Topic Models and Coordinate Ascent Variational Inference

STATS 305C: Applied Statistics

Scott Linderman

April 25, 2022

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.

Topic models

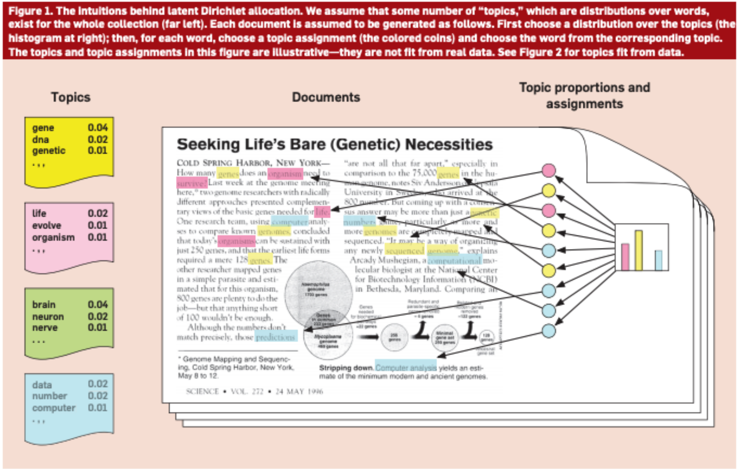
The most common mixed-membership model is the **topic model**, a generative model for documents.

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].

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 $\boldsymbol{\theta}_k \sim p(\boldsymbol{\theta}_k \mid \boldsymbol{\phi})$
- for each **document** $n = 1, \ldots, N$:
  - sample **topic proportions** $\pi_n \sim \mathrm{Dir}(\pi_n \mid \boldsymbol{\alpha})$
  - for each **word** $d = 1, \ldots, D$:
    - sample **topic assignment** $z_{n,d} \in \{1, \ldots, K\}$ from a cateogrical distribution $z_{n,d} \sim \mathrm{Cat}(\boldsymbol{\pi}_n)$
    - sample **word** $x_{n,d} \in \{1, \ldots, V\}$ from a cateogrical distribution, $x_{n,d} \sim \mathrm{Cat}(\boldsymbol{\theta}_{z_{n,d}})$

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,

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

As in mixture models, we can write this equivalently as

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

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 $\boldsymbol{\theta}_k \in \Delta_V$ and words $x_{n,d} \in \{1, \ldots, V\}$,

$$\boldsymbol{\theta}_k \overset{\text{iid}}{\sim} \mathrm{Dir}(\boldsymbol{\phi}), \tag{3}$$

$$\boldsymbol{\pi}_n \overset{\text{iid}}{\sim} \mathrm{Dir}(\boldsymbol{\alpha}), \tag{4}$$

$$z_{n,d} \overset{\text{iid}}{\sim} \mathrm{Cat}(\boldsymbol{\pi}_n) \tag{5}$$

$$x_{n,d} \overset{\text{iid}}{\sim} \mathrm{Cat}(\boldsymbol{\theta}_{z_{n,d}}) \tag{6}$$

Latent Dirichlet allocation II

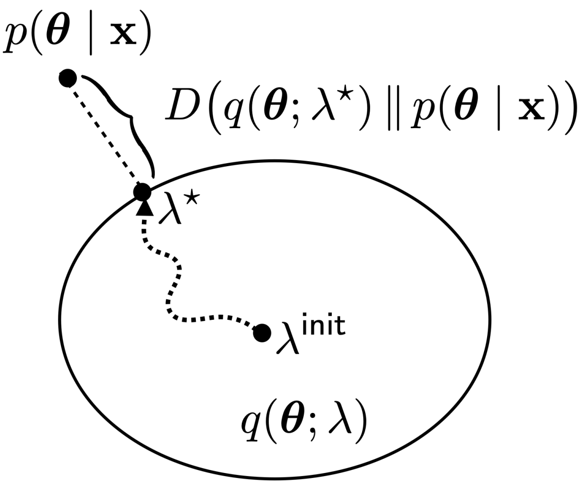
Plugging in these assumptions, the joint probability is,

