-th tensor of rank  $k$ :

$$A = \sum_{j=1}^k \underline{u_{1,j}} \otimes \underline{u_{2,j}} \otimes \dots \otimes \underline{u_{p,j}}$$

P-th symmetric tensor:

$$A = \sum_{j=1}^k \underline{u_j^{\otimes p}}$$

Vector product:

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

$$T = v_1^{\otimes 3} + v_2^{\otimes 3} + \dots$$

$$u \otimes v = uv^T$$



$$x_i = e_i$$

# Mixture of Unigrams: properties

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

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

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

*Handwritten notes:*  $\mathbb{E}[x_1, x_2^T]$  with an arrow pointing to the first equation. A bracket on the right side of the equations is labeled  $h=j$ .

Conditional mean:

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

*Handwritten notes:* A bracket on the right side of the equation is labeled  $h=j$ .

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.)

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

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

Proof:

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

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

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

$$= \sum_{i=1}^k w_i \mu_i \otimes \mu_i$$

$$\mathbb{E}[x_1 \otimes x_2]$$

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

$$\begin{bmatrix} 0 \end{bmatrix}$$

$$\hookrightarrow \Pr[h=i]$$

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

$$\sum \underbrace{\Pr[h=j]}_{w_j} \underbrace{\mathbb{E}[x_1 | h=j]}_{\mu_j} \otimes \mathbb{E}[x_2 | h=j]$$

# Spectral Methods: Mixture of Unigrams

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

$E[x_1 \otimes x_2]$

$$M_2 = \lambda_1 a_1 a_1^T + \lambda_2 a_2 a_2^T$$

Non-convex optimization, but guaranteed global solution!

Fast and scalable

Extensive linear algebra support

But ...

$$M = L R \leftarrow$$

$$\textcircled{M} = \textcircled{L} \textcircled{S} \textcircled{S^{-1}} \textcircled{R} = \textcircled{LS} \times \textcircled{S^{-1}R}$$

# Matrix Decomposition I

Matrix decomposition is not unique in general :(

Matrix decomposition is only unique for orthogonal factors

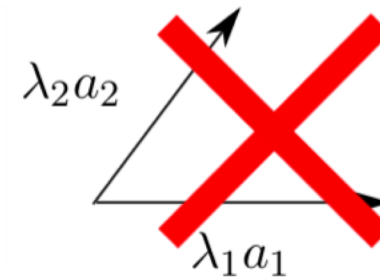
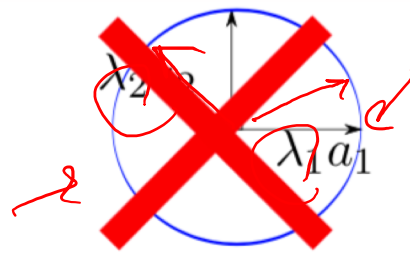
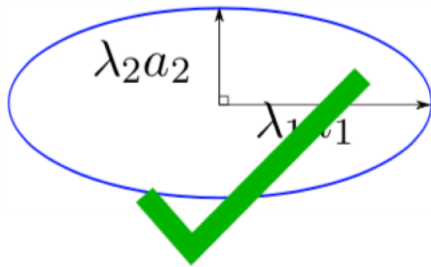
Guaranteed recovery needs the separability assumption:  $\lambda_1 > \lambda_2 > \dots$

$$\lambda_1 > \lambda_2 > \lambda_3 > \dots$$

$$\geq 0$$

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]

$$M_3 := \mathbb{E}[x_1 \otimes x_2 \otimes x_3]$$

$$M_3 = \sum_{i=1}^k \underbrace{w_i \mu_i \otimes \mu_i \otimes \mu_i}_{\text{red underline}}$$

