Topic Models and Coordinate Ascent Variational Inference

STATS 305C: Applied Statistics

Scott Linderman

May 4, 2023

Mixed Membership Models

Mixed membership models are designed for *grouped data*.

Each "data point" is itself a collection of observations. For example,

- ▶ in text analysis, a document is a collection of observed words.

- ▶ in social science, a survey is a collection of observed answers.

- ▶ in genetic sequencing, a genome is a collection of observed genes.

Mixed membership models look for patterns like the components of a mixture model, but allowing each data point to involve multiple components.

Notation for a mixed membership model

▶ Let $\boldsymbol{x}_n = (x_{n,1}, \ldots, x_{n,D})$ denote n-th **data point**. It contains a **collection of observations**. To simplify notation, assume all data points are length D.

▶ Each data point reflects a **combination of** K **mixture components** with parameters $\{\boldsymbol{\theta}_k\}_{k=1}^K$. These are shared by all documents.

▶ Each data point has its own **mixture proportions**, $\pi_n \in \Delta_K$.

▶ Each **observation is assigned to one of the components** by $z_{n,d} \in \{1, \ldots, K\}$.

Topic models

The most common mixed-membership model is the **topic model**, a generative model for documents.
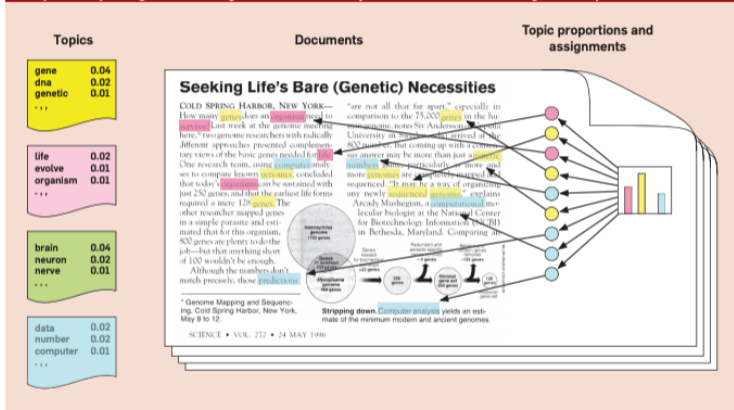
Topic models are so common, they have their own nomenclature:

data set	corpus	collection of documents
data point	document	collection of words
observation	word	one element of a document
mixture component	topic	distribution over words
mixture proportions	topic proportions	distribution over topics
mixture assignment	topic assignment	which topic produced a word

Table: Rosetta stone for translating between mixed membership model and topic model notation.

Topic modeling intuition

From Blei [2012].

Topic modeling of New York Times articles

❶	❷	❸	❹	❺
Game	Life	Film	Book	Wine
Season	Know	Movie	Life	Street
Team	School	Show	Books	Hotel
Coach	Street	Life	Novel	House
Play	Man	Television	Story	Room
Points	Family	Films	Man	Night
Games	Says	Director	Author	Place
Giants	House	Man	House	Restaurant
Second	Children	Story	War	Park
Players	Night	Says	Children	Garden

❻	❼	❽	❾	❿
Bush	Building	Won	Yankees	Government
Campaign	Street	Team	Game	War
Clinton	Square	Second	Mets	Military
Republican	Housing	Race	Season	Officials
House	House	Round	Run	Iraq
Party	Buildings	Cup	League	Forces
Democratic	Development	Open	Baseball	Iraqi
Political	Space	Game	Team	Army
Democrats	Percent	Play	Games	Troops
Senator	Real	Win	Hit	Soldiers

⓫	⓬	⓭	⓮	⓯
Children	Stock	Church	Art	Police
School	Percent	War	Museum	Yesterday
Women	Companies	Women	Show	Man
Family	Fund	Life	Gallery	Officer
Parents	Market	Black	Works	Officers
Child	Bank	Political	Artists	Case
Life	Investors	Catholic	Street	Found
Says	Funds	Government	Artist	Charged
Help	Financial	Jewish	Paintings	Street
Mother	Business	Pope	Exhibition	Shot

Figure 6.1: Posterior topics from the *The New York Times*.

Dynamic topic modeling of Science articles
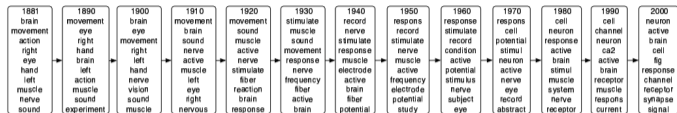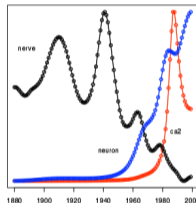
Figure 4. Examples from the posterior analysis of a 20-topic dynamic model estimated from the *Science* corpus. For two topics, we illustrate: (a) the top ten words from the inferred posterior distribution at ten year lags (b) the posterior estimate of the frequency as a function of year of several words from the same two topics (c) example articles throughout the collection which exhibit these topics. Note that the plots are scaled to give an idea of the shape of the trajectory of the words' posterior probability (i.e., comparisons across words are not meaningful).

