663 Final Project: LDA Collapsed Gibbs Modeling

David Chester Bernardo Dionis Altamash Rafiq

1. Abstract

We approach Latent Dirichlet Allocation (LDA) from the perspective of collapsed Gibbs sampling based off of the work presented by Griffiths and Steyvers, of whom built their work off the generative probabilistic model for documents introduced by Blei, NG, and Jordan. Similar to Griffiths and Styevers, we focused on an MCMC approach to sample the topic specific full conditionals where our samples approximate the true marginal posteriors. However, we do not explore varying nor growing our number of topics. Noting the work by Porteous et. al., we heuristically choose a total of 20 topics as the authors illustrate 20 topics can explain, on average, nearly ninety percent of a document within a corpus for a broad range of different corpora. We increase the speed of the algorithm by ***

2. Background

We are primarily using “Finding Scientific Topics” (Griffiths and Styevers), but we also reference “Latent Dirichlet Allocation” (Blei et. al.) and “Fast Collapsed Gibbs Sampling for Latent Dirichlet Allocation” (Porteous et. al.) and briefly discuss “Latent Dirichlet Allocation Model Training with Differential Privacy” (Zhao et. al.).

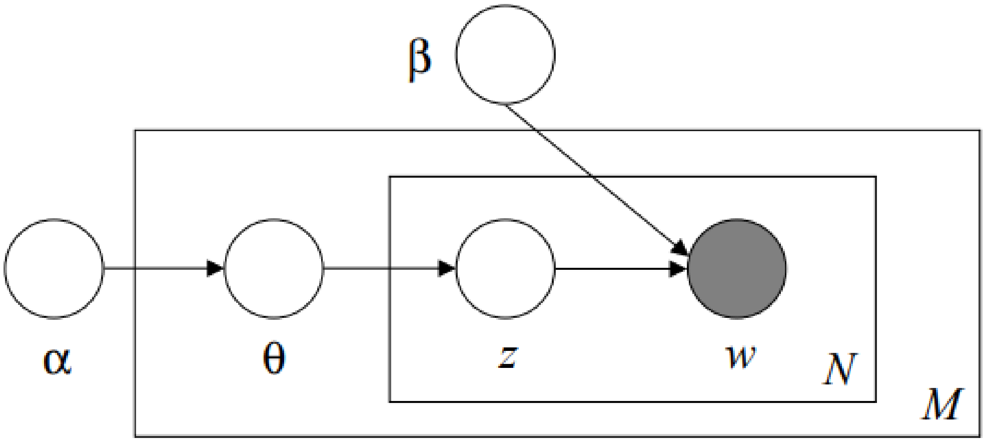
To begin, we are primarily interested in

$$p(\mathbf{z}|\mathbf{w}) = \frac{p(\mathbf{z},\mathbf{w})}{\Sigma_{\mathbf{z}} p(\mathbf{z},\mathbf{w})}$$

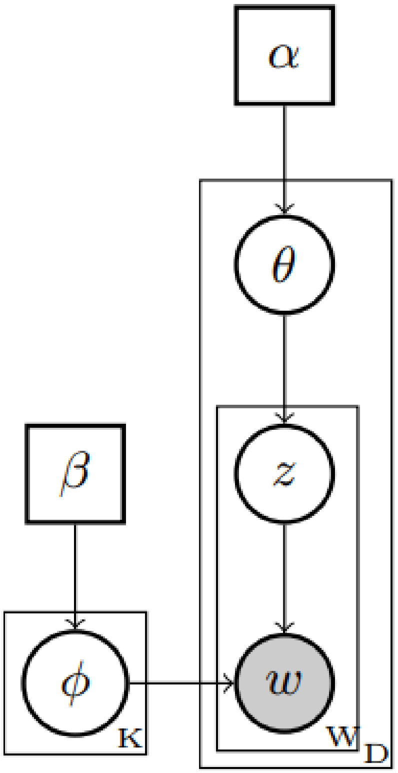
Where $p(\mathbf{z}|\mathbf{w})$ represents the posterior distribution over the assigments of words to topics. However, the denominator cannot be factorized and this quantity cannot be computed directly. The approach by Griffiths et. al. is to convert this to an MCMC set up on top of a slightly altered LDA setting first provided by Blei et. al.