Proof: similar to  $M_2$

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

$$\begin{aligned} \mathbb{E}[x_1 \otimes x_2 \otimes x_3] &= \mathbb{E}[\mathbb{E}[x_1 \otimes x_2 \otimes x_3 | h]] \\ &= \mathbb{E}[\mathbb{E}[x_1 | h] \otimes \mathbb{E}[x_2 | h] \otimes \mathbb{E}[x_3 | h]] \\ &= \sum_{i=1}^k \Pr[h = i] \mathbb{E}[x_1 | h = i] \otimes \mathbb{E}[x_2 | h = i] \otimes \mathbb{E}[x_3 | h = i] \\ &= \sum_{i=1}^k w_i \mu_i \otimes \mu_i \otimes \mu_i \end{aligned}$$

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

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

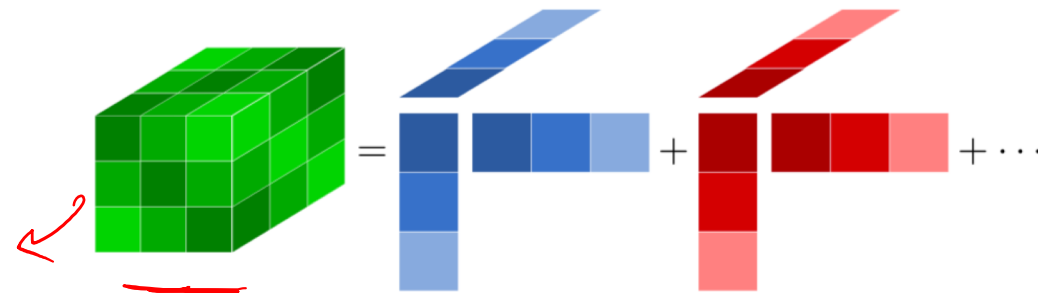
$$\mathbb{E}[x_1 | h] \otimes \mathbb{E}[x_2 | h] \otimes \mathbb{E}[x_3 | h]$$

$$\sum \Pr[h=i] \cdot \mu_i \otimes \mu_i \otimes \mu_i$$

# Spectral Methods: Tensor Decomposition

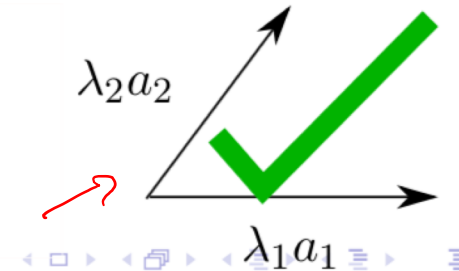
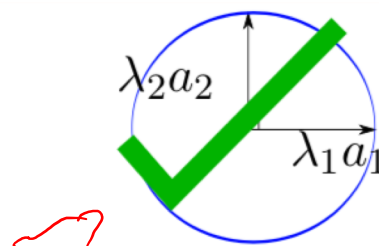
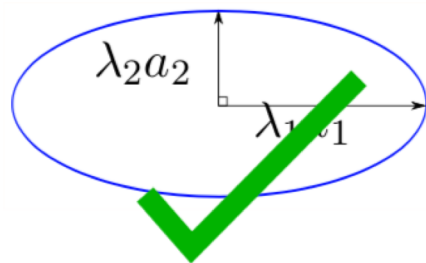
Training: Estimate  $M_3$  from the data and decompose into rank-1 components

$d \times d \times d$   
 $k \times k \times k$



$$T = v_1^{\otimes 3} + v_2^{\otimes 3} + \dots,$$

The decomposition is unique for **linearly independent** factors



# Training Algorithm

$(W) = U \Lambda^{-1/2}$   
 eigenvectors & eigenvalues

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

$[x_1 \otimes x_2]$

$M_2(W, W) = W^T M_2 W = I$  whitening

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

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

$\tilde{\mu}_i = \sqrt{w_i} \mu_i^T W$

$(W)$

$\tilde{M}_3 = M_3(W, W, W)$

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



# 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

0.3  
0.3  
0.2



SECTIONS HOME SEARCH The New York Times

COLLEGE FOOTBALL

## At Florida State, Football Clouds Justice

Now, an examination by The New York Times of police and court records, along with interviews with crime witnesses, has found that, far from an aberration, the treatment of the Winston complaint was in keeping with the way the police on numerous occasions have soft-pedaled allegations of wrongdoing by Seminoles football players. From criminal mischief and motor-vehicle theft to domestic violence, arrests have been avoided, investigations have stalled and players have escaped serious consequences.

