A Simple Stochastic Gradient Variational Bayes for the Correlated Topic Model

Tomonari Masada[1](✉) and Atsuhiro Takasu[2]

[1] Nagasaki University, 1-14 Bunkyo-machi, Nagasaki, Japan
masada@nagasaki-u.ac.jp
[2] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
takasu@nii.ac.jp

Abstract. This paper proposes a new inference for the correlated topic model (CTM) [3]. CTM is an extension of LDA [4] for modeling correlations among latent topics. The proposed inference is an instance of the stochastic gradient variational Bayes (SGVB) [7,8]. By constructing the inference network with the diagonal logistic normal distribution, we achieve a simple inference. Especially, there is no need to invert the covariance matrix explicitly. We performed a comparison with LDA in terms of predictive perplexity. The two inferences for LDA are considered: the collapsed Gibbs sampling (CGS) [5] and the collapsed variational Bayes with a zero-order Taylor expansion approximation (CVB0) [1]. While CVB0 for LDA gave the best result, the proposed inference achieved the perplexities comparable with those of CGS for LDA.

1 Introduction

Topic modeling is one of the outstanding text mining techniques that are based on unsupervised machine learning and have a wide variety of applications. After the proposal of LDA [4], many extensions are provided by considering more realistic situations. Especially, LDA cannot model correlations among latent topics. Therefore, the correlated topic model (CTM) has been proposed [3]. However, the inference for CTM is a bit complicated, because the logistic normal prior is not conjugate to the multinomial distribution. This paper proposes a new variational Bayesian inference for CTM. The main contribution is that we make the inference simple with the stochastic gradient variational Bayes (SGVB) [7,8]. While the proposed inference adopts a gradient-based optimization similar to the original one [3], no explicit inversion of the covariance matrix is required.

We briefly describe the variational inference for topic models. Let $\boldsymbol{x}_d = \{x_{d1}, \ldots, x_{dN_d}\}$ be the multiset of the words in the document d. z_{dn} denotes the topic to which the word token x_{dn} is assigned. Then the log evidence of the document d is lower bounded as $\log p(\boldsymbol{x}_d) \geq \mathbb{E}_{q(\boldsymbol{z}_d, \boldsymbol{\theta}_d)}[\log p(\boldsymbol{x}_d|\boldsymbol{z}_d, \boldsymbol{\Phi})p(\boldsymbol{z}_d|\boldsymbol{\theta}_d)p(\boldsymbol{\theta}_d)] - \mathbb{E}_{q(\boldsymbol{z}_d, \boldsymbol{\theta}_d)}[\log q(\boldsymbol{z}_d, \boldsymbol{\theta}_d)]$, where $\boldsymbol{\theta}_d$ is the topic probability distribution of the document d. $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K\}$ is the set of the per-topic word probability distributions, where K is the number of topics, and is MAP-estimated in this paper. Even when the posterior $q(\boldsymbol{z}_d, \boldsymbol{\theta}_d)$ is assumed to factorize as $q(\boldsymbol{z}_d)q(\boldsymbol{\theta}_d)$, closed

F. Li et al. (Eds.): APWeb 2016, Part II, LNCS 9932, pp. 424–428, 2016.
DOI: 10.1007/978-3-319-45817-5_39

form updates cannot be obtained for several parameters in CTM. Therefore, a gradient-based optimization is required. Further, since $\boldsymbol{\theta}_d$ is drawn from the logistic normal prior in CTM, the covariance matrix makes the inference complicated. The implementation by [3] explicitly inverts the covariance matrix. This paper proposes a simpler inference as an instance of SGVB [7,8]. The proposed method does not invert the covariance matrix explicitly thanks to SGVB.

