A Comparison of Variational Bayes and Markov Chain Monte Carlo Methods for Topic Models

Hossein Soleimani
HSOLEIMANI@PSU.EDU
Department of Electrical Engineering
Pennsylvania State University

December 2014

Abstract

Latent Dirichlet Allocation (LDA) is Bayesian hierarchical topic model which has been widely used for discovering topics from large collections of unstructured text documents. Estimating posterior distribution of topics as well as topic proportions for each document is the goal of inference in LDA. Since exact inference is analytically intractable for LDA, we need to use approximate inference approaches such as mean-field variational Bayes or MCMC methods. In this paper, efficiency and computational challenges of these approximate inference methods have been compared by conducting a simulation study on two high dimensional data sets. Experimental results show that, MCMC methods perform better than variational Bayes in terms of test-set likelihood and classification accuracy. But, this gain in performance comes with a huge increase in computation time. Results also suggest that mean-field variational Bayes can achieve a reasonable performance despite the high dimensionality of the problem and the factorized form of variational distribution. The source code for all our experiments is available from `https://github.com/hsoleimani/STAT540Project`.

1. Introduction

Topic models are a class of statistical models designed to discover latent patterns ("topics") that prevail in a collection of unstructured discrete data [1],[2],[3],[4]. These models have been widely used to analyze images [5],[6] and biomedical data [7] or to discover underlying semantic structure of different types of text documents such as scientific abstracts [8],[2],[9], news archives [10],[11], and tweets [12],[13]. Throughout the rest of this paper, we only focus on modelling text documents.

In the context of text mining, topic models provide powerful algorithms to automatically organize massive collections of documents by discovering the main themes ("topics") that occur in the documents. Each topic is defined by specifying a pattern of words; i.e. the words that appear more or less frequently than others under that topic.

In this paper, we use a hierarchical Bayesian topic model called Latent Dirichlet Allocation (LDA) [2]. In LDA, each topic is a Multinomial distribution over the set of words in a dictionary. LDA posits that each document is a mixture of multiple topics with its own specific mixing proportions (i.e. topic proportions). The goal of inference in LDA is to estimate posterior distribution of the latent topics as well as topic proportions for each document.

Exact inference is intractable in LDA and therefore we need to appeal to approximate inference methods such as mean-field variational Bayes [14] or Markov chain Monte Carlo (MCMC) methods [8]. Mean-field variational methods approximate posterior distribution of unobserved variables by a variational distribution which factorizes over latent variables. Breaking conditional dependencies among latent variables makes it possible to analytically compute the approximate posterior, but on the other hand, it may introduce bias in approximation and may result in suboptimal solutions.

MCMC approaches however, sample from the *true* posterior distribution and hence do not suffer from the source of bias in variational methods. But, obtaining a reliable estimate of the posterior distributions using MCMC methods requires huge computational resources. Moreover, since topic models are very high dimensional, using MCMC methods requires special treatments for checking convergence of the Markov chain.

In this paper, we investigate efficiency and computational challenges of variational Bayes and MCMC methods for performing inference in LDA. For each method, we study potential sources of inefficiency and possible challenges imposed by high dimensionality of the problem. We compare these methods by conducting a simulation study on a synthetic data set as well as a subset of Reuters news articles.

The rest of the paper is organized as follows: In section 2, we briefly review LDA and describe its inference process using variational Bayes and MCMC methods. Section 3, describes the simulation settings, performance criteria, and the data sets used in our experiments. Simulation results are reported in section 4. Finally, future works are discussed in section 5.

2. Model

In this section, we first introduce the notation we use throughout and then briefly review LDA model and its inference procedure using variational Bayes and MCMC methods.

2.1 Notation and Terminology

We observe a corpus $\mathcal{D}$ which is a collection of D documents and a dictionary consisting of N unique words. We index unique documents and unique words by $d \in \{1, ..., D\}$ and $n \in \{1, ..., N\}$, respectively. Our goal is to discover M topics which we index each one of them by $j \in \{1, ..., M\}$. We denote probability of word n under topic j by β_{jn}, and we have $\sum_{n=1}^{N} \beta_{jn} = 1 \; \forall j = 1, 2, ..., M$.

Each document d, consists of L_d words and we denote *i-th* word in this document by $w_{id} \in \{1, ..., N\}$, $i = 1, ..., L_d$. The proportion of topic j in document d is denoted by θ_{jd} where $\sum_{j=1}^{M} \theta_{jd} = 1$. Each word i in document d has an M-dimensional binary random vector $z_{id} = (z_{id}^{(1)}, z_{id}^{(2)}, ..., z_{id}^{(M)})$ which has only a single element equal to one and all other elements equal to zero. The non-zero element in z_{id} (i.e. j s.t. $z_{id}^{(j)} = 1$) is the topic of origin for word w_{id}.

Throughout the paper, we show vectors with dots; e.g. $\theta_{\cdot d} = (\theta_{1d}, \theta_{2d}, ..., \theta_{Md})$ and $\beta_{j\cdot} = (\beta_{j1}, \beta_{j2}, ..., \beta_{jN})$.

2.2 Latent Dirichlet Allocation

LDA posits that documents in corpus $\mathcal{D}$ are generated based on the following generative process:

1. For each topic $j = 1, 2, ..., M$, generate $\beta_{j\cdot} \sim \text{Dirichlet}(\nu)$.
2. For each document $d = 1, 2, ..., D$:
   (a) Generate topic proportions $\theta_{\cdot d} \sim \text{Dirichlet}(\alpha)$.
   (b) For each word $i = 1, 2, ..., L_d$:
      i. Choose a topic $z_{id} \sim \text{Multinomial}(\theta_{\cdot d})$.
      ii. Choose a word from topic z_{id}; i.e. $w_{id} \sim \text{Multinomial}(\beta_{z_{id}\cdot})$.