In a community whose self-image and economic well-being are so tightly bound to the fortunes of the nation's top-ranked college football team, law enforcement officers are finely attuned to a suspect's football connections. Those ties are cited repeatedly in police reports examined by The Times. What's more, dozens of officers work second jobs directing traffic and providing security at home football games, and many express their devotion to the Seminoles on social media.

On Jan. 10, 2013, a female student at Florida State spotted the man she believed had raped her the previous month. After learning his name, Jameis Winston, she reported him to the Tallahassee police.

In the 21 months since, Florida State officials have said little about how they handled the case, which is no As The Times reported last April, the Tallahassee police also failed to aggressively investigate the rape accusation. It did not become public until November, when a Tampa reporter, Matt Baker, acting on a tip, sought records of the police investigation.

Upon learning of Mr. Baker's inquiry, Florida State, having shown little curiosity about the rape accusation, suddenly took a keen interest in the journalist seeking to report it, according to emails obtained by The Times.

"Can you share any details on the requesting source?" David Perry, the university's police chief, asked the Tallahassee police. Several hours later, Mr.

Topics

Education



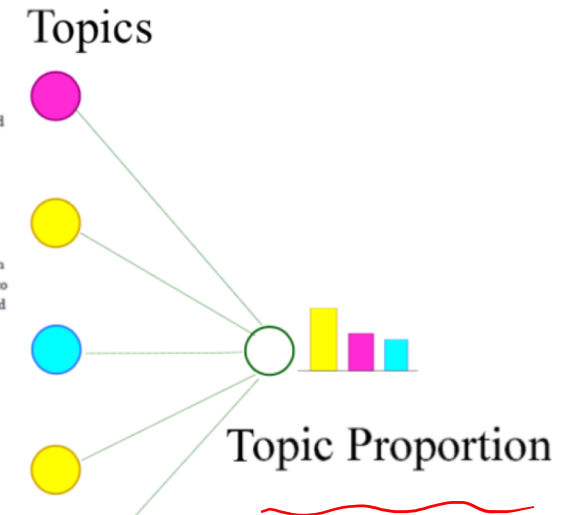
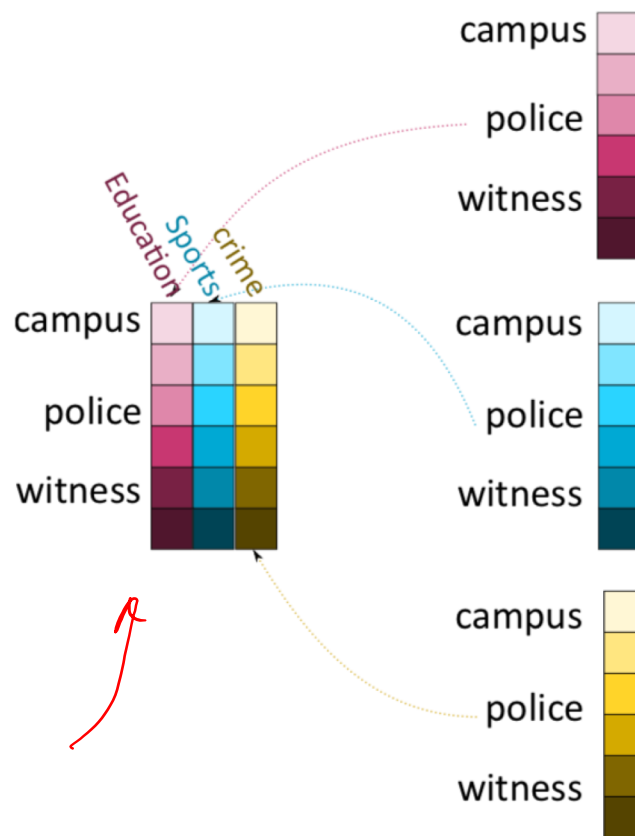
Crime