From Blei and Lafferty [2006].

Topic modeling of congressional voting records
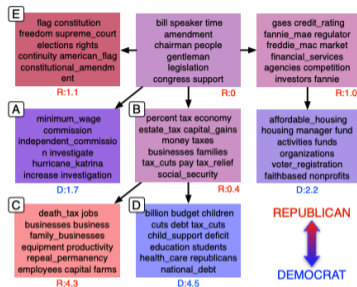
Figure 7.2: Topics discovered from Congressional floor debates using a downstream model to capture speaker's ideology. Many first-level topics are bipartisan (purple), while lower level topics are associated with specific ideologies (Democrats blue, Republicans red). For example, the "tax" topic (B) is bipartisan, but its Democratic-leaning child (D) focuses on social goals supported by taxes ("children", "education", "health care"), while its Republican-leaning child (C) focuses on business implications ("death tax", "jobs", "businesses"). The number below each topic denotes the magnitude of a learned regression parameter associated with that topic. Colors and the numbers beneath each topic show the regression parameter η associated with the topic. From Nguyen et al. [2013].

The generative process for a mixed membership model

The generative model process is:

- ▶ for each mixture component $k = 1, \ldots, K$, sample its parameter $\theta_k \sim p(\theta_k \mid \phi)$
- ▶ for each data point $n = 1, \ldots, N$:
    - ▶ sample mixture proportions $\pi_n \sim \text{Dir}(\pi_n \mid \alpha)$
    - ▶ for each observation $d = 1, \ldots, D$:
        - ▶ sample mixture assignment $z_{n,d} \in \{1, \ldots, K\}$ from a cateogrical distribution $z_{n,d} \sim \text{Cat}(\pi_n)$
        - ▶ sample observation $x_{n,d} \sim p(x \mid \theta_{z_{n,d}})$

The mixed membership model allows sharing at the dataset level (all data points share the same components) while allowing variability at the data point level (each data point has its own mixture proportions).

The generative process for a topic model

Slide 9 in the language of topic modeling:

▶ for each **topic** $k = 1, \ldots, K$, sample its parameter $\theta_k \sim p(\theta_k \mid \phi)$

▶ for each **document** $n = 1, \ldots, N$:

    ▶ sample **topic proportions** $\pi_n \sim \mathrm{Dir}(\pi_n \mid \alpha)$

    ▶ for each **word** $d = 1, \ldots, D$:

        ▶ sample **topic assignment** $z_{n,d} \in \{1, \ldots, K\}$ from a categorical distribution $z_{n,d} \sim \mathrm{Cat}(\pi_n)$

        ▶ sample **word** $x_{n,d} \in \{1, \ldots, V\}$ from a categorical distribution, $x_{n,d} \sim \mathrm{Cat}(\theta_{z_{n,d}})$

The topic model captures sharing at the corpus level (all documents share the same topics) while allowing variability at the data point level (each document weights topics differently).

The joint distribution

The joint probability for a general mixed membership model is,

$$p(\{\boldsymbol{\theta}_k\}_{k=1}^K, \{\boldsymbol{\pi}_n, \mathbf{z}_n, \mathbf{x}_n\}_{n=1}^N \mid \boldsymbol{\phi}, \boldsymbol{\alpha}) = \prod_{k=1}^K p(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}) \prod_{n=1}^N \left[p(\boldsymbol{\pi}_n \mid \boldsymbol{\alpha}) \prod_{d=1}^D p(z_{n,d} \mid \boldsymbol{\pi}_n) p(x_{n,d} \mid \boldsymbol{\theta}_{z_{n,d}}) \right] \tag{1}$$

As in mixture models, we can write this equivalently as

$$p(\{\boldsymbol{\theta}_k\}_{k=1}^K, \{\boldsymbol{\pi}_n, \mathbf{z}_n, \mathbf{x}_n\}_{n=1}^N \mid \boldsymbol{\phi}, \boldsymbol{\alpha}) = \prod_{k=1}^K p(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}) \prod_{n=1}^N \left[p(\boldsymbol{\pi}_n \mid \boldsymbol{\alpha}) \prod_{d=1}^D \prod_{k=1}^K \left[\pi_{n,k} p(x_{n,d} \mid \boldsymbol{\theta}_k) \right]^{\mathbb{I}[z_{n,d}=k]} \right] \tag{2}$$

Graphical Model

Exercise: Draw the graphical models for a mixture model and a mixed membership model.