2 Our Proposal: SGVB for CTM

In CTM, the per-document topic probabilities $\boldsymbol{\theta}_d$ are drawn from the logistic normal distribution, which is parameterized by the mean parameter $\boldsymbol{m}$ and the covariance matrix $\boldsymbol{\Sigma}$. We assume that the variational posterior $q(\boldsymbol{\theta}_d; \boldsymbol{\mu}_d, \boldsymbol{\sigma}_d)$ is the diagonal logistic normal, where the variational mean and standard deviation parameters for each pair (d, k) are referred to by μ_{dk} and σ_{dk}, respectively. The main contribution of this paper is that the inference is made simple by applying SGVB. A sample from $q(\boldsymbol{\theta}_d)$ is computed as $\theta_{dk} \propto \exp(\epsilon_{dk}\sigma_{dk} + \mu_{dk})$ with the reparameterization technique [7], where ϵ_{dk} is a noise distribution sample. The ELBO, i.e., the variational lower bound of the log evidence, is obtained as $\mathbb{E}_{q(\boldsymbol{z}_d;\boldsymbol{\gamma}_d)}\big[\log p(\boldsymbol{x}_d|\boldsymbol{z}_d;\boldsymbol{\Phi})\big] + \mathbb{E}_{q(\boldsymbol{z}_d;\boldsymbol{\gamma}_d)q(\boldsymbol{\theta}_d)}\big[\log p(\boldsymbol{z}_d|\boldsymbol{\theta}_d)\big] + \mathbb{E}_{q(\boldsymbol{\theta}_d)}\big[\log p(\boldsymbol{\theta}_d|\boldsymbol{m},\boldsymbol{\Sigma})\big] - \mathbb{E}_{q(\boldsymbol{z}_d;\boldsymbol{\gamma}_d)}\big[\log q(\boldsymbol{z}_d;\boldsymbol{\gamma}_d)\big] - \mathbb{E}_{q(\boldsymbol{\theta}_d)}\big[\log q(\boldsymbol{\theta}_d)\big]$ for the document d, where $q(\boldsymbol{z}_d;\boldsymbol{\gamma}_d)$ is assumed to be the discrete distribution. As no significant improvement was achieved by increasing the number of samples, the number of noise distribution samples was set to one in our experiment.

To estimate parameters, we take the partial derivatives of the ELBO L. The partial derivatives with respect to μ_{dk} and τ_{dk}, defined by $\tau_{dk} \equiv \log(\sigma^2)$, are $\frac{\partial L}{\partial \mu_{dk}} = \sum_{n=1}^{N_d} \gamma_{dnk} - N_d \exp(\epsilon_{dk}\sigma_{dk} + \mu_{dk})/\zeta_d - 2\sum_{k'=1}^{K-1} \mathit{\Lambda}_{kk'}(\epsilon_{dk}\sigma_{dk} + \mu_{dk} - m_{k'})$ and $\frac{\partial L}{\partial \tau_{dk}} = \frac{1}{2} + \frac{1}{2}\epsilon_{dk}\exp(\frac{\tau_{dk}}{2})\big[\sum_{n=1}^{N_d}\gamma_{dnk} - N_d \exp\{\epsilon_{dk}\exp(\frac{\tau_{dk}}{2}) + \mu_{dk}\}/\zeta_d - 2\sum_{k'=1}^{K-1}\mathit{\Lambda}_{kk'}\{\epsilon_{dk'}\exp(\frac{\tau_{dk'}}{2}) + \mu_{dk'} - m_{k'}\}\big]$, where $\zeta_d = 1 + \sum_{k=1}^{K-1}\exp(\epsilon_{dk}\sigma_{dk} + \mu_{dk})$. We skip the derivation due to the space limitation. Adam [6] is used for updating μ_{dk} and τ_{dk} repeatedly. Since the precision matrix $\boldsymbol{\Lambda}$, i.e., the inverse of the covariance matrix, appears only in the multiplication with a vector, the Cholesky decomposition can make the explicit inversion unnecessary. However, the analytic shrinkage [2] is used to make the computation stable. The $\boldsymbol{\phi}_k$s are MAP-estimated. The other parameters are updated by $\gamma_{dnk} \propto \theta_{dk}\phi_{kx_{dn}}$, $\boldsymbol{m} = \sum_{d=1}^{D}(\boldsymbol{\mu}_d + \boldsymbol{\epsilon}_d \circ \boldsymbol{\sigma}_d)/D$, and $\boldsymbol{\Sigma} = \sum_{d=1}^{D}(\boldsymbol{\epsilon}_d \circ \boldsymbol{\sigma}_d + \boldsymbol{\mu}_d - \boldsymbol{m})(\boldsymbol{\epsilon}_d \circ \boldsymbol{\sigma}_d + \boldsymbol{\mu}_d - \boldsymbol{m})^\top/2D$, where $\circ$ is the element-wise product.

3 Evaluation Experiment

The evaluation experiment was conducted over the four English document sets in Table 1. NYT is the first half of the New York Times news articles in "Bag of Words Data Set" of the UCI Machine Learning Repository.[1] MOVIE is the

[1] https://archive.ics.uci.edu/ml/datasets.html.