Sports

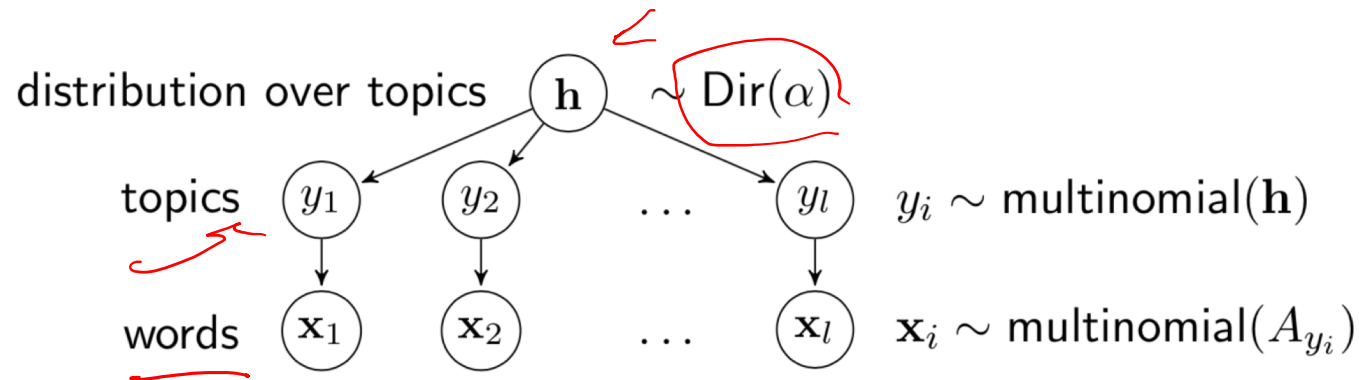


# Latent Dirichlet Allocation: LDA



# LDA: an Admixture Model

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



$$\mathbb{E}[x_i|h] = Oh$$

Topic distribution:

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

$$\alpha \propto (\alpha_1, \alpha_2, \dots, \alpha_k)$$

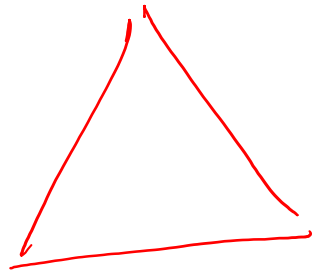
$$\alpha_0 := \alpha_1 + \alpha_2 + \dots + \alpha_k.$$

$$\underline{E[h \otimes h] + \eta E[h] \otimes E[h] = I}$$

$$\eta = \frac{\alpha_0}{\alpha_0 + 1}$$

## Spectral Methods: LDA (Cont.)

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



$$M_1 := E[x_1]$$

$$M_2 := E[x_1 \otimes x_2] - \frac{\alpha_0}{\alpha_0 + 1} M_1 \otimes M_1$$

$$M_2 = \sum_{i=1}^k \frac{\alpha_i}{(\alpha_0 + 1)\alpha_0} \mu_i \otimes \mu_i$$

$$E[E[h \otimes h] + \eta E[h] \otimes E[h]] = \sum \alpha_i \mu_i \mu_i^T + \eta \mu \mu^T = M_2$$

Raw moments are not directly diagonalizable.

Trick: Find the right off diagonals that compensate for it

$$\mu_1 = \mathbb{E}[x_1] = \mathbb{E}[\underbrace{\mathbb{E}[x_1|h]}_{\text{topic word} \rightarrow Oh}] = O \underbrace{\mathbb{E}[h]}_{\gamma}$$

## Spectral Methods: LDA (Cont.)

Proof:

$$\mathbb{E}[x_1|h] = Oh$$

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

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

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

$$\begin{aligned} M_2 &= O\mathbb{E}[h \otimes h]O^T - \frac{\alpha_0}{\alpha_0 + 1} O\mathbb{E}[h] \otimes \mathbb{E}[h]O^T \\ &= O \left[ \frac{1}{(\alpha_0 + 1)\alpha_0} (\text{diag}(\alpha) + \alpha \otimes \alpha) - \frac{1}{(\alpha_0 + 1)\alpha_0} \alpha \otimes \alpha \right] O^T \\ &= \sum \frac{\alpha_i}{(\alpha_0 + 1)\alpha_0} O_i \otimes O_i \end{aligned}$$