In practice, we only observe $w_{id} \; \forall i, d$ and want to estimate $z_{id}^{(j)}, \theta_{jd}$, and $\beta_{jn} \; \forall i, j, d, n$. If the observed data is in fact generated based on this process, we can obtain the optimal estimate of the parameters via maximum likelihood estimation or by estimating the posterior distributions through a Bayesian approach. In this paper, we choose the latter framework and integrate out all latent variables $z_{id}^{(j)}, \theta_{jd}$, and $\beta_{jn} \; \forall i, j, d, n$. Note that, dimension of the latent variable space is equal to $MN + MD + \sum_{d=1}^{D} ML_d$ which can be very large in any real-world data set.

The Dirichlet distribution used in steps 1 and 2-(a) is the conjugate prior on Multinomial distribution and has the following density function $p(x_\cdot) = \frac{\Gamma(\sum_{n=1}^{N} \nu_n)}{\prod_{n=1}^{N} \Gamma(\nu_n)} \prod_{n=1}^{N} x_n^{\nu_n - 1}$. In this paper, we use symmetric Dirichlet distribution where $\nu_1 = \nu_2 = ... = \nu_N = \nu$. Accordingly, ν and α are known scalar hyper-parameters.

Following this generative process and by integrating out latent variables, we obtain the likelihood of corpus $\mathcal{D}$:

$$p(\mathcal{D}|\alpha,\nu) = \int \prod_{j=1}^{M} p(\beta_{j\cdot}|\nu) \prod_{d=1}^{D} \left[\int p(\theta_{\cdot d}|\alpha) \Big(\prod_{i=1}^{L_d} \sum_{z_{id}} p(z_{id}|\theta_{\cdot d}) p(w_{id}|z_{id},\beta) \Big) d\theta_{\cdot d} \right] d\beta. \tag{1}$$

This likelihood function is discussed in more details in Appendix A.

Computing (1) is analytically intractable due to the coupling between θ and β. Therefore, since (1) is the normalizing constant of the posterior of latent variables, we cannot compute the posteriors analytically. Throughout the rest of this section, we describe how we can use variational Bayes and MCMC methods for conducting approximate inference.

2.3 Variational Bayes Method

The goal of variational Bayes method is to obtain a lower bound on the log-likelihood function. Suppose for now that we only have one latent variable Z and we observe X. The observed log-likelihood function is $\log p(X) = \log \int p(X,z)dz = \log E_{p(z|X)}[p(X)]$. In variational Bayes, we introduce a variational distribution $q(z)$ on latent variables and convert the log-likelihood function into logarithm of an expectation with respect to $q(z)$; i.e. we have $\log p(X) = \log \int \frac{p(X,z)q(z)}{q(z)} dz = \log E_q[p(X,Z)q(Z)]$. Then, we invoke the Jensen's inequality to obtain the surrogate function Q [14]:

$$Q = E_q[\log p(X,Z)] - E_q[\log q(Z)] \leq \log p(X) \tag{2}$$

It can be easily shown that the difference between the lower bound Q and the true log-likelihood function is equal to the Kullback-Leibler (KL) divergence between the variational distribution and the true posterior $p(Z|X)$ [14]. The tightest possible lower bound Q is then obtained by minimizing the KL divergence $KL\big(q(Z)||p(Z|X)\big)$ with respect to parameters of the variational distribution.

In LDA, we introduce variational distributions on latent variables θ, β and z. Moreover, to obtain a tractable surrogate function and to remove the troubling conditional dependencies in the true model, we assume that the variational distribution factorizes over all latent variables. The proposed variational distribution is of the following form:

$$q(\beta,\theta,z|\mu,\gamma,\phi) = \prod_{j=1}^{M} q(\beta_{j\cdot}|\mu_{j\cdot}) \prod_{d=1}^{D} q(\theta_{\cdot d}|\gamma_{\cdot d}) \prod_{i=1}^{L_d} q(z_{id}|\phi_{id}), \tag{3}$$

where $q(\beta_{j\cdot}|\mu_{j\cdot}) = \text{Dirichlet}(\mu_{j1},\mu_{j2},...,\mu_{jN})$, $q(\theta_{\cdot d}) = \text{Dirichlet}(\gamma_{1d},\gamma_{2d},...,\gamma_{Md})$, and $q(z_{id}|\phi_{id}) = \text{Multinomial}(\phi_{id}^{(1)},\phi_{id}^{(2)},...,\phi_{id}^{(M)})$, and $\mu_{jn} > 0$, $\gamma_{jd} > 0$, and $\phi_{id}^{(j)} > 0$ s.t. $\sum_{j=1}^{M} \phi_{id}^{(j)} = 1$ $\forall i,j,d$ are variational parameters.

Equation (3) for all feasible values of the variational parameters, defines a flexible family of factorized distributions on the latent variables. By minimizing the KL divergence with respect to variational parameters, we find the distribution from this family that best approximates the true posterior.