Table 1. Specifications of the four document sets used in the experiment

	# documents	# vocabulary words	# word tokens	Average doc. length
NYT	149,890	46,650	50,528,379	337.1
MOVIE	27,859	62,408	12,788,477	459.0
NSF	128,818	21,471	14,681,181	114.0
MEDLINE	125,490	42,830	17,610,749	140.3

set of movie reviews known as "Movie Review Data."[2] NSF is "NSF Research Award Abstracts 1990-2003 Data Set" of the UCI Machine Learning Repository. MEDLINE is a subset of the MEDLINE®/PUBMED®.[3] For all document sets, we applied the Porter stemming and removed high- and low-frequency words.

We compared the proposed SGVB for CTM to the collapsed Gibbs sampling (CGS) for LDA [5] and also to the collapsed variational Bayes with a zero-order Taylor expansion approximation (CVB0) for LDA [1]. The original VB for CTM has already been compared to LDA in [3]. Therefore, we did not repeat the comparison. However, CVB0 could not be considered in [3]. Therefore, we picked CVB0 up. The evaluation measure was the predictive perplexity. We first ran each compared method on the randomly selected 90 % training documents. We then ran each method on a randomly selected one third of the word tokens of each test document to obtain an estimation of the topic probabilities, where the per-topic word probabilities were never updated. By using the rest two thirds, the perplexity was computed by $\exp\{-\frac{1}{N_{\text{test}}}\sum_{d\in\mathcal{D}_{\text{test}}}\sum_{i\in\mathcal{I}_d}\log(\sum_{k=1}^{K}\theta_{dk}\phi_{kx_{di}})\}$, where $\mathcal{D}_{\text{test}}$ denotes the test document set, $\mathcal{I}_d$ the set of the indices of the test word tokens in the dth document, and N_{test} the total number of the test tokens. For each data set, the methods were compared for $K = 50, 100$, and 150.

Figure 1 depicts the mean and standard deviation of the perplexities obtained from ten different training/test random splits. The horizontal axis gives K, and the vertical axis the perplexity. CVB0 was better than the other methods. A similar result has been given in [1], though only the inferences for LDA is considered. With respect to the comparison of SGVB for CTM to CGS for LDA, the former was better than the latter for the following five cases: $K = 50$ and 100 for NYT (resp. $p = 0.00072$ and 0.001), $K = 50$ and 100 for MOVIE (resp. $p = 7.6 \times 10^{-7}$ and 0.00013), and $K = 50$ for MEDLINE ($p = 9.1 \times 10^{-6}$). The latter was better for the following five cases: $K = 150$ for MOVIE, $K = 50, 100$, and 150 for NSF, and $K = 150$ for MEDLINE. The former was comparable with the latter for the other two cases. The p values were obtained by the paired two-tailed t-test. In sum, the proposed SGVB for CTM was as good as CGS for LDA. However, the proposed method was far worse only for NSF, where it is suspected that the gradient-based optimization did not work well. Figure 2 gives topic correlations obtained from NYT for $K = 100$.

[2] http://www.cs.cornell.edu/people/pabo/movie-review-data/.

[3] We used the XML files from medline14n0770.xml to medline14n0774.xml.

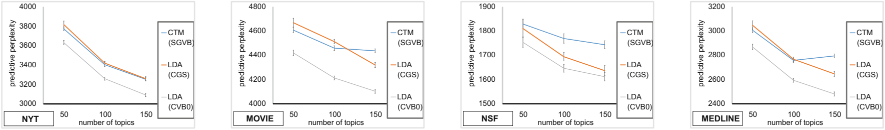

Fig. 1. Evaluation results in terms of predictive perplexity.

