A Revised Inference for Correlated Topic Model

Tomonari Masada[1] and Atsuhiro Takasu[2]

[1] Nagasaki University, 1-14 Bunkyo-machi, Nagasaki-shi, Nagasaki, 852-8521 Japan,
masada@nagasaki-u.ac.jp
[2] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430 Japan,
takasu@nii.ac.jp

Abstract. In this paper, we provide a revised inference for correlated topic model (CTM) [3]. CTM is proposed by Blei et al. for modeling correlations among latent topics more expressively than latent Dirichlet allocation (LDA) [2] and has been attracting attention of researchers. However, we have found that the variational inference of the original paper is unstable due to almost-singularity of the covariance matrix when the number of topics is large. This means that we may be reluctant to use CTM for analyzing a large document set, which may cover a rich diversity of topics. Therefore, we revise the inference and improve its quality. First, we modify the formula for updating the covariance matrix in a manner that enables us to recover the original inference by adjusting a parameter. Second, we regularize posterior parameters for reducing a side effect caused by the formula modification. While our method is based on a heuristic intuition, an experiment conducted on large document sets showed that it worked effectively in terms of perplexity.

Keywords: topic models, variational inference, covariance matrix

1 Introduction

Topic modeling is one of the dominant trends of recent data mining research. This approach uses latent variables for modeling topical diversity in document data and represents the data in a lower-dimensional "topic" space. This line of thinking reminds us that multilayer undirected graphical models also extract a lower-dimensional data representation. In fact, neural network researchers also propose a topic model, e.g. by using restricted Boltzmann machines [12] or by a hybridization [13], where we can find an affinity between both approaches.

However, many topic models, including latent Dirichlet allocation (LDA) proposed by the inaugural paper of Blei et al. [2], do not consider *correlations* among latent topics explicitly. It is natural to assume that latent topics interwoven into a semantic content of each document are correlated to some extent. For example, articles on worldwide energy consumption may often mention geopolitical conflicts among countries. Therefore, Blei et al. [3] have proposed correlated topic model (CTM) to offer an established way to model correlations among latent topics. While we know that recent sophisticated approaches can also model topic correlations [11, 4], we here concentrate on CTM.

In CTM, per-document topic distributions are obtained from a corpus-wide logistic normal distribution, where a covariance matrix $\boldsymbol{\Sigma}$ models correlations among latent topics. More precisely, we draw a K-dimensional vector $\boldsymbol{m}_d$ for document d from a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where K is the number of topics, and obtain a topic distribution $\boldsymbol{\theta}_d$ as $\theta_{dk} = \exp(m_{dk}) / \sum_{k'} \exp(m_{dk'})$. This construction makes CTM more expressive than LDA.

In the variational inference for CTM proposed in the original paper [3], the covariance matrix $\boldsymbol{\Sigma}$ is estimated by maximizing the variational lower bound of the log evidence. Further, $\boldsymbol{\Sigma}$ needs to be inverted in every iteration. Therefore, $\boldsymbol{\Sigma}$ should be kept less singular all through the iterations of the inference. However, our preliminary experiment conducted on large document sets has shown that the inference often gets unstable due to the fact that $\boldsymbol{\Sigma}$ is almost singular. When we apply CTM to a large data set, we would like to set the number of topics to a large value, say 300. Under that situation, the original inference gives a result disastrous in terms of perplexity [2]. Therefore, we have revised the inference. Through trying several heuristic methods for making the covariance matrix less singular, we have found an effective one, which is given in this paper.

However, after revising the inference, we have found another problem. It is known that variational inference for topic models is likely to give worse perplexity than inference by sampling [1], though it is not known whether there are any efficient sampling methods for CTM like collapsed Gibbs sampling (CGS) for LDA [6]. Therefore, for achieving as good perplexity as possible, we can take the following strategy: run CGS for LDA first and then run the variational inference for CTM. Fortunately, we can initialize the parameters of CTM based on a result of CGS for LDA. Therefore, we can start the variational inference for CTM with the parameter values giving a good perplexity. However, a preliminary experiment has shown that our revised inference is likely to make the initial perplexity, which is achieved by CGS for LDA, worse as the inference proceeds. Therefore, we regularize some of the variational posterior parameters so that they are kept close to their initial values inherited from CGS for LDA.