$$
\begin{aligned}
& p(\{\boldsymbol{\theta}_k\}_{k=1}^K, \{\boldsymbol{\pi}_n, \boldsymbol{z}_n, \boldsymbol{x}_n\}_{n=1}^N \mid \boldsymbol{\phi}, \boldsymbol{\alpha}) && (7)\\
&= \prod_{k=1}^K \mathrm{Dir}(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}) \prod_{n=1}^N \left[\mathrm{Dir}(\boldsymbol{\pi}_n \mid \boldsymbol{\alpha}) \prod_{d=1}^D \pi_{n, z_{n,d}} \, \theta_{z_{n,d}, x_{n,d}} \right] && (8)\\
&= \prod_{k=1}^K \mathrm{Dir}(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}) \prod_{n=1}^N \left(\mathrm{Dir}(\boldsymbol{\pi}_n \mid \boldsymbol{\alpha}) \prod_{d=1}^D \left[\prod_{k=1}^K \pi_{n,k}^{\mathbb{I}[z_{n,d}=k]} \right] \left[\prod_{k=1}^K \prod_{v=1}^V \theta_{k,v}^{\mathbb{I}[x_{n,d}=v]\mathbb{I}[z_{n,d}=k]} \right] \right)\\
&= \left[\prod_{k=1}^K \prod_{v=1}^V \theta_{k,v}^{\phi_v - 1} \right] \prod_{n=1}^N \left(\left[\prod_{k=1}^K \pi_{n,k}^{\alpha_k + N_{n,\cdot,k} - 1} \right] \left[\prod_{k=1}^K \prod_{v=1}^V \theta_{k,v}^{N_{n,v,k}} \right] \right) && (9)
\end{aligned}
$$

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.

A view of variational inference



Key questions

- *What parametric family should we use?*
  - This lecture: the **mean-field family**.
- *How should we measure closeness?*
  - This lecture: the **Kullback-Leibler (KL)** divergence.
- *How do we find the closest distribution in that family?*
  - This lecture: **coordinate ascent.**

These choices are what Blei et al. [2017] call **coordinate ascent variational inference** CAVI).

The mean-field family

The *mean-field family* gets its name from statistical mechanics. It treats each latent variable and parameter as independent with its own variational parameter,

$$q(\boldsymbol{\vartheta}; \boldsymbol{\lambda}) = \prod_{j=1}^{J} q(\vartheta_j; \lambda_j). \tag{14}$$

For example, in LDA the mean field approximation treats each topic, topic proportion, and topic assignment as independent,

$$q(\{\boldsymbol{\theta}_k\}_{k=1}^{K}, \{\boldsymbol{\pi}_n, \mathbf{z}_n\}_{n=1}^{N}; \boldsymbol{\lambda}) = \prod_{k=1}^{K} q(\boldsymbol{\theta}_k; \boldsymbol{\lambda}_k^{(\theta)}) \prod_{n=1}^{N} q(\boldsymbol{\pi}_n; \boldsymbol{\lambda}_n^{(\pi)}) \prod_{n=1}^{N} \prod_{d=1}^{D} q(z_{n,d}; \boldsymbol{\lambda}_{n,d}^{(z)}) \tag{15}$$

Question: Is this a good approximation to the posterior?