By taking the derivative of the KL divergence and setting them to zero, we obtain closed-form updates for all variational parameters:

$$\mu_{jn} = \nu + \sum_{d=1}^{D} \sum_{i=1}^{L_d} \delta(w_{id} = n)\phi_{id}^{(j)}, \qquad \gamma_{jd} = \alpha + \sum_{i=1}^{L_d} \phi_{id}^{(j)},$$

$$\phi_{id}^{(j)} \propto \exp\left(\Psi(\gamma_{jd}) - \Psi(\sum_{j'=1}^{M} \gamma_{j'd}) + \Psi(\mu_{jw_{id}}) - \Psi(\sum_{n=1}^{N} \mu_{nd}) \right), \quad \sum_{j=1}^{M} \phi_{id}^{(j)} = 1, \tag{4}$$

where $\Psi(\cdot)$ is the first derivative of the $\log \Gamma(\cdot)$ function and $\delta(x)$ is the Kronecker delta function which is equal to 1 if $x = 0$ and 0 otherwise. Derivation of the surrogate function is discussed in Appendix B.

Since the update equation for each variational parameter depends on the value of others, we need an iterative algorithm to cycle over all variational parameters and update them one by one. This iterative process is described in Algorithm 1. Convergence in this algorithm is achieved if the relative change in Q is less 10^{-4}. Note that, there are two nested loops (on the whole corpus and on each document) in Algorithm 1 and each are terminated when the convergence criterion is met. At every iteration of the algorithm, we re-initialize ($\gamma_{\cdot d}$ and $\phi_{\cdot d}$) uniformly for each document while keeping μ fixed from the previous iteration. Experimentally, we have found that this helps to avoid poor local optima and achieve higher likelihood.

After convergence, we obtain variational estimate of the posterior mean of latent variables as follows: $\hat{\theta}_{jd} \approx \frac{\gamma_{jd}}{\sum_{j'=1}^{M} \gamma_{j'd}}$ $\forall j, d$, $\hat{\beta}_{jn} \approx \frac{\mu_{jn}}{\sum_{n'=1}^{N} \mu_{jn'}}$, $\forall j, n$, and $E\hat{z}_{id}^{(j)} \approx \phi_{id}^{(j)}$, $\forall i, j, d$.

Algorithm 1 Variational Bayes

```
Initialize $\mu_{jn}$ $\forall j, n$.
repeat
    for $d = 1$ to $D$ do
        Initialize $\phi^{j}_{id} = 1/M$ $\forall i, j$.
        Initialize $\gamma_{jd} = \alpha + L_d/M$ $\forall j$.
        repeat
            Update $\phi^{(j)}_{id}$ $\forall j$.
            Update $\gamma_{jd}$ $\forall j$.
        until convergence
    end for
    Update $\mu_{jn}$ $\forall j, n$.
until convergence
```

Algorithm 2 Gibbs Sampling

```
Initialize $\beta_{jn}, \theta_{jd}, z_{id}$ $\forall i, j, d, n$.
repeat
    for $d = 1$ to $D$ do
        Sample $z_{id}$ $\forall i$.
        Sample $\theta_{\cdot d}$.
    end for
    Sample $\beta_{j\cdot}$ $\forall j$.
until convergence
```

Algorithm 3 Collapsed Gibbs Sampling

```
Initialize $z_{id}$ $\forall i, d$.
repeat
    for $d = 1$ to $D$ do
        Sample $z_{id}$ $\forall i$.
    end for
until convergence
```

2.4 Markov Chain Monte Carlo Method

In this section, we describe how we can estimate posterior mean of the latent variables using a Markov Chain Monte Carlo approach. We have Gibbs update for sampling each variable given other latent variables and the corpus $\mathcal{D}$:

$$\begin{aligned} p(\beta_{j\cdot}|\beta_{-j\cdot}, z, \theta, \mathcal{D}) &= \text{Dirichlet}(\nu + t_{j1}, ..., \nu + t_{jN}) && (5)\\ p(\theta_{\cdot d}|\theta_{-\cdot d}, z, \beta, \mathcal{D}) &= \text{Dirichlet}(\alpha + m_{1d}, ..., \alpha + m_{Md}) && (6)\\ p(z_{id}|z_{-id}, \theta, \beta, \mathcal{D}) &= \text{Multinomial}(\theta_{1d}\beta_{1w_{id}}, ..., \theta_{Md}\beta_{Mw_{id}}) && (7) \end{aligned}$$

where $t_{jn} = \sum_{d=1}^{D} \sum_{i=1}^{L_d} \delta(w_{id} = n) z_{id}^{(j)}$ and $m_{jd} = \sum_{i=1}^{L_d} z_{id}^{(j)}$. The "$-$" subscript indicates all variables excluding the one from which we sample at the current step. The sampling algorithm is described in Algorithm 2. Detailed derivation of the Gibbs updates are provided in Appendix C.

Because of the high dimensionality of the problem, handling all the samples that we generate from the Markov chain is a challenging task. Here, instead of keeping all the original samples in memory, at each step of the algorithm, we recursively update the posterior mean estimate for each parameter. Also, we need a mechanism to automatically check convergence of the Markov chain. In order to do this, we use a modified version of the batch means algorithm [15] to *recursively* compute Monte Carlo standard error. The recursive nature of our computation requires us to set the batch length in the batch means algorithm initially and keep it fixed throughout the algorithm. Here, the batch length is set to 50 for all variables. Using this recursive method, we compute Monte Carlo standard error for each latent variable and stop the sampling algorithm when the maximum of the Monte Carlo standard errors of all latent variables is less than 0.03.

2.4.1 Collapsed Gibbs Sampling

We can further simplify the Gibbs sampling algorithm by performing collapsed Gibbs sampling [16] in which we integrate out θ and β and only sample from z. This will slightly reduce the computational cost and runtime of the sampling algorithm. However, since z has the highest dimension among all latent variables ($\sum_{d=1}^{D} L_d M$), this method will have similar computational challenges to the full Gibbs sampling method. By sampling only from z, we can still estimate posterior mean of θ and β by $\frac{m_{jd}}{\sum_{j'=1}^{M} m_{j'd}}$ $\forall j, d$ and $\frac{t_{jn}}{\sum_{n'=1}^{N} t_{jn'}}$ $\forall j, n$, respectively.