In sum, our method consists of the following two features: 1) Keep the covariance matrix less singular; and 2) Keep the values of some posterior parameters close to their values initialized based on a result of CGS for LDA. With respect to the former, we modify the update formula of the covariance matrix by taking an approach similar to shrinkage estimation. With respect to the latter, we add a regularization term so that the parameter values do not deviate significantly from their initial values. The rest of the paper is organized as follows. Section 2 describes the details of the original variational inference for CTM. Section 3 presents our proposal. Section 4 contains the results of an experiment conducted over large document sets. Section 5 concludes the paper with discussions.

2 Correlated Topic Models

With correlated topic model (CTM) [3], we can explicitly model correlations among latent topics. The variety of correlations that can be modeled in CTM is

richer than that in LDA, because Dirichlet prior distribution used in LDA is not that powerful in modeling correlations among drawn multinomial probabilities. To make the paper self-contained, we describe CTM in detail. In the following, we denote the number of topics, different words, and documents by K, W, and D, respectively, and identify each entity with its index number.

As in latent Dirichlet allocation (LDA) [2], we represent each document d as a mixture of K latent topics by using a multinomial distribution Multi($\boldsymbol{\theta}_d$) defined over topics, also in CTM. The multinomial parameters $\boldsymbol{\theta}_d = (\theta_{d1}, \ldots, \theta_{dK})$ satisfy $\sum_k \theta_{dk} = 1$, where θ_{dk} is a probability that topic k is expressed by a word token in document d. Topic k is in turn represented by a multinomial distribution Multi($\boldsymbol{\phi}_k$) defined over words. The parameters $\boldsymbol{\phi}_k = (\phi_{k1}, \ldots, \phi_{kW})$ satisfy $\sum_w \phi_{kw} = 1$, where ϕ_{kw} is a probability that a token of word w expresses topic k. Both in LDA and in CTM, we assume that $\boldsymbol{\phi}_k$s are drawn from a corpus-wide Dirichlet prior distribution Dir($\boldsymbol{\beta}$), where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_W)$ is a set of its hyperparameters. However, with respect to per-document topic multinomial distributions Multi($\boldsymbol{\theta}_d$), LDA and CTM show a difference.

In LDA, per-document topic distributions are drawn from a corpus-wide Dirichlet distribution Dir($\boldsymbol{\alpha}$). However, Dirichlet distribution can only model a restricted variety of correlations among topics. Therefore, CTM adopts logistic-normal distribution as the prior for $\boldsymbol{\theta}_d$s. We below describe CTM generatively.

1. For each topic $k \in \{1, \ldots, K\}$, draw parameters $\boldsymbol{\phi}_k = (\phi_{k1}, \ldots, \phi_{kW})$ of a multinomial distribution Multi($\boldsymbol{\phi}_k$) defined over words $\{1, \ldots, W\}$ from the corpus-wide Dirichlet prior distribution Dir($\boldsymbol{\beta}$).
2. For each document $d \in \{1, \ldots, D\}$,
   (a) Draw a K-dimensional vector $\boldsymbol{m}_d = (m_{d1}, \ldots, m_{dK})$ from the corpus-wide K-dimensional Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
   (b) Obtain a topic distribution $\boldsymbol{\theta}_d = (\theta_{d1}, \ldots, \theta_{dK})$ as $\theta_{dk} = \frac{\exp(m_{dk})}{\sum_{k'} \exp(m_{dk'})}$.
   (c) Let n_d be the length (i.e., the number of word tokens) of document d. For the ith word token in document d, where $i \in \{1, \ldots, n_d\}$,
      i. Draw a topic from the topic multinomial distribution Multi($\boldsymbol{\theta}_d$). Let the drawn topic be the value of a latent variable z_{di}, which gives the topic to which the ith word token in document d is assigned.
      ii. Draw a word from the word multinomial distribution Multi($\boldsymbol{\phi}_{z_{di}}$). Let the drawn word be the value of an observed variable x_{di}, which gives the word appearing as the ith word token of document d.