Latent Dirichlet allocation

Latent Dirichlet Allocation (LDA) [Blei et al., 2003] is the most widely used topic model.

It assumes conjugate Dirichlet-Categorical model for the topics $\theta_k \in \Delta_V$ and words $x_{n,d} \in \{1, \ldots, V\}$,

$$\theta_k \overset{\text{iid}}{\sim} \text{Dir}(\phi), \tag{3}$$

$$\pi_n \overset{\text{iid}}{\sim} \text{Dir}(\alpha), \tag{4}$$

$$z_{n,d} \overset{\text{iid}}{\sim} \text{Cat}(\pi_n) \tag{5}$$

$$x_{n,d} \overset{\text{iid}}{\sim} \text{Cat}(\theta_{z_{n,d}}) \tag{6}$$

Latent Dirichlet allocation II

Plugging in these assumptions, the joint probability is,

$$p(\{\boldsymbol{\theta}_k\}_{k=1}^{K}, \{\boldsymbol{\pi}_n, \mathbf{z}_n, \mathbf{x}_n\}_{n=1}^{N} \mid \boldsymbol{\phi}, \boldsymbol{\alpha}) \tag{7}$$

$$= \prod_{k=1}^{K} \mathrm{Dir}(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}) \prod_{n=1}^{N} \left[\mathrm{Dir}(\boldsymbol{\pi}_n \mid \boldsymbol{\alpha}) \prod_{d=1}^{D} \pi_{n,z_{n,d}} \theta_{z_{n,d}, x_{n,d}} \right] \tag{8}$$

$$= \prod_{k=1}^{K} \mathrm{Dir}(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}) \prod_{n=1}^{N} \left(\mathrm{Dir}(\boldsymbol{\pi}_n \mid \boldsymbol{\alpha}) \prod_{d=1}^{D} \left[\prod_{k=1}^{K} \pi_{n,k}^{\mathbb{I}[z_{n,d}=k]} \right] \left[\prod_{k=1}^{K} \prod_{v=1}^{V} \theta_{k,v}^{\mathbb{I}[x_{n,d}=v]\mathbb{I}[z_{n,d}=k]} \right] \right)$$

$$\propto \left[\prod_{k=1}^{K} \prod_{v=1}^{V} \theta_{k,v}^{\phi_v-1} \right] \prod_{n=1}^{N} \left(\left[\prod_{k=1}^{K} \pi_{n,k}^{\alpha_k+N_{n,\cdot,k}-1} \right] \left[\prod_{k=1}^{K} \prod_{v=1}^{V} \theta_{k,v}^{N_{n,v,k}} \right] \right) \tag{9}$$

where

▶ $N_{n,v,k} = \sum_{d=1}^{D} \mathbb{I}[x_{n,d}=v]\mathbb{I}[z_{n,d}=k]$ is the number of instances of word v in document n assigned to topic k, and,

▶ $N_{n,\cdot,k} = \sum_{v=1}^{V} N_{n,v,k} = \sum_{d=1}^{D} \mathbb{I}[z_{n,d}=k]$ is the number of words in document n assigned to topic k.

Gibbs sampling for LDA

As usual, sample each variable from its conditional distribution, holding the rest fixed.

▶ **Topic assignments:** The assignments are conditionally independent given the topic parameters and proportions,

$$p(z_{n,d} = k \mid x_{n,d} = v, \{\boldsymbol{\theta}_k\}_{k=1}^{K}, \boldsymbol{\pi}_n) \propto \pi_{n,k}\theta_{k,v} \tag{10}$$

▶ **Topic proportions:** Let $N_{n,k} = \sum_{d=1}^{D} \mathbb{I}[z_{n,d} = k]$ denote the number of words in document n assigned to topic k. Then,

$$p(\boldsymbol{\pi}_n \mid \boldsymbol{\alpha}, \mathbf{z}_n) \propto \mathrm{Dir}(\boldsymbol{\pi}_n \mid \boldsymbol{\alpha}) \prod_{k=1}^{K} \pi_{n,k}^{N_{n,\cdot,k}} = \mathrm{Dir}([\alpha_1 + N_{n,\cdot,1}, \ldots, \alpha_K + N_{n,\cdot,K}]). \tag{11}$$

▶ **Topic parameters:** Let $N_{n,k,v} = \sum_{d=1}^{D} \mathbb{I}[z_{n,d} = k]\mathbb{I}[x_{n,d} = v]$ denote the number of times word v was assigned to topic k in document n. Then,

$$p(\boldsymbol{\theta}_k \mid \{\mathbf{z}_n, \mathbf{x}_n\}, \boldsymbol{\phi}) \propto \mathrm{Dir}(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}) \prod_{n=1}^{N} \prod_{v=1}^{V} \theta_{k,v}^{N_{n,v,k}} = \mathrm{Dir}([\phi_1 + \sum_{n=1}^{N} N_{n,v,k}, \ldots, \phi_V + \sum_{n=1}^{N} N_{n,V,k}]).$$

Gibbs sampling for LDA

As usual, sample each variable from its conditional distribution, holding the rest fixed.