For each word i in document d, random vector z_{id} conditioned on other latent variables is a Multinomial random variable:

$$p(z_{id}|z_{-id}, \mathcal{D}) = \text{Multinomial}(s_{id}^{(1)}, ..., s_{id}^{(M)}), \tag{8}$$

where $s_{id}^{(j)} = \frac{\nu + t_{jw_{id},-i}}{\nu N + \sum_{n'=1}^{N} t_{jn',-i}} (\alpha + m_{jd,-i})$, $t_{jw_{id},-i} = t_{jw_{id}} - z_{id}^{(j)-}$, $m_{jd,-i} = m_{jd} - z_{id}^{(j)-}$, and $z_{id}^{(j)-}$ is the value of $z_{id}^{(j)}$ at the previous iteration. The sampling algorithm is described in Algorithm 3. Appendix D derives the collapsed Gibbs update for z.

3. Simulation Settings

We conduct simulation studies to compare efficiency of variational Bayes and MCMC methods on synthetic data and a subset of Reuters news articles. On synthetic data, we generate documents from a known distribution and on the Reuters corpus we use parametric bootstrap to compare different methods.

3.1 Performance Criteria

We compare variational Bayes and MCMC methods with respect to four performance criteria:

- Test set log-likelihood: We estimate posterior mean of θ and β in the inference step. Then, we compute log-likelihood of the test documents using: $\sum_{d=1}^{D_{test}} \sum_{i=1}^{L_d} \log \big(\sum_{j=1}^{M} \hat{\theta}_{jd} \hat{\beta}_{jw_{id}} \big)$.

- Test set classification accuracy: Each document d in our data sets has a ground-truth class label $c_d \in \{1, 2, ..., C\}$. Using these class labels and posterior mean of topic proportions on the training set, we learn a multinomial class label distribution for each topic j; i.e. $p_j(c) = \frac{\sum_{d=1}^{D} \delta(c_d = c) \hat{\theta}_{jd}}{\sum_{c'=1}^{C} \sum_{d=1}^{D} \delta(c_d = c') \hat{\theta}_{jd}}$ $\forall c \in \{1, 2, ..., C\}$. Then, we assign the label with highest probability $\hat{c}_d = \underset{c}{\text{argmax}} \sum_{j=1}^{M} \hat{\theta}_{jd} p_j(c)$ to each test document and compute correct classification rate on the test set. High classification accuracy indicates that the discovered topics highly match the ground-truth class labels.

- MSE of topic proportions of the test documents

- Runtime: We also compare training time of different methods.

3.2 Synthetic Data[1]

The synthetic data set consists of, respectively, 2000 and 1000 documents in the training and test sets. The documents are generated by choosing words from a dictionary of 2000 unique words and 10 topics. Under each topic, 1% of the words are randomly selected as "high probability" words whose probability parameter is then selected from Uniform$(0.7, 0.8)$. Probability parameters of the rest of the words are generated from Uniform$(0, 0.1)$. Then, word probabilities under each topic are normalized to form a Multinomial distribution. For each document, one topic is chosen as the dominant topic whose topic proportion is set to 50/59 while topic proportions of the rest of the topics are set to 1/59. Each topic is dominant in respectively, 200 and 100 training and test documents. Each document is labelled as the index of its dominant topic. Words in the documents are then generated using the generative process of LDA described in section 2.2.

LDA hyper-parameters ν and α are both set to 0.1. For each method, we generate a batch of synthetic data, conduct the inference step, and measure the performance criteria. We repeat this process on T batches of synthetic data. T for each inference method is chosen such that the Monte Carlo standard error of classification accuracy is less than 0.05.

3.3 Reuters Corpus

We also compare methods on a small subset of the Reuters[2] data set. The corpus used here has respectively, 5214 and 2069 documents in the training and test sets. Each document is labelled with a single ground-truth label, and in total there are 8 class labels. Accordingly, we set number of topics in LDA to 8. In total, there are 6468 unique words in the dictionary.

On this data set, we use a parametric bootstrap to compare variational Bayes and MCMC methods. For each method, we first estimate posterior mean of β and θ and then use them to generate T bootstrap samples. Bootstrap sample size is selected to obtain 0.05 Monte Carlo standard error on classification accuracy. LDA hyper-parameters ν and α are both set to 0.1

4. Results

Figure 1 shows the results of our simulation study. The first and second rows of Figure 1 are respectively, results on the synthetic and the Reuters data sets. Experimental results show that on both data sets, the Gibbs sampling methods perform better than variational Bayes. Gibbs sampling and collapsed Gibbs sampling methods achieve almost 99% accuracy on the synthetic data and higher than 84% on the Reuters data set. On both data sets, variational Bayes achieves lower classification accuracy than the Gibbs sampling methods. This pattern is almost the same in all other performance criteria. Gibbs sampling methods achieve lower MSE and higher likelihood than the variational Bayes method.

This result is not unexpected, since unlike variational Bayes approach, there is no approximation in sampling methods. It is possible to achieve accurate estimates of the posterior distributions with an arbitrary standard error if the Markov Chain is run long enough. However, this gain in performance comes with a huge computational cost. As shown on the right column in Figure 1, runtime for Gibbs sampling methods is much higher than variational Bayes. This can potentially restrict applicability of MCMC methods in higher dimensional problems specially when Gibbs updates are not available.