Based on this description, we obtain the full joint distribution of CTM as follows:

$$
\begin{aligned}
&p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\phi}, \boldsymbol{m} | \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = p(\boldsymbol{\phi}|\boldsymbol{\beta}) p(\boldsymbol{m}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{z}|\boldsymbol{m}) p(\boldsymbol{x}|\boldsymbol{\phi}, \boldsymbol{z}) \\
&= \prod_k \frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \phi_{kw}^{\beta_w - 1} \cdot \prod_d \frac{1}{(2\pi)^{K/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\Big\{ -\frac{1}{2} (\boldsymbol{m}_d - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{m}_d - \boldsymbol{\mu}) \Big\} \\
&\quad \cdot \prod_d \prod_i \frac{\exp(m_{d z_{di}})}{\sum_k \exp(m_{dk})} \phi_{z_{di} x_{di}} \ . \qquad (1)
\end{aligned}
$$

The variational inference proposed in the original paper [3] approximates the posterior $p(\boldsymbol{z}, \boldsymbol{\phi}, \boldsymbol{m}|\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ by a factorized variational posterior $q(\boldsymbol{z})q(\boldsymbol{\phi})q(\boldsymbol{m})$. Consequently, a lower bound of the log evidence $\ln p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is obtained by applying Jensen's inequality as follows:

$$\begin{aligned}\ln p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \geq & \int \sum_{\boldsymbol{z}} q(\boldsymbol{z})q(\boldsymbol{m}) \ln p(\boldsymbol{z}|\boldsymbol{m}) d\boldsymbol{m} + \int q(\boldsymbol{\phi}) \ln p(\boldsymbol{\phi}|\boldsymbol{\beta}) d\boldsymbol{\phi} \\ & + \int \sum_{\boldsymbol{z}} q(\boldsymbol{z})q(\boldsymbol{\phi}) \ln p(\boldsymbol{x}|\boldsymbol{\phi}, \boldsymbol{z}) d\boldsymbol{\phi} + \int q(\boldsymbol{m}) \ln p(\boldsymbol{m}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\boldsymbol{m} \\ & - \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \ln q(\boldsymbol{z}) - \int q(\boldsymbol{\phi}) \ln q(\boldsymbol{\phi}) d\boldsymbol{\phi} - \int q(\boldsymbol{m}) \ln q(\boldsymbol{m}) d\boldsymbol{m} \ . \end{aligned} \tag{2}$$

We denote the right hand side of Eq. (2) by $\mathcal{L}$, which needs to be maximized. With respect to each variational posterior distribution, we assume the followings.

- $q(\boldsymbol{z})$ is factorized as $\prod_d \prod_i \gamma_{diz_{di}}$, where γ_{dik} is an approximated probability that the ith word token in document d expresses topic k.
- $q(\boldsymbol{\phi})$ is factorized as $\prod_k q(\boldsymbol{\phi}_k|\boldsymbol{\zeta}_k)$, where $q(\boldsymbol{\phi}_k|\boldsymbol{\zeta}_k)$ is the density of an approximated Dirichlet posterior whose parameters are $\boldsymbol{\zeta}_k = (\zeta_{k1}, \ldots, \zeta_{kW})$.
- $q(\boldsymbol{m})$ is factorized as $\prod_d \prod_k q(m_{dk}|r_{dk}, s_{dk})$, where $q(m_{dk}|r_{dk}, s_{dk})$ is a density of a univariate Gaussian distribution whose mean and standard deviation parameters are r_{dk} and s_{dk}, respectively.

The variational parameters γ_{dik}s, ζ_{kw}s, and ν_ds can be updated by a closed formula: $\gamma_{dik} \propto \exp(r_{dk}) \cdot \frac{\exp \Psi(\zeta_{kx_{di}})}{\exp \Psi(\sum_w \zeta_{kw})}$; $\zeta_{kw} = \beta_w + \sum_d \sum_i \sum_k \gamma_{dik}$; and $\nu_d = \sum_k \exp(r_{dk} + s_{dk}^2/2)$, where ν_ds are the parameters introduced to make the inference tractable. Details of the derivation are referred to the original paper [3].

However, the parameters r_{dk}s and s_{dk}s cannot be updated by any closed formulas. In this paper, we use L-BFGS [10, 7] for maximizing the relevant terms in $\mathcal{L}$ and update these parameters. The target functions are given below.

$$\begin{aligned}\mathcal{L}(r_{dk}) = & \, n_{dk} r_{dk} - \frac{n_d}{\nu_d} \exp(r_{dk} + s_{dk}^2/2) \\ & + \frac{1}{2} r_{dk}^2 (\boldsymbol{\Sigma}^{-1})_{kk} - r_{dk} \sum_{k'} (r_{dk'} - \mu_{k'}) (\boldsymbol{\Sigma}^{-1})_{kk'} \end{aligned} \tag{3}$$

$$\mathcal{L}(s_{dk}) = -\frac{n_d}{\nu_d} \exp(r_{dk} + s_{dk}^2/2) - \frac{1}{2} s_{dk}^2 (\boldsymbol{\Sigma}^{-1})_{kk} + \ln s_{dk} \tag{4}$$

Since the parameters $r_{d1}, \ldots, r_{dK}$ and $s_{d1}, \ldots, s_{dK}$ for a fixed d are dependent on each other, we update them not separately but in concert by using L-BFGS.