▶ **Topic assignments:** The assignments are conditionally independent given the topic parameters and proportions,

$$p(z_{n,d} = k \mid x_{n,d} = v, \{\boldsymbol{\theta}_k\}_{k=1}^K, \boldsymbol{\pi}_n) \propto \pi_{n,k} \theta_{k,v} \tag{10}$$

▶ **Topic proportions:** Let $N_{n,k} = \sum_{d=1}^D \mathbb{I}[z_{n,d} = k]$ denote the number of words in document n assigned to topic k. Then,

$$p(\boldsymbol{\pi}_n \mid \boldsymbol{\alpha}, \boldsymbol{z}_n) \propto \mathrm{Dir}(\boldsymbol{\pi}_n \mid \boldsymbol{\alpha}) \prod_{k=1}^K \pi_{n,k}^{N_{n,\cdot,k}} = \mathrm{Dir}([\alpha_1 + N_{n,\cdot,1}, \dots, \alpha_K + N_{n,\cdot,K}]). \tag{11}$$

▶ **Topic parameters:** Let $N_{n,k,v} = \sum_{d=1}^D \mathbb{I}[z_{n,d} = k]\mathbb{I}[x_{n,d} = v]$ denote the number of times word v was assigned to topic k in document n. Then,

$$p(\boldsymbol{\theta}_k \mid \{\boldsymbol{z}_n, \boldsymbol{x}_n\}, \boldsymbol{\phi}) \propto \mathrm{Dir}(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}) \prod_{n=1}^N \prod_{v=1}^V \theta_{k,v}^{N_{n,v,k}} = \mathrm{Dir}([\phi_1 + \sum_{n=1}^N N_{n,v,k}, \dots, \phi_V + \sum_{n=1}^N N_{n,V,k}]).$$

Gibbs sampling for LDA

As usual, sample each variable from its conditional distribution, holding the rest fixed.

▶ **Topic assignments:** The assignments are conditionally independent given the topic parameters and proportions,

$$p(z_{n,d} = k \mid x_{n,d} = v, \{\boldsymbol{\theta}_k\}_{k=1}^K, \pi_n) \propto \pi_{n,k}\theta_{k,v} \tag{10}$$

▶ **Topic proportions:** Let $N_{n,k} = \sum_{d=1}^D \mathbb{I}[z_{n,d} = k]$ denote the number of words in document n assigned to topic k. Then,

$$p(\pi_n \mid \boldsymbol{\alpha}, \boldsymbol{z}_n) \propto \mathrm{Dir}(\pi_n \mid \boldsymbol{\alpha}) \prod_{k=1}^K \pi_{n,k}^{N_{n,\cdot,k}} = \mathrm{Dir}([\alpha_1 + N_{n,\cdot,1}, \ldots, \alpha_K + N_{n,\cdot,K}]). \tag{11}$$

▶ **Topic parameters:** Let $N_{n,k,v} = \sum_{d=1}^D \mathbb{I}[z_{n,d} = k]\mathbb{I}[x_{n,d} = v]$ denote the number of times word v was assigned to topic k in document n. Then,

$$p(\boldsymbol{\theta}_k \mid \{\boldsymbol{z}_n, \boldsymbol{x}_n\}, \boldsymbol{\phi}) \propto \mathrm{Dir}(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}) \prod_{n=1}^N \prod_{v=1}^V \theta_{k,v}^{N_{n,v,k}} = \mathrm{Dir}([\phi_1 + \sum_{n=1}^N N_{n,v,k}, \ldots, \phi_V + \sum_{n=1}^N N_{n,V,k}]).$$

Why does LDA produce sharp topics?

Consider the log posterior as a function of θ_k and π_n. From eq. 8, it is,

$$\sum_{k=1}^{K} \log p(\theta_k \mid \phi) + \sum_{n=1}^{N} \log p(\pi_n \mid \alpha) + \sum_{n=1}^{N} \sum_{d=1}^{D} \Big(\log \pi_{n,z_{n,d}} + \log \theta_{z_{n,d},x_{n,d}} \Big) \tag{13}$$

The double sum over *n* and *d* dominates this expression.