Variational Bayes method on the other hand, achieves reasonable performance on both data sets at a much smaller computation time. This in fact suggests that despite the factorized form of the variational distribution, by minimizing Kullback-Leibler divergence, variational Bayes can

[1] The python code to generate synthetic documents is available from https://github.com/hsoleimani/STAT540Project/blob/master/synthetic/datagen.py

[2] https://github.com/hsoleimani/STAT540Project/tree/master/Reuters/data. The original big corpus is available from http://www.daviddlewis.com/resources/testcollections/reuters21578/

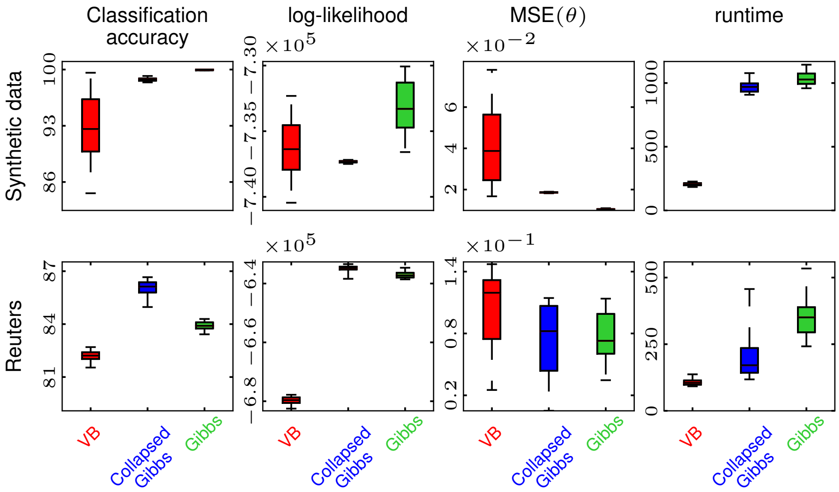

Figure 1: Results of comparison between variational Bayes and MCMC methods on synthetic and Reuters data sets.

still provide a sensible approximation of the true posterior. The main benefit of variational Bayes compared to MCMC is again the small runtime which is more critical in bigger data sets.

We can also see that performing collapsed Gibbs sampling instead of the full Gibbs sampling reduces the runtime. But, the runtime for collapsed Gibbs sampling is still much higher than the runtime of variational Bayes. Also, note that, despite the high dimensionality of the problem, the data is very sparse; each word on average appears in respectively 29 and 95 documents in the Reuters and synthetic data sets. Because of this sparsity, Gibbs sampling methods are practically feasible and the Markov chain can in fact converge in reasonable amount of time.

Figure 1 also shows that the variational Bayes method has larger variance in almost all performance criteria. This is partly due to the sensitivity of this method to initialization of variational parameters. A poor initialization may cause the algorithm to be trapped in poor local optima. Gibbs sampling methods on the other hand, are less sensitive to initialization and can effectively explore all the latent variable space.

Besides smaller runtime, another potential benefit of variational Bayes over MCMC methods is that it provides an analytical form for the objective function which can then be optimized to find the best value of hyper-parameters. In this project, we fixed the hyper-parameters of LDA ν and α to some known numbers. But in practice, these hyper-parameters can hugely affect performance of LDA and quality of discovered topics. In variational Bayes, we can easily optimize the surrogate function via any derivative-based optimization technique and find the optimal value of hyper-parameters. But this is more difficult in MCMC methods. Method of moments has been suggested as a way to estimate hyper-parameters in MCMC methods [17].

5. Future work

In a simple topic model such as LDA, Gibbs updates are available for all variables in an MCMC approach. However, in slightly more complicated models, we should use the general Metropolis-Hastings algorithm. In such models, it is more complicated and challenging to use MCMC methods. Comparison of MCMC methods and variational Bayes in these situations is an interesting future direction of this work. Also, studying how variational Bayes and MCMC methods scale as the size of data increases is another aspect to consider in future works.

References

[1] D. Blei, L. Carin, and D. Dunson, "Probabilistic Topic Models," *IEEE Signal Processing Magazine*, pp. 77–84, Nov. 2010.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[3] H. Soleimani and D. J. Miller, "Parsimonious Topic Models with Salient Word Discovery," *Knowledge and Data Engineering, IEEE Transaction on*, vol. 27, pp. 824–837, 2015.

[4] H. Soleimani and D. J. Miller, "Sparse topic models by parameter sharing," in *Machine Learning for Signal Processing (MLSP), 2014 IEEE International Workshop on*, pp. 1–6, IEEE, 2014.

[5] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, pp. 524–531, 2005.

[6] J. Sivic and B. Russell, "Discovering objects and their location in images," in *Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005*, pp. 370–377, 2005.

[7] J. K. Pritchard, M. Stephens, and P. Donnelly, "Inference of population structure using multilocus genotype data.," *Genetics*, vol. 155, pp. 945–59, June 2000.

[8] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101 Suppl, pp. 5228–35, Apr. 2004.

[9] E. M. Talley, D. Newman, D. Mimno, B. W. Herr, H. M. Wallach, G. A. P. C. Burns, A. G. M. Leenders, and A. McCallum, "Database of NIH grants using machine-learned categories and graphical clustering.," *Nature methods*, vol. 8, pp. 443–4, June 2011.

[10] X. Wei and W. Croft, "LDA-based document models for ad-hoc retrieval," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 178–185, 2006.

[11] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, (New York, New York, USA), p. 497, ACM Press, June 2009.

[12] D. Ramage, S. Dumais, and D. Liebling, "Characterizing Microblogs with Topic Models.," in *Fourth International AAAI Conference on Weblogs and Social Media*, pp. 130–137, 2010.