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

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

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

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

$$M_2 = \mathbb{E}[x_1 \otimes x_2] - \frac{\alpha_0}{\alpha_0 + 1} \mathbb{E}[h] \otimes \mathbb{E}[h]$$

$$= \sum \bigcirc \mu_i \otimes \mu_i$$

$$O\mathbb{E}[h \otimes h]O^T$$

# Spectral Methods: LDA (Cont.)

Theorem: [Anandkumar et al., 2012a]

$$M_3 := \mathbb{E}[x_1 \otimes x_2 \otimes x_3]$$

$$= \frac{\alpha_0}{\alpha_0 + 2} \left( \mathbb{E}[x_1 \otimes x_2 \otimes M_1] + \mathbb{E}[x_1 \otimes M_1 \otimes x_2] + \mathbb{E}[M_1 \otimes x_1 \otimes x_2] \right) + \frac{2\alpha_0^2}{(\alpha_0 + 2)(\alpha_0 + 1)} M_1 \otimes M_1 \otimes M_1.$$

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

Raw moments are not directly diagonalizable.

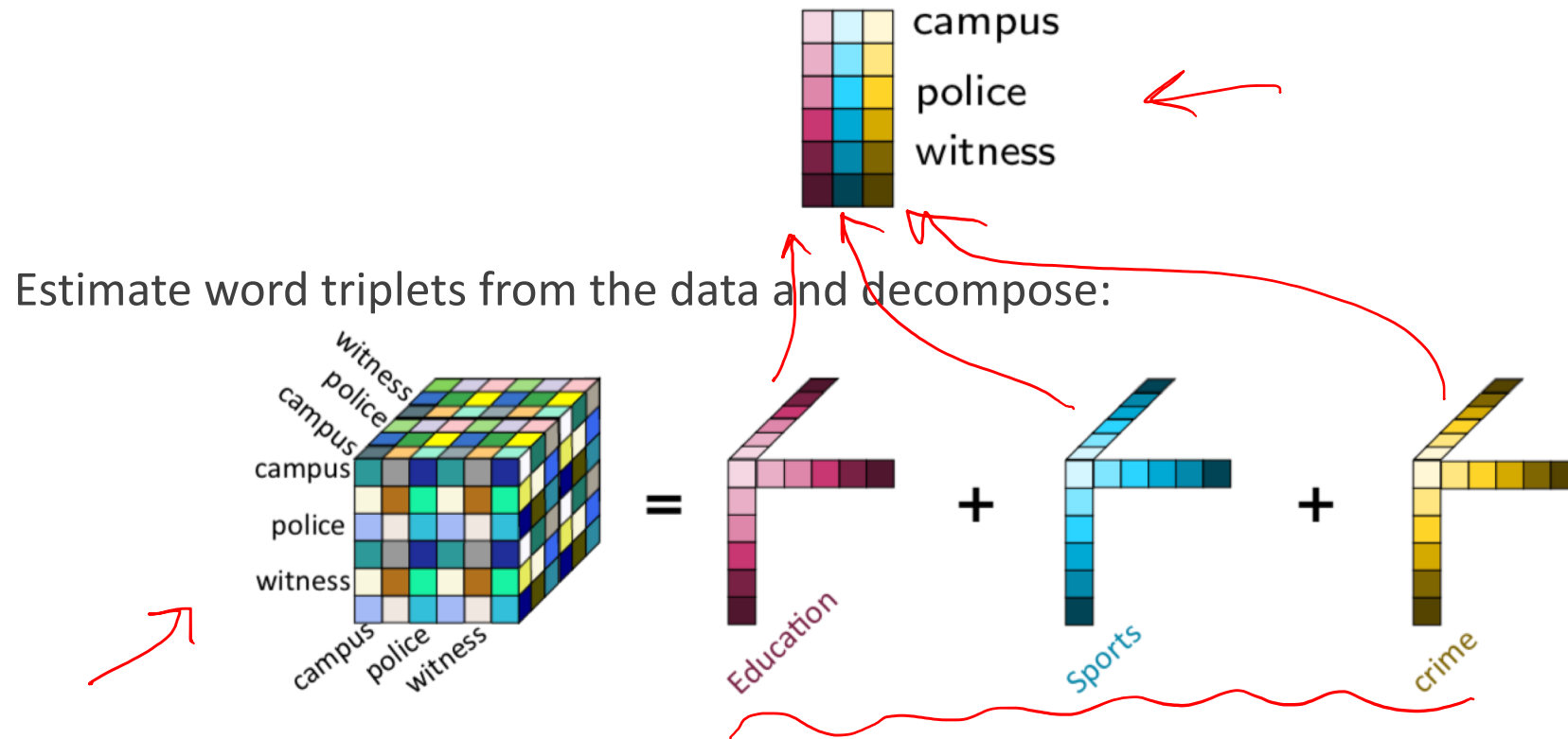
Trick: Find the right off diagonals that compensate for it