It encourages topic proportions and topic probabilities to both be large, but recall that both π_n and θ_k are constrained to the simplex.

For $\log \pi_{n,z_{n,d}}$ to be large, the posterior should assign all words to as few topics as possible.

For $\log \theta_{z_{n,d},x_{n,d}}$ to be large, the topics should put high probability on as few words as possible.

These goals are at odds. The LDA posterior balances these goals to find topics with sharply co-occuring words.

Coordinate Ascent Variational Inference for LDA

Let's derive the CAVI updates for LDA.

Assume a mean field family, and assume each factor is of the same exponential family form as the corresponding prior:

$$q(z_{n,d}; \boldsymbol{\lambda}_{n,d}^{(z)}) = \text{Cat}(z_{n,d}; \boldsymbol{\lambda}_{n,d}^{(z)}) \tag{14}$$

$$q(\boldsymbol{\pi}_n; \boldsymbol{\lambda}_n^{(\pi)}) = \text{Dir}(\boldsymbol{\pi}_n; \boldsymbol{\lambda}_n^{(\pi)}) \tag{15}$$

$$q(\boldsymbol{\theta}_k; \boldsymbol{\lambda}_k^{(\theta)}) = \text{Dir}(\boldsymbol{\theta}_k; \boldsymbol{\lambda}_k^{(\theta)}). \tag{16}$$

so $\boldsymbol{\lambda}_{n,d}^{(z)} \in \Delta_K$, $\boldsymbol{\lambda}_n^{(\pi)} \in \mathbb{R}_+^K$, and $\boldsymbol{\lambda}_k^{(\theta)} \in \mathbb{R}_+^V$ are the variational parameters.

(It turns out, for conjugate exponential family models, the optimal variational factors are of the same form as the prior anyway!)

CAVI updates for the topic assignments

Recall that the optimal CAVI updates are of the form in Eq. **??**.

We already derived the conditional distributions for Gibbs sampling (Slide 15).

For the topic assignments, the CAVI update is,

$$\log q(z_{n,d} = k; \lambda_{n,d}^{(z)}) = \mathbb{E}_{q(\pi_n)q(\theta_k)} \big[\log \pi_{n,k} + \log p(x_{n,d} \mid \theta_k) \big] + c \tag{17}$$

$$= \mathbb{E}_{q(\pi_n)} \big[\log \pi_{n,k} \big] + \mathbb{E}_{q(\theta_k)} \big[\log \theta_{k,x_{n,d}} \big] + c \tag{18}$$

$$= \log \operatorname{Cat}(z_{n,d} = k; \lambda_{n,d}^{(z)}) \tag{19}$$

$$\Rightarrow \log \lambda_{n,d,k}^{(z)} = \mathbb{E}_{q(\pi_n)} \big[\log \pi_{n,k} \big] + \mathbb{E}_{q(\theta_k)} \big[\log \theta_{k,x_{n,d}} \big] + c \tag{20}$$

Since $\lambda_{n,d}^{(z)}$ must sum to one,

$$\lambda_{n,d,k}^{(z)} = \frac{\exp \left\{ \mathbb{E}_{q(\pi_n)} \big[\log \pi_{n,k} \big] + \mathbb{E}_{q(\theta_k)} \big[\log \theta_{k,x_{n,d}} \big] \right\}}{\sum_{j=1}^{K} \exp \left\{ \mathbb{E}_{q(\pi_n)} \big[\log \pi_{n,j} \big] + \mathbb{E}_{q(\theta_j)} \big[\log \theta_{j,x_{n,d}} \big] \right\}} \tag{21}$$

Expectations under Dirichlet distributions

The variational factors for π_n and θ_k are Dirichlet distributions.

The necessary expectations have closed form expressions:

$$\mathbb{E}_{\mathrm{Dir}(\pi;\boldsymbol{\alpha})}[\log \pi_k] = \psi(\alpha_k) - \psi\Big(\sum_{j=1}^{K} \alpha_j\Big) \tag{22}$$

where $\psi(\cdot)$ is the *digamma function*, the logarithmic derivative of the gamma function.