[13] L. Hong and B. Davison, "Empirical study of topic modeling in twitter," in *Proceedings of the First Workshop on Social Social Media Analytics*, pp. 80–88, ACM, 2010.

[14] M. Jordan, *Learning in Graphical Models:*. Boston, MA, USA: Kluwer Academic Publishers, 1998.

[15] G. Jones and M. Haran, "Fixed-width output analysis for Markov chain Monte Carlo," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1537–1547, 2006.

[16] R. Neal, "Markov chain sampling methods for Dirichlet process mixture models," *Journal of computational and graphical statistics*, vol. 9, no. 2, pp. 249–265, 2000.

[17] T. P. Minka, "Estimating a Dirichlet distribution." http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/minka-dirichlet.pdf, 2012.

Appendix A. Latent Dirichlet Allocation

In this appendix, we describe details of likelihood function derivation in LDA. Given topics β, joint probability of w, z, θ for each document is:

$$\begin{aligned} p(w_{\cdot d}, z_{\cdot d}, \theta_{\cdot d}|\beta,\alpha) &= p(\theta_{\cdot d}|\alpha)\prod_{i=1}^{L_d} p(z_{id}|\theta_{\cdot d})p(w_{id}|z_{id},\beta) \\ &= \left[\frac{\Gamma(\alpha M)}{\Gamma(\alpha)^M}\prod_{j=1}^{M}\theta_{jd}^{\alpha-1}\right]\left[\prod_{i=1}^{L_d}\prod_{j=1}^{M}\left(\theta_{jd}\beta_{jw_{id}}\right)^{z_{id}^{(j)}}\right] \end{aligned} \tag{9}$$

Complete data likelihood is then computed by taking product of (9) for all documents and multiplying by probability of β:

$$\begin{aligned} p(\mathcal{D}, z, \theta, \beta|\alpha,\nu) &= \prod_{j=1}^{M} p(\beta_{j\cdot}|\nu)\prod_{d=1}^{D} p(w_{\cdot d}, z_{\cdot d}, \theta_{\cdot d}|\beta,\alpha) \\ &= \prod_{j=1}^{M} p(\beta_{j\cdot}|\nu)\prod_{d=1}^{D}\left[p(\theta_{\cdot d}|\alpha)\prod_{i=1}^{L_d} p(z_{id}|\theta_{\cdot d})p(w_{id}|z_{id},\beta)\right] \\ &= \prod_{j=1}^{M}\left[\frac{\Gamma(\nu N)}{\Gamma(\nu)^N}\prod_{n=1}^{N}\beta_{jn}^{\nu-1}\right]\prod_{d=1}^{D}\left[\left[\frac{\Gamma(\alpha M)}{\Gamma(\alpha)^M}\prod_{j=1}^{M}\theta_{jd}^{\alpha-1}\right]\left[\prod_{i=1}^{L_d}\prod_{j=1}^{M}\left(\theta_{jd}\beta_{jw_{id}}\right)^{z_{id}^{(j)}}\right]\right] \end{aligned} \tag{10}$$

Observed data likelihood (1) is computed by integrating out latent variables from (10).

Appendix B. Variational Bayes

In this appendix, we derive the surrogate function used in variational inference. The variational distributions defined in (3) are:

$$q(\beta_{j\cdot}|\mu_{j\cdot}) = \frac{\Gamma(\sum_{n=1}^{N}\mu_{jn})}{\prod_{n=1}^{N}\Gamma(\mu_{jn})}\prod_{n=1}^{N}\beta_{jn}^{\mu_{jn}-1}, \quad \forall j \tag{11}$$

$$q(\theta_{\cdot d}|\gamma_{\cdot d}) = \frac{\Gamma(\sum_{j=1}^{M}\gamma_{jd})}{\prod_{j=1}^{M}\Gamma(\gamma_{jd})}\prod_{j=1}^{M}\theta_{jd}^{\gamma_{jd}-1}, \quad \forall d \tag{12}$$

$$q(z_{id}|\phi_{id}) = \prod_{j=1}^{M}(\phi_{id}^{(j)})^{z_{id}^{(j)}}, \quad \forall i, d. \tag{13}$$

The surrogate function is computed by:

$$\begin{aligned} Q &= E_q[\log p(\mathcal{D}, z, \theta, \beta|\alpha,\nu)] - E_q[\log q(z,\theta,\beta|\mu,\gamma,\phi)] \\ &= \sum_{j=1}^{M} E_q \log p(\beta_{j\cdot}|\nu) + \sum_{d=1}^{D}\left[E_q \log p(\theta_{\cdot d}|\alpha) + \sum_{i=1}^{L_d}\left(E_q \log p(z_{id}|\theta_{\cdot d}) + E_q \log p(w_{id}|z_{id},\beta)\right)\right] \\ &- \sum_{j=1}^{M} E_q \log q(\beta_{j\cdot}|\mu_{j\cdot}) - \sum_{d=1}^{D}\left[E_q \log q(\theta_{\cdot d}|\gamma_{\cdot d}) + \sum_{i=1}^{L_d} E_q \log q(z_{id}|\phi_{id})\right]. \end{aligned} \tag{14}$$