We first discuss the hierarchical Bayesian model before discussing the use of conjugacy in order to delve into how Griffiths and Steyvers obtained various results in more detail than the authors originally provide.

Consider the graphical model representation of LDA from Blei et. al.:



We contrast this to the graphical LDA model used by Griffiths et. al. and Porteous et. al.:



Where we noticably see that the two constructions differ by ϕ over K topics.

Seen in a different light, the complete probability model by Griffiths et. al. is:

$$\begin{aligned} w_i|z_i,\phi^{(z_i)} &\sim Discrete(\phi^{(z_i)}) \\ \phi &\sim Dirichlet(\beta) \\ z_i|\theta^{(d_i)} &\sim Discrete(\theta^{(d_i)}) \\ \theta &\sim Dirichlet(\alpha) \end{aligned}$$

However, we can view this in the multivariate setting where $\mathbf{w}|\mathbf{z},\phi \sim Multinomial(\phi)$, and similarly, $\mathbf{z}|\theta \sim Multinomial(\theta)$, however, the above depiction provides a nice relationship between each particular

$w_i, z_i, \phi^{(z_i)}, \theta^{(d_i)}$. Also note that while $\boldsymbol{\alpha}$ and β could be vector specified, here we assume one value for $\alpha = \mathbf{1}\alpha$ and similarly $\beta = \mathbf{1}\beta$, i.e. we have symmetric Dirichlet distributions. The approach by Griffiths et. al. was to integrate out the latent variables $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ allowing us to avoid a much longer chain structure. However, they leave much of the calculations to the reader, of which we attempt to spell out below. I.e. we can note that:

$$p(\mathbf{z}|\mathbf{w}) = \frac{p(\mathbf{z}, \mathbf{w})}{\Sigma_{\mathbf{z}} p(\mathbf{z}, \mathbf{w})}$$

where

$$p(\mathbf{w}, \mathbf{z}) = p(\mathbf{w}|\mathbf{z}) p(\mathbf{z})$$

where

$$p(\mathbf{z}, \mathbf{w}, \phi) = p(\mathbf{w}|\mathbf{z}, \phi) p(\mathbf{z}|\phi) p(\phi)$$

where integrating ϕ out of $p(\mathbf{w}|\mathbf{z}, \phi)p(\phi)$ is equivalent to $p(\mathbf{w}|\mathbf{z})$. However, we know via Multinomial-Dirichlet conjugacy, that $p(\phi|\mathbf{w}, \mathbf{z}) \; \alpha \; p(\mathbf{w}|\mathbf{z}, \phi)p(\phi)$ is the same as $\phi \sim Dir(\beta + \boldsymbol{\Sigma}\mathbf{w_i})$. By integrating ϕ we have equivalently:

$$\int_\phi \frac{\Gamma(\Sigma(\beta_i + \Sigma\mathbf{w_i}))}{\Pi\,\Gamma(\beta_i + \Sigma\mathbf{w_i})} \Pi\phi_i^{(\beta_i + \Sigma\mathbf{w_i})-1} d\phi$$

Griffifths et. al. does not provide the term above, nor any steps to the final form below with ϕ integrated out, i.e.

$$P(\mathbf{w} \mid \mathbf{z}) = \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^T \prod_{j=1}^{T} \frac{\prod_w \Gamma\left(n_j^{(w)} + \beta\right)}{\Gamma\left(n_j^{(\cdot)} + W\beta\right)},$$