CAVI updates for the topic proportions

Referring back to Slide 15, we see the CAVI update is,

$$\log q(\pi_n; \lambda_n^{(\pi)}) = \mathbb{E}_{q(\boldsymbol{z}_n)}\left[\log \mathrm{Dir}(\pi_n; \boldsymbol{\alpha}) + \sum_{k=1}^{K} N_{n,k} \log \pi_{n,k}\right] + c \tag{23}$$

$$= \sum_{k=1}^{K}(\alpha_k - 1 + \mathbb{E}_{q(\boldsymbol{z}_n)}[N_{n,k}]) \log \pi_{n,k} \tag{24}$$

$$= \log \mathrm{Dir}\left(\pi_n; \lambda_n^{(\pi)}\right) \tag{25}$$

$$\Rightarrow \lambda_n^{(\pi)} = \left[\alpha_1 + \mathbb{E}_{q(\boldsymbol{z}_n)}[N_{n,1}], \ldots, \alpha_K + \mathbb{E}_{q(\boldsymbol{z}_n)}[N_{n,K}]\right], \tag{26}$$

where

$$\mathbb{E}_{q(\boldsymbol{z}_n)}[N_{n,k}] = \sum_{d=1}^{D} \mathbb{E}_{q(z_{n,d})}[\mathbb{I}[z_{n,d} = k]] = \sum_{d=1}^{D} \lambda_{n,d,k}^{(z)}. \tag{27}$$

CAVI updates for the topic parameters

The topic parameter updates are similar

$$\log q(\boldsymbol{\theta}_k; \boldsymbol{\lambda}_k^{(\theta)}) = \mathbb{E}_{q(\boldsymbol{z})}\left[\log \mathrm{Dir}(\boldsymbol{\theta}_k; \boldsymbol{\phi}) + \sum_{v=1}^{V} N_{k,v} \log \theta_{k,v}\right] + c \tag{28}$$

$$= \sum_{v=1}^{V}(\phi_v - 1 + \mathbb{E}_{q(\boldsymbol{z})}[N_{k,v}]) \log \theta_{k,v} \tag{29}$$

$$= \log \mathrm{Dir}\left(\boldsymbol{\theta}_k; \boldsymbol{\lambda}_k^{(\theta)}\right) \tag{30}$$

$$\Rightarrow \boldsymbol{\lambda}_k^{(\theta)} = \left[\phi_1 + \mathbb{E}_{q(\boldsymbol{z})}[N_{k,1}], \ldots, \phi_V + \mathbb{E}_{q(\boldsymbol{z})}[N_{k,V}]\right], \tag{31}$$

where

$$\mathbb{E}_{q(\boldsymbol{z})}[N_{k,v}] = \sum_{n=1}^{N}\sum_{d=1}^{D}\mathbb{E}_{q(z_{n,d})}[\mathbb{I}[z_{n,d}=k]]\mathbb{I}[x_{n,d}=v] = \sum_{n=1}^{N}\sum_{d=1}^{D}\lambda_{n,d,k}^{(z)}\mathbb{I}[x_{n,d}=v]. \tag{32}$$

Calculating the ELBO

Dropping hyperparameters and variational parameters, the ELBO is,

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_q[\log p(\{\boldsymbol{x}_n, \boldsymbol{z}_n, \boldsymbol{\pi}_n\}_{n=1}^N, \{\boldsymbol{\theta}_k\}_{k=1}^K)] - \mathbb{E}_q[\log q(\{\boldsymbol{z}_n, \boldsymbol{\pi}_n\}_{n=1}^N, \{\boldsymbol{\theta}_k\}_{k=1}^K)] \tag{33}$$

Thanks to the factorization of the joint distribution and the variational posterior, this simplifies,

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\lambda}) = \sum_{n=1}^N \mathbb{E}_q[\log p(\boldsymbol{\pi}_n)] &+ \sum_{k=1}^K \mathbb{E}_q[\log p(\boldsymbol{\theta}_k)] \\
&+ \sum_{n=1}^N \sum_{d=1}^D \mathbb{E}_q[\log p(z_{n,d} \mid \boldsymbol{\pi}_n)] + \mathbb{E}_q[\log p(x_{n,d} \mid z_{n,d}, \boldsymbol{\theta})] \\
&- \sum_{k=1}^K \mathbb{E}_q[\log q(\boldsymbol{\theta}_k)] - \sum_{n=1}^N \mathbb{E}_q[\log q(\boldsymbol{\pi}_n)] - \sum_{n=1}^N \sum_{d=1}^D \mathbb{E}_q[\log q(z_{n,d})] \tag{34}
\end{aligned}$$

The first two terms are Dirichlet cross entropies and the last three terms are entropies. Tensorflow probability has already implemented these functions for you, so this is easier to calculate than you might think!