We compute each term in (14) by taking expectation with respect to the variational distribution. For $\beta_{j\cdot}$ which is a Dirichlet random variable, we have:

$$\begin{aligned} E_q \log p(\beta_{j\cdot}|\nu) &= E_q \log \Gamma(\nu N) - N E_q \log \Gamma(\nu) + \sum_{n=1}^{N} (\nu - 1) E_q \log \beta_{jn} \\ &= \log \Gamma(\nu N) - N \log \Gamma(\nu) + \sum_{n=1}^{N} (\nu - 1)\big(\Psi(\mu_{jn}) - \Psi(\sum_{n'=1}^{N} \mu_{jn'})\big), \end{aligned} \tag{15}$$

where we used $E_q \log \beta_{jn} = \Psi(\mu_{jn}) - \Psi(\sum_{n'=1}^{N} \mu_{jn'})$. Accordingly, for all other Dirichlet random variables in (14) we have:

$$E_q \log p(\theta_{\cdot d}|\alpha) = \log \Gamma(\alpha M) - M \log \Gamma(\alpha) + \sum_{j=1}^{M} (\alpha - 1)\big(\Psi(\gamma_{jd}) - \Psi(\sum_{j'=1}^{M} \gamma_{j'd})\big), \tag{16}$$

$$E_q \log q(\theta_{\cdot d}|\gamma_{\cdot d}) = \log \Gamma(\sum_{j=1}^{M} \gamma_{jd}) - \sum_{j=1}^{M} \log \Gamma(\gamma_{jd}) + \sum_{j=1}^{M} (\gamma_{jd} - 1)\big(\Psi(\gamma_{jd}) - \Psi(\sum_{j'=1}^{M} \gamma_{j'd})\big), \tag{17}$$

$$E_q \log q(\beta_{j\cdot}|\mu_{j\cdot}) = \log \Gamma(\sum_{n=1}^{N} \mu_{jn}) - \sum_{n=1}^{N} \log \Gamma(\mu_{jn}) + \sum_{n=1}^{N} (\mu_{jn} - 1)\big(\Psi(\mu_{jn}) - \Psi(\sum_{n'=1}^{N} \mu_{jn'})\big) \tag{18}$$

For the multinomial random variable z, we have:

$$\begin{aligned} E_q \log p(z_{id}|\theta_{\cdot d}) &= \sum_{j=1}^{M} [E_q z_{id}^{(j)}][E_q \log \theta_{jd}] \\ &= \sum_{j=1}^{M} \phi_{id}^{(j)} \big(\Psi(\gamma_{jd}) - \Psi(\sum_{j'=1}^{M} \gamma_{j'd})\big), \end{aligned} \tag{19}$$

where we used the fact that z_{id} and $\theta_{\cdot d}$ are independent under variational distribution q. This in fact is one instance where the factorized form of variational distribution helps to compute the surrogate function analytically. Similarly, we have:

$$E_q \log q(z_{id}|\phi_{id}) = \sum_{j=1}^{M} [E_q z_{id}^{(j)}][E_q \log \phi_{id}^{(j)}] = \sum_{j=1}^{M} \phi_{id}^{(j)} \log \phi_{id}^{(j)}. \tag{20}$$

Finally,

$$E_q \log p(w_{id}|z_{id}, \beta) = \sum_{j=1}^{M} E_q z_{id}^{(j)} E_q \log \beta_{j w_{id}} = \sum_{j=1}^{M} \phi_{id}^{(j)} \big(\Psi(\mu_{j w_{id}}) - \Psi(\sum_{n'=1}^{N} \mu_{jn'})\big). \tag{21}$$

By substituting (15)-(21), we obtain the surrogate function Q which can then be maximized with respect to the variational parameters to find the tightest possible lower bound for the observed log-likelihood function.

Appendix C. Gibbs Sampling

In this section, we derive the Gibbs updates for all latent variables.

From (10), we can see that posterior of $\beta_{j\cdot}$ conditioned on other variables is:

$$\begin{aligned} p(\beta_{j\cdot}|z,\beta_{-j\cdot},\theta,\mathcal{D}) &\propto \prod_{n=1}^{N}\beta_{jn}^{\nu-1}\prod_{d=1}^{D}\prod_{i=1}^{L_d}\beta_{jw_{id}}^{z_{id}^{(j)}} \\ &\propto \prod_{n=1}^{N}\beta_{jn}^{\nu-1+\sum_{d=1}^{D}\sum_{i=1}^{L_d}\delta(w_{id}=n)z_{id}^{(j)}} = \prod_{n=1}^{N}\beta_{jn}^{\nu+t_{jn}-1} \\ p(\beta_{j\cdot}|z,\beta_{-j\cdot},\theta,\mathcal{D}) &= \text{Dirichlet}(\nu+t_{j1},\nu+t_{j2},...,\nu+t_{jN}). \end{aligned} \tag{22}$$

Similarly, posterior of $\theta_{\cdot d}$ is:

$$\begin{aligned} p(\theta_{\cdot d}|z,\theta_{-\cdot d},\beta,\mathcal{D}) &\propto \prod_{j=1}^{M}\theta_{jd}^{\alpha-1}\prod_{i=1}^{L_d}\theta_{jd}^{z_{id}^{(j)}} \\ &\propto \prod_{j=1}^{M}\theta_{jd}^{\alpha-1+\sum_{i=1}^{L_d}z_{id}^{(j)}} = \prod_{j=1}^{M}\theta_{jd}^{\alpha+m_{jd}-1} \\ p(\theta_{\cdot d}|z,\theta_{-\cdot d},\beta,\mathcal{D}) &= \text{Dirichlet}(\alpha+m_{1d},\alpha+m_{2d},...,\alpha+m_{Md}). \end{aligned} \tag{23}$$