A number of steps seem to missing which we attempt to uncover below by reverse engineering from the results to find forms that might be of interest, i.e

$$\begin{aligned}
P(w \mid z) =& \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^\top \prod_{j=1}^{T} \frac{\Pi_w \Gamma\left(n_j^{(\omega)} + \beta\right)}{\Gamma\left(n_j^{(\cdot)} + W\beta\right)} \\
&= \int_\phi \frac{\Gamma(\Sigma(\beta_i + \Sigma\mathbf{w_i}))}{\Pi\,\Gamma(\beta_i + \Sigma\mathbf{w_i})} \Pi\phi_i^{(\beta_i + \Sigma\mathbf{w_i})-1} d\phi \\
&= \underbrace{\prod_{j=1}^{T} \frac{\Pi_w \Gamma\left(n_j^{(w)} + \beta\right)}{\Gamma\left(n_j^{(\cdot)} + W\beta\right)} \cdot \prod_{j=1}^{T} \frac{\Gamma\left(n_j^{(\cdot)} + W\beta\right)}{\Pi_w \Gamma\left(n_j^{(\omega)} + \beta\right)}}_{\text{multiply by 1.}} \int_\phi \frac{\Gamma(\Sigma(\beta_i + \Sigma\mathbf{w_i}))}{\Pi\,\Gamma(\beta_i + \Sigma\mathbf{w_i})} \Pi\phi_i^{(\beta_i + \Sigma\mathbf{w_i})-1} d\phi \\
&= \prod_{j=1}^{T} \frac{\Pi_w \Gamma\left(n_j^{(w)} + \beta\right)}{\Gamma\left(n_j^{(\cdot)} + W\beta\right)} \frac{\Gamma(\Sigma(\beta_i + \Sigma\mathbf{w_i}))}{\Pi\,\Gamma(\beta_i + \Sigma\mathbf{w_i})} \int_\phi \prod_{j=1}^{T} \frac{\Gamma\left(n_j^{(\cdot)} + W\beta\right)}{\Pi_w \Gamma\left(n_j^{(\omega)} + \beta\right)} \Pi\phi_i^{(\beta_i + \Sigma\mathbf{w_i})-1} d\phi
\end{aligned}$$

and if $\left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^\top = \frac{\Gamma(\Sigma(\beta_i + \Sigma\mathbf{w_i}))}{\Pi\,\Gamma(\beta_i + \Sigma\mathbf{w_i})}$ then this would imply that $\int_\phi \prod_{j=1}^{T} \frac{\Gamma\left(n_j^{(\cdot)} + W\beta\right)}{\Pi_w \Gamma\left(n_j^{(\omega)} + \beta\right)} \Pi\phi_i^{(\beta_i + \Sigma\mathbf{w_i})-1} d\phi$ integrates to 1, but this seems potentially reasonable since $\prod_{j=1}^{T} \frac{\Gamma\left(n_j^{(\cdot)} + W\beta\right)}{\Pi_w \Gamma\left(n_j^{(\omega)} + \beta\right)}$ seems to resemble a product that would yield $\frac{\Gamma(\Sigma(\beta_i + \Sigma\mathbf{w_i}))}{\Pi\,\Gamma(\beta_i + \Sigma\mathbf{w_i})}$. However in the interest of not speculating, we can again “multiply by 1”.