Calculating the ELBO II

The middle terms are,

$$\mathbb{E}_q[\log p(z_{n,d} \mid \pi_n)] = \sum_{k=1}^{K} \mathbb{E}_{q(z_{n,d})}[\mathbb{I}[z_{n,d} = k]] \cdot \mathbb{E}_{q(\pi_n)}[\log \pi_{n,k}] \tag{35}$$

$$= \sum_{k=1}^{K} \lambda_{n,d,k}^{(z)} \cdot \mathbb{E}_{q(\pi_n)}[\log \pi_{n,k}] \tag{36}$$

and

$$\mathbb{E}_q[\log p(x_{n,d} \mid z_{n,d}, \boldsymbol{\theta})] = \sum_{k=1}^{K} \mathbb{E}_{q(z_{n,d})}[\mathbb{I}[z_{n,d} = k]] \cdot \mathbb{E}_{q(\boldsymbol{\theta}_k)}[\log \theta_{k,x_{n,d}}] \tag{37}$$

$$= \sum_{k=1}^{K} \lambda_{n,d,k}^{(z)} \cdot \mathbb{E}_{q(\boldsymbol{\theta}_k)}[\log \theta_{k,x_{n,d}}] \tag{38}$$

Again, these involve only simple expectations of Dirichlet distributions.

Bag of word counts assumption and exchangeability

LDA models the words as **exchangeable** random variables. That is, the joint distribution is invariant to permutations:

$$p(\mathbf{x}_n) = p(x_{n,1}, \ldots, x_{n,D}) \equiv p(x_{n,\sigma(1)}, \ldots, p(x_{n,\sigma(D)})) \tag{39}$$

where $\sigma(\cdot)$ is any permutation of the indices $\{1, \ldots, D\}$.

In text modeling, this is called the **bag of words** assumption.

In LDA, this manifests in the joint probability (eq. 9, reproduced below) only depending on **word counts**,

$$p(\{\boldsymbol{\theta}_k\}_{k=1}^K, \{\boldsymbol{\pi}_n, \mathbf{z}_n, \mathbf{x}_n\}_{n=1}^N \mid \boldsymbol{\phi}, \boldsymbol{\alpha})$$
$$\propto \left[\prod_{k=1}^K \prod_{v=1}^V \theta_{k,v}^{\phi_v - 1} \right] \prod_{n=1}^N \left(\left[\prod_{k=1}^K \pi_{n,k}^{\alpha_k + N_{n,\cdot,k} - 1} \right] \left[\prod_{k=1}^K \prod_{v=1}^V \theta_{k,v}^{N_{n,v,k}} \right] \right) \tag{40}$$

Working with word counts

The fact that the joint probability depends only on word counts suggests an alternative representation of the data.

Let, $\mathbf{y}_n \in \mathbb{N}^V$ denote the n-th document represented as a vector of word counts; i.e.,

$$y_{n,v} = \sum_{d=1}^{D} \mathbb{I}[x_{n,d} = v]. \tag{41}$$

Typically these vector will be **sparse**.

Likewise, let $\mathbf{c}_{n,v} \in \mathbb{N}^K$ be a vector of **latent counts** denoting how many times word v in document n was attributed to each of the K topics; i.e.,

$$c_{n,v,k} = \sum_{d=1}^{D} \mathbb{I}[x_{n,d} = v] \, \mathbb{I}[z_{n,d} = k] \tag{42}$$

We must have that $\sum_{k=1}^{K} c_{n,v,k} = y_{n,v}$.

LDA with word counts and the corresponding data types

Picture: Draw the data types for y, c, π, and θ.

Gibbs and CAVI updates with word counts

In terms of word counts, the conditional distribution of $c_{n,v} \in \mathbb{N}^K$ is,

$$p(c_{n,v} \mid y_n, \{\theta_k\}_{k=1}^K, \pi_n) = \mathrm{Mult}\left(c_{n,v} \,\middle|\, y_n, \left[\frac{\pi_{n,1}\theta_{1,v}}{\sum_{k=1}^K \pi_{n,k}\theta_{k,v}}, \ldots, \frac{\pi_{n,K}\theta_{K,v}}{\sum_{k=1}^K \pi_{n,k}\theta_{k,v}} \right] \right) \tag{43}$$

Instead of sampling each $z_{n,d}$ in the Gibbs sampler, we can directly sample count vectors $c_{n,v}$.

Likewise, for CAVI, instead of having parameters for each $q(z_{n,d} \mid \lambda_{n,d}^{(z)})$, only store one for each term v,

$$q(c_{n,v}; \lambda_{n,v}^{(c)}) = \mathrm{Mult}(c_{n,v} \mid y_{n,v}, \lambda_{n,v}^{(c)}) \tag{44}$$

$$\lambda_{n,v,k}^{(c)} = \frac{\exp\left\{ \mathbb{E}_{q(\pi_n)}\left[\log \pi_{n,k}\right] + \mathbb{E}_{q(\theta_k)}\left[\log \theta_{k,v}\right] \right\}}{\sum_{j=1}^K \exp\left\{ \mathbb{E}_{q(\pi_n)}\left[\log \pi_{n,j}\right] + \mathbb{E}_{q(\theta_j)}\left[\log \theta_{j,v}\right] \right\}} \tag{45}$$

From these variational parameters, it's easy to compute the expected summary counts,

$$\mathbb{E}_q[N_{n,\cdot,k}] = \sum_{v=1}^V y_{n,v} \lambda_{n,v,k}^{(c)} \qquad\qquad \mathbb{E}_q[N_{\cdot,k,v}] = \sum_{n=1}^N y_{n,v} \lambda_{n,v,k}^{(c)} \tag{46}$$

Scaling up to very large datasets

There are a few tricks to make LDA much more scalable.