The mean parameter $\boldsymbol{\mu}$ of the K-dimensional Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, which models the correlation among latent topics, can be updated as $\boldsymbol{\mu} = \frac{1}{D} \sum_d \boldsymbol{r}_d$. The covariance matrix $\boldsymbol{\Sigma}$ can be updated as

$$\boldsymbol{\Sigma} = \frac{1}{D} \sum_d \boldsymbol{S}_d + \frac{1}{D} \sum_d (\boldsymbol{r}_d - \boldsymbol{\mu})(\boldsymbol{r}_d - \boldsymbol{\mu})^T \ , \tag{5}$$

where $\boldsymbol{S}_d$ is a $K \times K$ diagonal matrix whose kth diagonal entry is s_{dk}^2. However, Eq. (5) is likely to give an almost singular matrix when K is large. This is a serious problem, because we need the inverse of $\boldsymbol{\Sigma}$ in Eqs. (3) and (4). When we apply CTM to a large document set, we would like to set the number of topics K to a large number, say 300. This type of situation is likely to make $\boldsymbol{\Sigma}$ almost singular and thus to make the entire inference unstable. Therefore, we propose a revised inference for CTM.

3 A revised inference for CTM

Our proposal aims to achieve a stable inference for CTM by making the covariance matrix $\boldsymbol{\Sigma}$ far from singular. The proposal has the following two features:

- We modify the update formula Eq. (5) in a manner that we can recover the original inference by adjusting a parameter called *interpolation parameter*.
- We initialize the parameters of CTM by using a result of collapsed Gibbs sampling (CGS) for LDA and regularize some parameters lest they deviate substantially from their initial values. The strength of regularization can be adjusted by a parameter called *regularization parameter*.

3.1 Covariance matrix update

First, we discuss a revised update of $\boldsymbol{\Sigma}$. We focus on the $K \times K$ matrix appearing as the second term of the right hand side of Eq. (5), i.e., $\hat{\boldsymbol{T}} = \frac{1}{D}\sum_d (\boldsymbol{r}_d - \boldsymbol{\mu})(\boldsymbol{r}_d - \boldsymbol{\mu})^T$. We can view $\hat{\boldsymbol{T}}$ as the maximum likelihood estimator of the covariance matrix $\boldsymbol{T}$ of a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{T})$, from which $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_D$ are drawn. Chen et al. [5] give the following matrix as a "naive but most well-conditioned estimate" for $\boldsymbol{T}$: $\hat{\boldsymbol{F}} = \frac{\mathrm{Tr}(\hat{\boldsymbol{T}})}{K}\boldsymbol{I}$, where $\mathrm{Tr}(\cdot)$ is the trace of a matrix and $\boldsymbol{I}$ is the identity matrix. $\hat{\boldsymbol{F}}$ is a diagonal matrix whose diagonal entries are all equal to the average of the diagonal elements of $\hat{\boldsymbol{T}}$. Our approach uses $\hat{\boldsymbol{F}}$ in place of the first term of the right hand side of Eq. (5) and update $\boldsymbol{\Sigma}$ as $\boldsymbol{\Sigma} = \frac{\mathrm{Tr}(\hat{\boldsymbol{T}})}{K}\boldsymbol{I} + \hat{\boldsymbol{T}}$. Further, we obtain a linear interpolation of this equation and Eq. (5) by introducing an *interpolation parameter* π as follows:

$$\boldsymbol{\Sigma} = \left\{(1-\pi)\cdot\frac{1}{D}\sum_d \boldsymbol{S}_d + \pi\cdot\frac{\mathrm{Tr}(\hat{\boldsymbol{T}})}{K}\boldsymbol{I}\right\} + \frac{1}{D}\sum_d (\boldsymbol{r}_d - \boldsymbol{\mu})(\boldsymbol{r}_d - \boldsymbol{\mu})^T \ . \qquad (6)$$

When $\pi = 0$, we can recover the original inference. It can be said that we use $\hat{\boldsymbol{F}}$ to conduct a shrinkage operation on the average of the covariance matrices $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_D$ of the variational Gaussian posteriors $q(\boldsymbol{m}_d|\boldsymbol{r}_d, \boldsymbol{S}_d)$, $d = 1, \ldots, D$.