$$
\begin{aligned}
P(w \mid z) &= \prod_{j=1}^{T} \frac{\Pi_w \Gamma\left(n_j^{(w)} + \beta\right)}{\Gamma\left(n_j^{(\cdot)} + W\beta\right)} \cdot \prod_{j=1}^{T} \frac{\Gamma\left(n_j^{(\cdot)} + W\beta\right)}{\Pi_w \Gamma\left(n_j^{(\omega)} + \beta\right)} \int_{\phi} \frac{\Gamma(\Sigma(\beta_i + \Sigma \mathbf{w_i}))}{\Pi\, \Gamma(\beta_i + \Sigma \mathbf{w_i})} \Pi \phi_i^{(\beta_i + \Sigma \mathbf{w_i}) - 1} d\phi \\
&= \prod_{j=1}^{T} \frac{\Pi_w \Gamma\left(n_j^{(w)} + \beta\right)}{\Gamma\left(n_j^{(\cdot)} + W\beta\right)} \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^{\top} \int_{\phi} \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^{-T} \prod_{j=1}^{T} \frac{\Gamma\left(n_j^{(\cdot)} + W\beta\right)}{\Pi_w \Gamma\left(n_j^{(\omega)} + \beta\right)} \frac{\Gamma(\Sigma(\beta_i + \Sigma \mathbf{w_i}))}{\Pi\, \Gamma(\beta_i + \Sigma \mathbf{w_i})} \Pi \phi_i^{(\beta_i + \Sigma \mathbf{w_i}) - 1} d\phi
\end{aligned}
$$

implying that everything in the integral integrates to one. We do not find very much intuition to what we've found, but felt an attempt to fill in the wholes not presented within the literature mentioned might be useful, and could be worth returning to at a later point.

Note that $n_j^{(\cdot)}$ is the numer of times word w was assigned to topic j in vector of assignments $\mathbf{z}$ above.

Similarly, we note that $p(\mathbf{z})$ is the same as integrating ϕ out of the joint $p(\mathbf{z}, \theta) = p(\mathbf{z}|\theta)p(\theta)$, but just as before we have a Multinomial-Dirichlet conjugacy and we can procede in the same manner as before to obtain

$$
P(\mathbf{z}) = \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^D \prod_{d=1}^{D} \frac{\prod_j \Gamma\left(n_j^{(d)} + \alpha\right)}{\Gamma\left(n^{(d)} + T\alpha\right)},
$$

The authors then jump to provide the form for the full conditional for $p(z_i|\mathbf{z_{-i}}, \mathbf{w})$ such that

$$
P(z_i = j \mid \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,n}^{(d_i)} + T\alpha}
$$

Where at first glance we can notice a blend of the forms from $P(\mathbf{w}|\mathbf{z})$ and $P(\mathbf{z})$ on a univariate-like scale as we might expect since the full conditional is proportionate to the joint, and the joint is equal to the product of $P(\mathbf{w}|\mathbf{z})P(\mathbf{z})$. Given all of the $\mathbf{z_{-i}}$ are constants, we note that many terms can be cancelled out through the proportionate sign, however we see information from the words still included. Of course, this would make sense as we are thinking of the words as a likelihood function within our Bayesian machinery. Note that $n_{-i}^{(\cdot)}$ is the count that does not include current assignment ot z_i.

Moreover, the full conditional has an intuitive nature such that the left term can be interpreted as the probability of w_i under topic j, and the right term can be interpreted as the probability of topic j in document d_i. Of course, once our chain has converged, we can take our samples to be independent samples of the posterior topic marginals.

The authors note that this is actually a relatively effecient calculation since we just have to sum a relatively small number of nonzero counts.

Lastly, Griffiths et. al. then discuss the estimation of θ_j and ϕ_j. The authors note their estimates

$$
\begin{aligned}
\hat{\phi}_j^{(w)} &= \tfrac{n_j^{(w)} + \beta}{n_j^{(\cdot)} + W\beta} \\
\hat{\theta}_j^{(d)} &= \tfrac{n_j^{(d)} + \alpha}{n^{(d)} + T\alpha}
\end{aligned}
$$

correspond to the predictive distributions over new words w and new topics z, conditioned on $\mathbf{w}$ and $\mathbf{z}$, but leaves it for the reader to infer how. We illustrate the idea noting that $p(w_{+1}|\mathbf{w})$ is the same as integrating $\mathbf{z}$ out of $p(w_{+1}, \mathbf{z}|\mathbf{w})$ which is the same as integrating out $\mathbf{z}$ from $p(w_{+1}|\mathbf{z}, \mathbf{w})p(\mathbf{z}|\mathbf{w})$. Since we have the samples from $p(\mathbf{z}|\mathbf{w})$, and since we can sample from w_{+1} now that we are conditioning on $\mathbf{z}$, we can simply apply monte carlo sampling to approximate the posterior predictive distribution for w_{+1}. From here we can take the average of these samples to represent the estimate of the posterior predictive mean, which we are saying is $\hat{\phi}$. Of course, the estimates provided above are with respect to the jth $\hat{\phi}_j$ which requires the same procedure. Note that we can follow the same method to obtain the estimate $\hat{\theta}_j$.