First, to save memory, you only need to track, $\lambda_{n,v}^{(c)} = [\lambda_{n,v,1}^{(c)}, \ldots, \lambda_{n,v,K}^{(c)}]$ if $y_{n,v} > 0$.

Likewise, you can process documents in rolling fashion, discarding $\lambda_{n,v}^{(c)}$ once you've updated $\mathbb{E}_q[N_{k,v}]$ and $\lambda_n^{(\pi)}$.

Finally, you can use **stochastic variational inference** [Hoffman et al., 2013] to work with mini-batches of documents to get Monte Carlo estimates of $\mathbb{E}_q[N_{k,v}]$.

SVI can be seen as **stochastic gradient ascent** on the ELBO using **natural gradients** Amari [1998]; i.e., gradient descent preconditioned with the Fisher information matrix.

Evaluating topic models

The key hyperparameter is *K*, the number of topics. By now, we've seen a few different ways of setting these "complexity knobs."

Question: what approaches could we take?

Blei recommends another method that differs slightly from what we've seen thus far. He suggests evaluating,

$$p(\mathbf{x}_{n'}^{\text{out}} \mid \mathbf{x}_{n'}^{\text{in}}, \{\mathbf{x}_n\}_{n=1}^N) = \int p(\mathbf{x}_{n'}^{\text{out}} \mid \boldsymbol{\pi}_{n'}, \{\boldsymbol{\theta}_k\}_{k=1}^K) \, p(\boldsymbol{\pi}_{n'} \mid \mathbf{x}_{n'}^{\text{in}}, \{\boldsymbol{\theta}_k\}_{k=1}^K) \, p(\{\boldsymbol{\theta}_k\}_{k=1}^K \mid \{\mathbf{x}_n\}_{n=1}^N) \, \mathrm{d}\boldsymbol{\pi}_{n'}$$

(47)

where

▶ $\mathbf{x}_{n'}^{\text{out}}$ consists of a subset of words in a held-out document n'

▶ $\mathbf{x}_{n'}^{\text{in}}$ are the remaining words in that document, which are used to estimate the topic proportions, and

▶ $\{\mathbf{x}_n\}_{n=1}^N$ are the training documents used to estimate the topics $\{\boldsymbol{\theta}_k\}_{k=1}^K$.

Evaluating topic models II

Question: why not simply compare the ELBO for different values of K? It's related to the marginal likelihood, after all.

Other mixed membership models

- ▶ LDA evolved from a long line of work on topic modeling. Deerwester et al. [1990] proposed **latent semantic analysis** and Hofmann [1999] proposed a probabilistic version called the **aspect model**.

- ▶ Pritchard et al. [2000] developed MM models in **population genetics**.

- ▶ Erosheva et al. [2007] used MM models for **survey data**.

- ▶ Airoldi et al. [2008] developed MM models for **community detection in networks**. Gopalan and Blei [2013] developed a stochastic variational inference algorithm for this model.

- ▶ Gopalan et al. [2013] proposed **Poisson matrix factorization**, which is closely related to LDA.

References I

David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120, 2006.

Jordan Boyd-Graber, Yuening Hu, and David Mimno. Applications of topic models. *Found. Trends® Inf. Retr.*, 11(2-3):143–296, 2017.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(5), 2013.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

References II

Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1105–1112, 2009.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.*, 41(6):391–407, September 1990.

T Hofmann. Probabilistic latent semantic indexing. *Proceedings of the 22nd annual international ACM*, 1999.

J K Pritchard, M Stephens, and P Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, June 2000.

Elena A Erosheva, Stephen E Fienberg, and Cyrille Joutard. Describing disability through individual-level mixture models for multivariate binary data. *Ann. Appl. Stat.*, 1(2):346–384, 2007.

Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.*, 9:1981–2014, September 2008.

References III

Prem K Gopalan and David M Blei. Efficient discovery of overlapping communities in massive networks. *Proc. Natl. Acad. Sci. U. S. A.*, 110(36):14534–14539, September 2013.

Prem Gopalan, Jake M Hofman, and David M Blei. Scalable recommendation with Poisson factorization. November 2013.