$$\mathbb{E}[h \otimes h \otimes h] + \eta_1 \left[ \mathbb{E}[h \otimes h] \otimes \mathbb{E}[h] + \mathbb{E}[h] \otimes \mathbb{E}[h \otimes h] + \mathbb{E}[h \otimes h] \otimes \mathbb{E}[h] \right]$$

$$+ \eta_2 \left[ \mathbb{E}[h \otimes h \otimes h] + \mathbb{E}[h \otimes h] \otimes \mathbb{E}[h] + \mathbb{E}[h] \otimes \mathbb{E}[h \otimes h] + \mathbb{E}[h \otimes h] \otimes \mathbb{E}[h] \right]$$

# Spectral Methods: Training LDA

Goal: recover the topic word matrix:



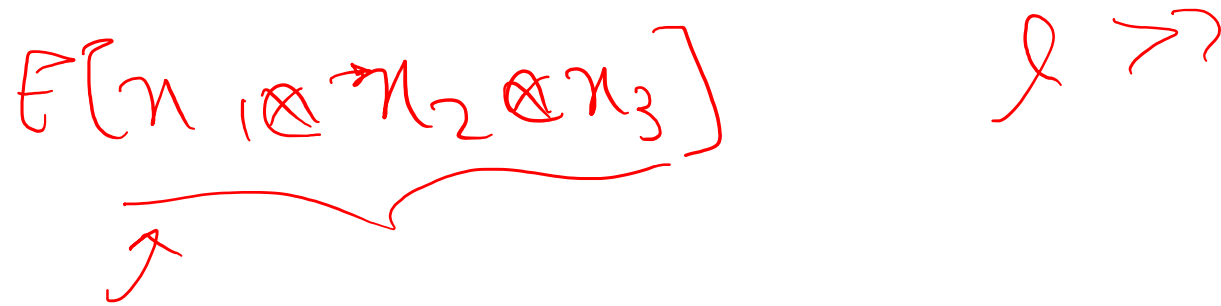
# Practical Notes (Cont.)

---

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

Should we average over all  $\binom{\ell}{3}$ . 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

$$F[n_1 \otimes n_2 \otimes n_3] \quad \Rightarrow$$




# Practical Notes (Cont.)

$$c \in \mathbb{R}^d$$

Let  $c \in \mathbb{R}^d$  be the word count vector for the document, then its contribution to the moment is:

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

Which is equal to:

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

# Practical Notes

The same can be observed for the second moment.

$$E[x_1 \otimes x_2]$$

$$\frac{1}{\binom{\ell}{2}} \cdot \frac{1}{2!} \cdot (c \otimes c - \text{diag}(c))$$