Backtracking in time, Blei et. al. had an interest in $max\; p(\mathbf{w}|\phi, \alpha) = \int p(\mathbf{w}|\theta, \phi)p(\theta|\alpha)d(\theta)$ but noting its intractability resorted to variational bayes, whereas others have resorted to expectation propogation. Despite the slightly different aim of these authors from our objective, we still note that Blei et. al. discuss a key component surrounding the concept of exchangeability when calculating $p(\mathbf{w}, \mathbf{z})$. These authors claim that since the topics are infinitely exchangeable within a document in LDA, then by de Finetti's representation theorem, the joint distribution of the topics conditioned on $\boldsymbol{\theta}$ are then *independent and identically distributed*, such that since

$$p(z_1, \ldots, z_N) = p\big(z_{\pi(1)}, \ldots, z_{\pi(N)}\big)$$

then

$$p(\mathbf{w}, \mathbf{z}) = \int p(\theta) \left(\prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n) \right) d\theta$$

In other words, why we are able to express the joint as a product in the manner presented above. This notion of exchangeability and di Finetti's theorem was not mentioned within Griffiths, but we thought it important to discuss as it was a critical tool in developing the LDA model within Blei et. al. from which Griffiths et. al. built upon.

Clearly, the approach in Griffiths et. al. provides a huge speed up as we can integrate out the latent variables and simply sum a relatively small number of nonzero counts in our quest to obtain posterior samples. However, one of the disadvantages to this algorithm is the number of full conditionals that need to be sampled. Following the work from Griffiths et. al. we note that Porteous et. al. have developed a way to "reduce the inner-loop" or reducing the number of K topics to sample from to a much smaller number of full conditionals. The argument is based on the notion that "for any particular word and document, the sampling distributions of interst are frequently skewed such that most of the probability mass is concentrated on a small fraction of the total number of topics K." The authors provide an alogorithm "Fast LDA", which takes advantage of this concentrated probability notion mentioned above.

We draw slightly on the works of Porteous et. al., noting that the authors claim that this concentration occurs "for most real data sets after several iterations of the Gibbs sampler", and note their experiment with the New York Times Corpus, and the very different PubMed Corpus. Despite the very different nature of the Corpora, the findings were quite similar as illustrated below