The Kullback-Leibler (KL) divergence

The KL divergence is a measure of closeness between two distributions. It is defined as,

$$D_{\mathrm{KL}}\left(q(\boldsymbol{\vartheta};\boldsymbol{\lambda}) \parallel p(\boldsymbol{\vartheta} \mid \boldsymbol{x}\right) = \mathbb{E}_{q(\boldsymbol{\vartheta};\boldsymbol{\lambda})}\left[\log \frac{q(\boldsymbol{\vartheta};\boldsymbol{\lambda})}{p(\boldsymbol{\vartheta} \mid \boldsymbol{x})}\right] \tag{16}$$

$$= \mathbb{E}_{q(\boldsymbol{\vartheta};\boldsymbol{\lambda})}\left[\log q(\boldsymbol{\vartheta};\boldsymbol{\lambda})\right] - \mathbb{E}_{q(\boldsymbol{\vartheta};\boldsymbol{\lambda})}\left[\log p(\boldsymbol{\vartheta} \mid \boldsymbol{x})\right] \tag{17}$$

It has some nice properties:

- It is non-negative.
- It is zero iff $q(\boldsymbol{\vartheta};\boldsymbol{\lambda}) \equiv p(\boldsymbol{\vartheta} \mid \boldsymbol{x})$.
- It is defined in terms of expectations wrt q.

But it's also a bit weird...

- It's asymmetric ($D_{\mathrm{KL}}\left(q \parallel p\right) \neq D_{\mathrm{KL}}\left(p \parallel q\right)$).

The evidence lower bound (ELBO) from another angle

More concerning, the KL divergence involves the posterior $p(\boldsymbol{\vartheta} \mid \boldsymbol{x})$, which we cannot compute!

But notice that...

$$\begin{aligned} D_{\mathrm{KL}}\left(q(\boldsymbol{\vartheta};\boldsymbol{\lambda}) \,\|\, p(\boldsymbol{\vartheta} \mid \boldsymbol{x})\right) &= \mathbb{E}_{q(\boldsymbol{\vartheta};\boldsymbol{\lambda})}\left[\log q(\boldsymbol{\vartheta};\boldsymbol{\lambda})\right] - \mathbb{E}_{q(\boldsymbol{\vartheta};\boldsymbol{\lambda})}\left[\log p(\boldsymbol{\vartheta} \mid \boldsymbol{x})\right] && (18) \\ &= \mathbb{E}_{q(\boldsymbol{\vartheta};\boldsymbol{\lambda})}\left[\log q(\boldsymbol{\vartheta};\boldsymbol{\lambda})\right] - \mathbb{E}_{q(\boldsymbol{\vartheta};\boldsymbol{\lambda})}\left[\log p(\boldsymbol{\vartheta}, \boldsymbol{x})\right] + \mathbb{E}_{q(\boldsymbol{\vartheta};\boldsymbol{\lambda})}\left[\log p(\boldsymbol{x})\right] && (19) \\ &= \underbrace{\mathbb{E}_{q(\boldsymbol{\vartheta};\boldsymbol{\lambda})}\left[\log q(\boldsymbol{\vartheta};\boldsymbol{\lambda})\right] - \mathbb{E}_{q(\boldsymbol{\vartheta};\boldsymbol{\lambda})}\left[\log p(\boldsymbol{\vartheta}, \boldsymbol{x})\right]}_{\text{negative ELBO,}-\mathscr{L}(\boldsymbol{\lambda})} + \underbrace{\log p(\boldsymbol{x})}_{\text{evidence}} && (20) \end{aligned}$$

The first term involves the log joint, which we can compute, and the last term is independent of the variational parameters!

Rearranging, we see that $\mathscr{L}(\boldsymbol{\lambda})$ is a lower bound on the marginal likelihood, aka the evidence,

$$\mathscr{L}(\boldsymbol{\lambda}) = \log p(\boldsymbol{x}) - D_{\mathrm{KL}}\left(q(\boldsymbol{\vartheta};\boldsymbol{\lambda}) \,\|\, p(\boldsymbol{\vartheta} \mid \boldsymbol{x})\right) \leq \log p(\boldsymbol{x}). \tag{21}$$

That’s why we call it the **evidence lower bound (ELBO)**.

Optimizing the ELBO with coordinate ascent

We want to find the variational parameters $\boldsymbol{\lambda}$ that minimize the KL divergence or, equivalently, maximize the ELBO.