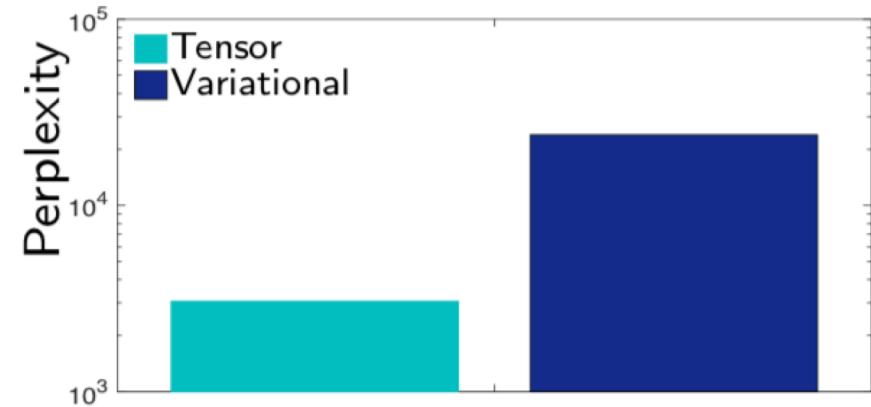
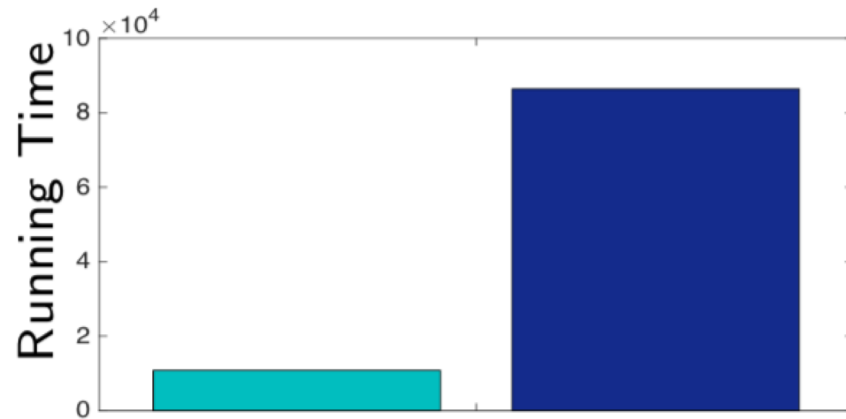
Why?

$$\begin{aligned} \mathbb{E}[c_n(i)^2 - c_n(i)] &= \mathbb{E}\left[\left(\sum_{p=1}^{\ell_n} x_{n,p}(i)\right)^2 - \sum_{p=1}^{\ell_n} x_{n,p}(i)\right] \\ &= \mathbb{E}\left[\sum_{p=1}^{\ell_n} x_{n,p}(i)^2 + 2 \sum_{p < q} x_{n,p}(i) x_{n,q}(i) - \sum_{p=1}^{\ell_n} x_{n,p}(i)\right] \\ &= 2 \sum_{p < q} \mathbb{E}[x_{n,p}(i) x_{n,q}(i)] \quad (\text{since } x_{n,p}(i)^2 = x_{n,p}(i)) \\ &= \ell_n(\ell_n - 1) [M_2^f]_{i,i} \rightarrow E[x_1 \otimes x_2] \end{aligned}$$

# 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

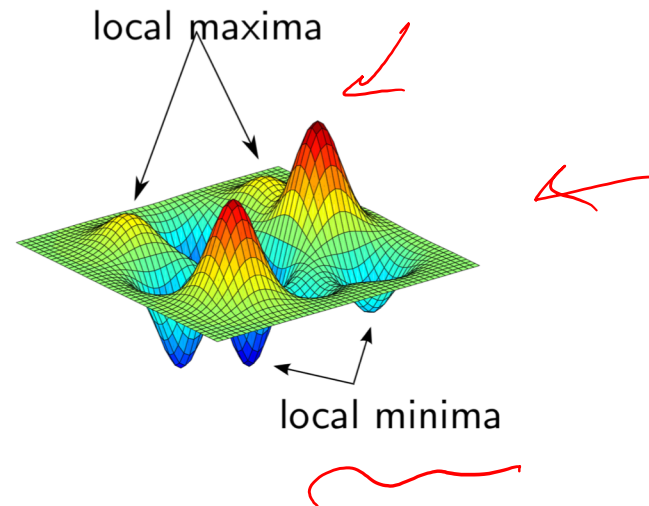
Not known if it can be derived for arbitrary models

# 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)