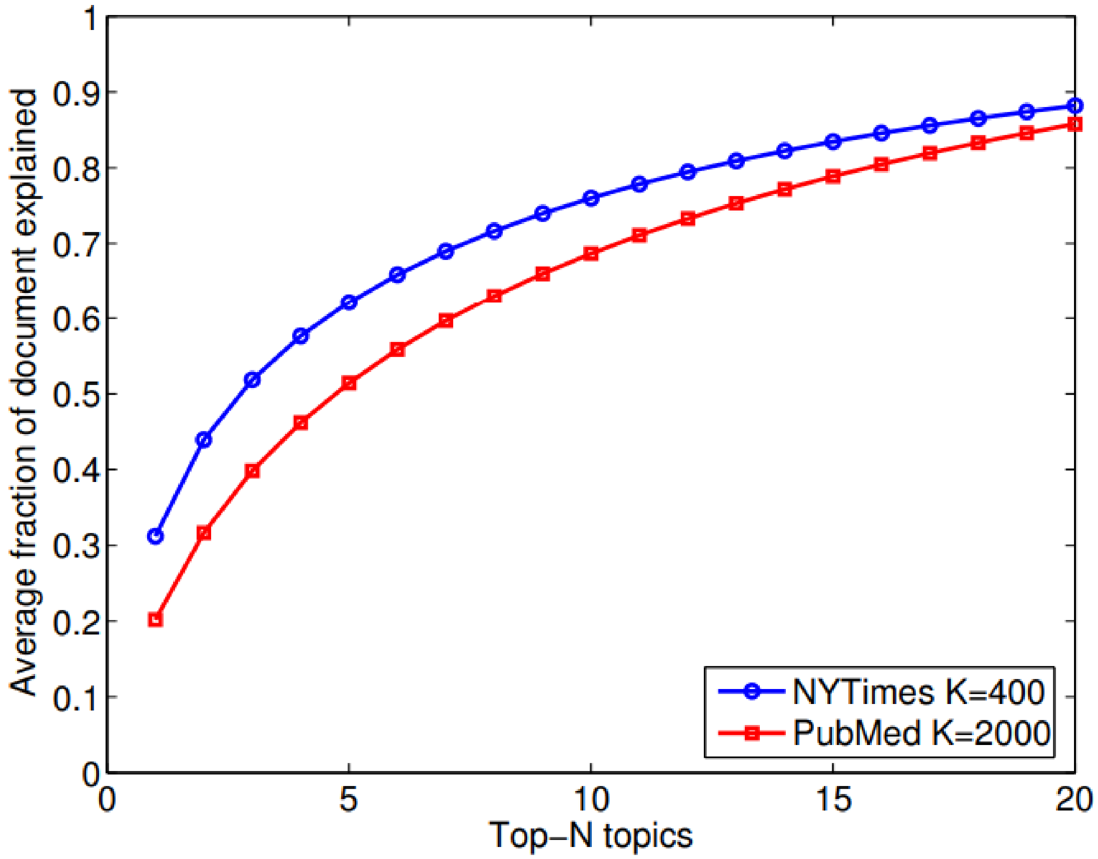


Figure 4: Average fraction of a document explained by top-20 topics, for NYTimes (K=400 topics) and PubMed (K=2000 topics). We see that, on average, the top-20 topics in any document account for approximately 90% of the words in the document.

We note that while our model draws slightly on Porteous et. al., we are manually setting our topic number which may not be the most advantageous aspect of our model. One might look into the BNP approaches that allow for allow for the number of topics to grow. Moreover, we still have to sample each topic whereas the Porteous et. al. model is drastically faster since it does not have to do this (and peforms even better with even larger K).

For Griffiths et. al. the purpose of the algorithm was centered on the title of the paper. That is, they were focused on “Finding Scientific Topics”. Griffiths et. al. use Bayesian model selection to establish the number of topics (where as we chose 20), and illustrate that the extracted topics from a corpora of abstracts from PNAS was capable of providing meaningful structure in the data consistent the class designations provided by the authors of the articles. This is quite useful for experts as a “first-order approximation to the kind of knowledge avialable” within a corpora of documents. In other words, it can help scientists decide whether or not to invest valuable time investigating a particular corpora.

Known possible applications also include identifying “hot topics” as Griffiths and Porteous point out by examining dynamics and tagging abstracts to illustrate semantic content.

These are concepts that could help anyone hoping to better expend their valuable research time. In this vein, whatever the topic area might be, it might be wise to begin the research process by first running an LDA model on a few relevant corpora before digging in one’s specified research.

One other way to utilize this method to learn generally about what topics are being covered in other fields. That is, for areas that might be too distant from one’s particular area of interest, and given finite time, one might augment their intellectual development by running this method on scientific domains completely separate from one’s particular area of interest.

Moreover, LDA methods are of particular relevance to genomic research and recommender systems.

3. Algorithm

4. Optimization for Performance

5. Applications to Simulated Data Sets

6. Applications to Real Data Sets

7. Comparative Analysis with Competing Algorithms

8. Discussion / Conclusion

9. References / Bibliography