For the mean-field family, we can typically do this via **coordinate ascent**.

Consider optimizing the parameters for one factor $q(\vartheta_j; \lambda_j)$. As a function of λ_j, the ELBO is,

$$
\begin{aligned}
\mathscr{L}(\boldsymbol{\lambda}) &= \mathbb{E}_{q(\vartheta_j;\lambda_j)}\left[\mathbb{E}_{q(\boldsymbol{\vartheta}_{\neg j};\boldsymbol{\lambda}_{\neg j})}\left[\log p(\boldsymbol{\vartheta}, \boldsymbol{x})\right]\right] - \mathbb{E}_{q(\vartheta_j;\lambda_j)}[\log q(\vartheta_j; \lambda_j)] + c && (22)\\
&= \mathbb{E}_{q(\vartheta_j;\lambda_j)}\left[\mathbb{E}_{q(\boldsymbol{\vartheta}_{\neg j};\boldsymbol{\lambda}_{\neg j})}\left[\log p(\vartheta_j \mid \boldsymbol{\vartheta}_{\neg j}, \boldsymbol{x})\right]\right] - \mathbb{E}_{q(\vartheta_j;\lambda_j)}[\log q(\vartheta_j; \lambda_j)] + c' && (23)\\
&= -D_{\mathrm{KL}}\left(q(\vartheta_j; \lambda_j) \,\|\, \tilde{p}(\vartheta_j)\right) + c'' && (24)
\end{aligned}
$$

where

$$
\tilde{p}(\vartheta_j) \propto \exp\left\{\mathbb{E}_{q(\boldsymbol{\vartheta}_{\neg j};\boldsymbol{\lambda}_{\neg j})}\left[\log p(\vartheta_j \mid \boldsymbol{\vartheta}_{\neg j}, \boldsymbol{x})\right]\right\} \tag{25}
$$

The ELBO is maximized wrt λ_j when this KL is minimized; i.e. when $q(\vartheta_j; \lambda_j) = \tilde{p}(\vartheta_j)$, the exponentiated expected log conditional probability, holding all other factors fixed.

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:

$$q(z_{n,d}; \boldsymbol{\lambda}_{n,d}^{(z)}) = \mathrm{Cat}(z_{n,d}; \boldsymbol{\lambda}_{n,d}^{(z)}) \tag{26}$$

$$q(\boldsymbol{\pi}_n; \boldsymbol{\lambda}_n^{(\pi)}) = \mathrm{Dir}(\boldsymbol{\pi}_n; \boldsymbol{\lambda}_n^{(\pi)}) \tag{27}$$

$$q(\boldsymbol{\theta}_k; \boldsymbol{\lambda}_k^{(\theta)}) = \mathrm{Dir}(\boldsymbol{\theta}_k; \boldsymbol{\lambda}_k^{(\theta)}). \tag{28}$$

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!)

Bag of word counts assumption and exchangeability

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

$$p(\boldsymbol{x}_n) = p(x_{n,1}, \ldots, x_{n,D}) \equiv p(x_{n,\sigma(1)}, \ldots, p(x_{n,\sigma(D)})) \tag{51}$$

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**,

$$\begin{aligned} p(\{\boldsymbol{\theta}_k\}_{k=1}^K, \{\boldsymbol{\pi}_n, \boldsymbol{z}_n, \boldsymbol{x}_n\}_{n=1}^N \mid \boldsymbol{\phi}, \boldsymbol{\alpha}) \\ \propto \left[\prod_{k=1}^K \prod_{v=1}^V \theta_{k,v}^{\phi_v - 1}\right] \prod_{n=1}^N \left(\left[\prod_{k=1}^K \pi_{n,k}^{\alpha_k + N_{n,\cdot,k} - 1}\right] \left[\prod_{k=1}^K \prod_{v=1}^V \theta_{k,v}^{N_{n,v,k}}\right] \right) \end{aligned} \tag{52}$$

Working with word counts

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

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

$$y_{n,v} = \sum_{d=1}^{D} \mathbb{I}[x_{n,d} = v]. \tag{53}$$

Typically these vector will be **sparse**.