3.2 Variational mean regularization

Second, we discuss a regularization of parameters. We update the variational means, i.e., r_{dk}s, in a regularized manner, because r_{dk}s are likely to deviate from their initial values when we use Eq. (6), in place of Eq. (5), for updating $\boldsymbol{\Sigma}$.

Table 1. Document set specifications.

	# docs	# words	# training tokens (# test tokens)
CORA[3]	36,183	8,542	2,127,005 (235,566)
MOVREV[4]	27,859	18,616	8,145,228 (905,427)
TDT4[5]	96,246	15,153	15,070,250 (1,674,304)
NSF[6]	128,181	19,066	12,284,568 (1,366,399)
MEDLINE[7]	2,495,210	134,615	225,844,065 (25,097,215)

The probability of topic k in document d is calculated as $\frac{\exp(m_{dk})}{\sum_{k'}\exp(m_{dk'})}$, and r_{dk} is of the same dimension with m_{dk}. Therefore, based on a result of CGS for LDA, we can initialize r_{dk} as follows: $r_{dk} = \log(n_{dk} + \alpha_k)$, where n_{dk} is the number of the word tokens in document d that are assigned to topic k. α_k is a Dirichlet hyperparameter corresponding to topic k and is updated by Minka's method [8, 1] in our experiment. To keep r_{dk}s close to their initial values, we add a regularization term to the right hand side of Eq. (3) as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{reg}}(r_{dk}) = n_{dk}r_{dk} - \frac{n_d}{\nu_d}\exp(r_{dk} + s_{dk}^2/2) \\ + \frac{1}{2}r_{dk}^2(\boldsymbol{\Sigma}^{-1})_{kk} - r_{dk}\sum_{k'}(r_{dk'} - \mu_{k'})(\boldsymbol{\Sigma}^{-1})_{kk'} \\ - \pi\rho\big\{r_{dk} - \log(n_{dk} + \alpha_k)\big\}^2 \ . \end{aligned} \tag{7}$$

We call ρ *regularization parameter*, which determines the strength of the regularization and takes a non-negative value. We maximize $\mathcal{L}_{\mathrm{reg}}(r_{dk})$ in Eq. (7) in place of $\mathcal{L}(r_{dk})$ in Eq. (3). Note that ρ is multiplied by the interpolation parameter π. Therefore, even when $\rho > 0$, the inference is reduced to the original one as long as $\pi = 0$. When ρ takes a large positive value, r_{dk}s are kept close to their initial values. This means that we keep staying close to the result of CGS for LDA. When we applied this regularization to our revised inference in our experiment, we always set ρ to 1.0, because other settings gave no interesting differences. Therefore, we have only one parameter π to be adjusted by hand.

Consequently, $\pi = 0.0$ means that we use the original inference for CTM and, at the same time, do not regularize r_{dk}s. Further, $\pi = 1.0$ means that we use the revised inference for CTM in its full capacity and, at the same time, regularize r_{dk}s with strength 1.0. However, we needed to directly apply the regularization to the original inference in our experiment for comparison. Therefore, in this case, we set ρ to 1.0 after eliminating π from Eq. (7).

[3] `http://people.cs.umass.edu/~mccallum/data.html`

[4] `http://www.cs.cornell.edu/people/pabo/movie-review-data/polarity_html.zip`

[5] `http://projects.ldc.upenn.edu/TDT4/`

[6] `http://archive.ics.uci.edu/ml/datasets/NSF+Research+Award+Abstracts+1990-2003`

[7] This is the set of the abstracts extracted from the XML files whose names range from `medline12n0600.xml` to `medline12n0699.xml`.

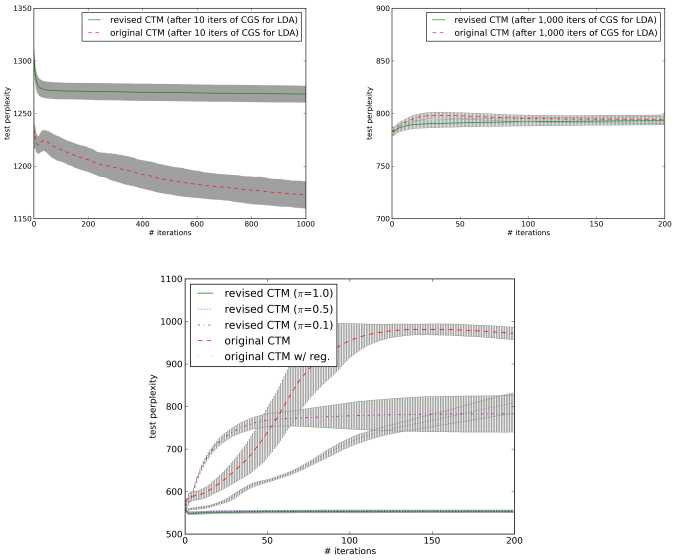

Fig. 1. Comparing the revised inference with the original one on CORA document set.