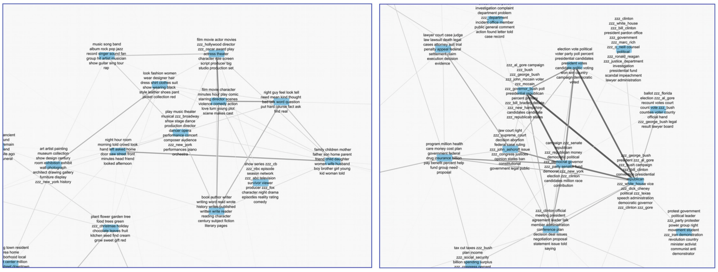
music song band album rock pop jazz record singer sound fan group hit artist musician show guitar sing tour rap
film movie actor movies zzz_hollywood director zzz_oscar award play actress theater character role screen script producer big studio production set
look fashion women wear designer hair dress shirt clothes suit show wearing black style leather shoes pant jacket collection red
film movie character minutes hour play comic starring director scenes violence comedy action love turn young plot scene makes cast
right guy feel look tell need mean kind thought bad talk word question put hard course fact ask find real
play music theater musical zzz_broadway show stage dance production director dancer opera performance concert composer audience zzz_new_york performances piano orchestra
night hour room morning told crowd took hand left asked home door saw street front minutes head friend looked afternoon
art artist painting museum collection show design century room exhibition exhibit wall photograph architect drawing gallery furniture display zzz_new_york history
family children mother father son home parent friend child daughter women wife husband boy brother girl young kid woman told
show series zzz_cb zzz_nbc episode season network zzz_abc television survivor viewer producer zzz_fox character night drama episodes reality rating comedy
book author writer writing word read wrote history writes published written write reader reading character century subject fiction literary pages
plant flower garden tree food trees green zzz_christmas holiday chocolate leaves fruit kitchen seed find cream grow sweet gift red
investigation complaint department problem zzz_department incident office member public general comment action found letter told case record
lawyer court case judge law lawsuit death legal cases attorney suit trial penalty appeal federal settlement claim execution decision evidence
zzz_clinton zzz_white_house zzz_bill_clinton president pardon office zzz_government zzz_marc_rich zzz_o_neill counsel political zzz_ronald_reagan zzz_justice_department investigation presidential fund scandal impeachment lawyer administration
election vote political voter party poll percent presidential candidates president votes candidate politic voting won win country campaign democratic voted
zzz_al_gore campaign zzz_bush zzz_george_bush zzz_john_mccain voter zzz_mccain zzz_governor_bush poll presidential republican percent primary zzz_bill_bradley debate zzz_new_hampshire candidates candidate zzz_republican states
ballot zzz_florida election zzz_al_gore recount votes court count vote zzz_bush counties voter county official hand zzz_george_bush legal result lawyer board
law court right zzz_supreme_court decision abortion federal case ruling zzz_john_ashcroft issue zzz_congress justices opinion states ban constitutional government legal public
program million health care money cost plan government federal drug insurance billion pay benefit percent help fund group need proposal
campaign zzz_senate republican zzz_republican money democratic political zzz_democrat governor zzz_party senator fund democrat zzz_new_york election zzz_clinton candidates million race contribution
zzz_george_bush president zzz_al_gore zzz_bush campaign zzz_bill_clinton convention presidential republican zzz_white_house vice zzz_dick_cheney political zzz_texas speech administration democratic governor zzz_clinton zzz_gore
zzz_clinton official meeting president agreement leader talk member administration conference plan decision deal issues negotiation proposal statement issue told saying
protest government political leader zzz_party protester power group right movement student zzz_iran demonstration revolution country minister activist communist anti demonstrator
tax cut taxes zzz_bush plan income zzz_social_security billion spending surplus zzz_congress percent

Fig. 2. Topic correlations obtained by our method from the NYT data set. The left and right panels contain the topics seemingly relating to art and politics, respectively.

The graph is drawn by Cytoscape[4]. The edge thickness represents the magnitude of the corresponding entry of the covariance matrix. The left panel contains the topics seemingly relating to art, and the right those seemingly relating to politics.

4 Conclusion

This paper proposed a new inference for CTM. We apply SGVB to CTM and obtain a set of simple updating formulas. The experiment showed that the proposed method was comparable with CGS for LDA in terms of perplexity, though CVB0 for LDA was the best for all settings. While the proposed method was as good as CGS for LDA, it did not work well for some cases. A further elaboration seems required for a more effective gradient-based optimization.

References

1. Asuncion, A., Welling, M., Smyth, P., Teh, Y.W.: On smoothing and inference for topic models. In: UAI, pp. 27–34 (2009)
2. Bartz, D., Müller, K.R.: Generalizing analytic shrinkage for arbitrary covariance structures. In: NIPS 26, pp. 1869–1877 (2013)
3. Blei, D.M., Lafferty, J.D.: Correlated topic models. In: NIPS, pp. 147–154 (2005)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. JMLR **3**, 993–1022 (2003)

[4] http://www.cytoscape.org/.

5. Griffiths, T.L., Steyvers, M.: Finding scientific topics. PNAS **101**(Suppl 1), 5228–5235 (2004)
6. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (2015)
7. Kingma, D.P., Welling, M.: Stochastic gradient VB and the variational auto-encoder. In: ICLR (2014)
8. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: ICML, pp. 1278–1286 (2014)