Likewise, let $\boldsymbol{c}_{n,v} \in \mathbb{N}^K$ 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.,

$$c_{n,v,k} = \sum_{d=1}^{D} \mathbb{I}[x_{n,d} = v]\, \mathbb{I}[z_{n,d} = k] \tag{54}$$

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 $\boldsymbol{y}$, $\boldsymbol{c}$, π, and θ.

Gibbs and CAVI updates with word counts

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

$$p(\boldsymbol{c}_{n,v} \mid y_n, \{\boldsymbol{\theta}_k\}_{k=1}^K, \boldsymbol{\pi}_n) = \mathrm{Mult}\left(\boldsymbol{c}_{n,v} \,\middle|\, y_n, \left[\frac{\pi_{n,1}\theta_{1,v}}{\sum_{k=1}^K \pi_{n,k}\theta_{k,v}}, \ldots, \frac{\pi_{n,K}\theta_{K,v}}{\sum_{k=1}^K \pi_{n,k}\theta_{k,v}}\right]\right) \tag{55}$$

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

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

$$q(\boldsymbol{c}_{n,v}; \boldsymbol{\lambda}_{n,v}^{(c)}) = \mathrm{Mult}(\boldsymbol{c}_{n,v} \mid y_{n,v}, \boldsymbol{\lambda}_{n,v}^{(c)}) \tag{56}$$

$$\lambda_{n,v,k}^{(c)} = \frac{\exp\left\{\mathbb{E}_{q(\boldsymbol{\pi}_n)}\left[\log \pi_{n,k}\right] + \mathbb{E}_{q(\boldsymbol{\theta}_k)}\left[\log \theta_{k,v}\right]\right\}}{\sum_{j=1}^K \exp\left\{\mathbb{E}_{q(\boldsymbol{\pi}_n)}\left[\log \pi_{n,j}\right] + \mathbb{E}_{q(\boldsymbol{\theta}_j)}\left[\log \theta_{j,v}\right]\right\}} \tag{57}$$

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

$$\mathbb{E}_q[N_{n,\cdot,k}] = \sum_{v=1}^V y_{n,v}\lambda_{n,v,k}^{(c)} \qquad \mathbb{E}_q[N_{\cdot,k,v}] = \sum_{n=1}^N y_{n,v}\lambda_{n,v,k}^{(c)} \tag{58}$$

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, $\boldsymbol{\lambda}_{n,v}^{(c)} = [\lambda_{n,v,1}^{(c)}, \ldots, \lambda_{n,v,K}^{(c)}]$ if $y_{n,v} > 0$.

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

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,

$$p(\boldsymbol{x}_{n'}^{\text{out}} \mid \boldsymbol{x}_{n'}^{\text{in}}, \{\boldsymbol{x}_n\}_{n=1}^N) = \int p(\boldsymbol{x}_{n'}^{\text{out}} \mid \boldsymbol{\pi}_{n'}, \{\boldsymbol{\theta}_k\}_{k=1}^K)\, p(\boldsymbol{\pi}_{n'} \mid \boldsymbol{x}_{n'}^{\text{in}}, \{\boldsymbol{\theta}_k\}_{k=1}^K)\, p(\{\boldsymbol{\theta}_k\}_{k=1}^K \mid \{\boldsymbol{x}_n\}_{n=1}^N)\, \mathrm{d}\boldsymbol{\pi}_{n'} \tag{59}$$

where

- $\boldsymbol{x}_{n'}^{\text{out}}$ consists of a subset of words in a held-out document n'
- $\boldsymbol{x}_{n'}^{\text{in}}$ are the remaining words in that document, which are used to estimate the topic proportions, and
- $\{\boldsymbol{x}_n\}_{n=1}^N$ 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.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, April 2017.

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

References II

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

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.

References III

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.

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.