Finally, posterior of z_{id} conditioned on other variables is:

$$\begin{aligned} p(z_{id}|z_{-id},\beta,\theta,\mathcal{D}) &\propto \prod_{j=1}^{M}\left(\theta_{jd}\beta_{jw_{id}}\right)^{z_{id}^{(j)}} \\ p(z_{id}|z_{-id},\beta,\theta,\mathcal{D}) &= \text{Multinomial}(\theta_{1d}\beta_{1w_{id}},\theta_{2d}\beta_{2w_{id}},...,\theta_{Md}\beta_{Mw_{id}}). \end{aligned} \tag{24}$$

Appendix D. Collapsed Gibbs Sampling

Here, we derive updates for the collapsed Gibbs sampling algorithm. Using our definition of m_{jd} and t_{jn}, we can write (10) as:

$$p(\mathcal{D},z,\theta,\beta|\alpha,\nu) = \prod_{j=1}^{M}\Big[\frac{\Gamma(\nu N)}{\Gamma(\nu)^N}\prod_{n=1}^{N}\beta_{jn}^{\nu+t_{jn}-1}\Big]\prod_{d=1}^{D}\Big[\frac{\Gamma(\alpha M)}{\Gamma(\alpha)^M}\prod_{j=1}^{M}\theta_{jd}^{\alpha+m_{jd}-1}\Big]. \tag{25}$$

By taking integral of (25) with respect to θ and β, we have:

$$\begin{aligned} p(w,z|\alpha,\nu) &= \prod_{j=1}^{M}\Big[\frac{\Gamma(\nu N)}{\Gamma(\nu)^N}.\frac{\prod_{n=1}^{N}\Gamma(\nu+t_{jn})}{\Gamma(\nu N+\sum_{n=1}^{N}t_{jn})}\Big]\prod_{d=1}^{D}\Big[\frac{\Gamma(\alpha M)}{\Gamma(\alpha)^M}.\frac{\prod_{j=1}^{M}\Gamma(\alpha+m_{jd})}{\Gamma(\alpha M+\sum_{j=1}^{M}m_{jd})}\Big] \\ &= p(w|z,\nu)p(z|\alpha), \end{aligned} \tag{26}$$

where $p(w|z,\nu)=\prod_{j=1}^{M}\left[\frac{\Gamma(\nu N)}{\Gamma(\nu)^N}.\frac{\prod_{n=1}^{N}\Gamma(\nu+t_{jn})}{\Gamma(\nu N+\sum_{n=1}^{N}t_{jn})}\right]$ and $p(z|\alpha)=\prod_{d=1}^{D}\left[\frac{\Gamma(\alpha M)}{\Gamma(\alpha)^M}.\frac{\prod_{j=1}^{M}\Gamma(\alpha+m_{jd})}{\Gamma(\alpha M+\sum_{j=1}^{M}m_{jd})}\right]$.

Here, for simplicity, we change the notation for $\mathcal{D}$ and denote all words in the documents in $\mathcal{D}$ by w.

Then, posterior of $z_{id}^{(j)}$ conditioned on other variables is:

$$\begin{aligned}
p(z_{id}^{(j)}|z_{-id}, w, \alpha, \nu) &= \frac{p(z, w|\alpha, \nu)}{p(z_{-id}, w|\alpha, \nu)} = \frac{p(w|z, \nu)p(z|\alpha)}{p(w_{-id}|z_{-id}, \nu)p(w_{id})p(z_{-id}|\alpha)} \ (*) \\
&\propto \frac{p(w|z, \nu)}{p(w_{-id}|z_{-id}, \nu)} \frac{p(z|\alpha)}{p(z_{-id}|\alpha)} \ (**) \\
&\propto \frac{\Gamma(\nu + t_{jw_{id}})}{\Gamma(\nu N + \sum_{n'=1}^{N} t_{jn'})} \frac{\Gamma(\nu N + \sum_{n'=1}^{N} t_{jn',-i})}{\Gamma(\nu + t_{jw_{id},-i})} \frac{\Gamma(\alpha + m_{jd})}{\Gamma(\alpha M + \sum_{j'=1}^{M} m_{j'd})} \frac{\Gamma(\alpha M + \sum_{j'=1}^{M} m_{j'd,-i})}{\Gamma(\alpha + m_{jd,-i})},
\end{aligned}$$

where in (*), we used the fact that each word w_{id} is independent of others conditioned on z_{id}. Also, in (**) we drop $p(w_{id})$ since it is independent of z_{id}. Then, we use the property of Gamma function ($\Gamma(x + 1) = x\Gamma(x)$) and write:

$$p(z_{id}^{(j)}|z_{-id}, w, \alpha, \nu) \propto \frac{\nu + t_{jw_{id},-i}}{\nu N + \sum_{n'=1}^{N} t_{jn',-i}} \frac{\alpha + m_{jd,-i}}{\alpha M + \sum_{j'=1}^{M} m_{j'd,-i}}. \tag{27}$$

Therefore,

$$\begin{aligned}
p(z_{id}|z_{-id}, w, \alpha, \nu) &= \prod_{j=1}^{M} \left[\frac{(\nu + t_{jw_{id},-i})(\alpha + m_{jd,-i})}{\nu N + \sum_{n'=1}^{N} t_{jn',-i}} \right]^{z_{id}^{(j)}}, \\
p(z_{id}|z_{-id}, w, \alpha, \nu) &= \text{Multinomial}(s_{id}^{(1)}, s_{id}^{(2)}, ..., s_{id}^{(M)}),
\end{aligned} \tag{28}$$

where

$$s_{id}^{(j)} = \frac{(\nu + t_{jw_{id},-i})(\alpha + m_{jd,-i})}{\nu N + \sum_{n'=1}^{N} t_{jn',-i}}. \tag{29}$$