4 Comparison experiment

We compared our revised inference with the original one on the document sets in Table 1. For each document set, we conducted a series of appropriate preprocessings, e.g. changing text case to lower case, stemming, removing high and low frequency words, etc. The evaluation measure is *test perplexity* [2], which represents predictive power of topic models and is defined as $\exp(-\frac{\sum_d \sum_i \log \sum_k \lambda_{dk} \zeta_{k x_{di}}}{N_{test}})$, where N_{test} is the number of the word tokens used for calculating perplexity, called test word tokens. λ_{dk} is defined as $\sum_{i=1}^{n_d} \gamma_{dik}$. We randomly select 90% word tokens from every document for running inference and use the rest for calculating test perplexity. Note that smaller perplexity is better.

Fig. 1 gives test perplexities of the revised and the original inferences conducted on CORA document set under various settings. The horizontal axis represents number of iterations, and the vertical axis represents perplexity. For each setting, we repeated inference 10 times and calculated a mean and a standard deviation of the corresponding 10 perplexities. The mean and the standard deviation, which is depicted by the error bar, are given at every iteration.

First, we clarify when the original inference gave a better perplexity. The top left panel of Fig. 1 presents the result obtained when the number of topics was 30, a fairly small number, and we ran only 10 iterations of CGS for LDA before initiating the revised or the original inference for CTM. The number of iterations of the inference for CTM was set to 1,000, because CGS for LDA iterates only 10 times and could not reduce perplexity enough. The interpolation parameter π was set to 1.0 for the revised inference and to 0.0 for the original inference.

Since 10 iterations of CGS for LDA led to a perplexity nearly equal to 1,320, both of the line graphs, each corresponding to the revised inference (green solid line) and the original one (red dashed line), start from the perplexity of around 1,320. Obviously, the original inference is significantly better. It can be said that the original inference works without any modification when we set the number of topics to be relatively small and start the variational inference for CTM after a small number of iterations of CGS for LDA.

Second, we increased the number of iterations of CGS for LDA. The top right panel of Fig. 1 gives the result when we ran 1,000 iterations of CGS for LDA and then ran 200 iterations of the revised or the original inference for CTM. π was set to 1.0 for the revised inference and to 0.0 for the original inference. 1,000 iterations of CGS reduced perplexity to around 780. In this manner, CGS for LDA reduced perplexity significantly before we initiated the inference for CTM. Consequently, both of the revised and the original inferences could not improve the perplexity achieved by CGS for LDA,[8] though they behave a little differently. However, it is clear that the original inference gave almost the same perplexity with the revised inference, and thus that the revised inference could show no advantage also in this case.

Third, we increased the number of topics to 300. The bottom panel of Fig. 1 gives the corresponding result, where we ran 1,000 iterations of CGS for LDA and ran 200 iterations of the revised or the original inference for CTM. This panel contains five line graphs, each corresponding to the following cases: 1) $\pi = 1.0$ (green solid line), 2) $\pi = 0.5$ (blue dotted line), 3) $\pi = 0.1$ (magenta dash-dot line), 4) the original inference (red dashed line), and 5) the original inference with regularization (black pixel marker). The regularization parameter ρ was set to 1.0 for the revised inferences. We also applied our regularization to the original inference by setting $\pi\rho = 1.0$ in Eq. (7) and, at the same time, by setting $\pi = 0.0$ in Eq. (6). As this panel shows, the perplexities for the cases $\pi = 1.0$ and $\pi = 0.5$ gave almost the same perplexity, which is a little better than that achieved by CGS for LDA. However, the case $\pi = 0.1$ and the original inference gave significantly worse perplexities than those two cases. Further, even when we apply the regularization to the original inference, the improvement was not remarkable.[9] That is, our regularization did not work for the original inference. At the same time, this result also shows that the regularization with $\pi\rho = 1.0$ is not so strong to forcibly keep r_{dk}s almost the same with their initial values.

In sum, the original inference for CTM works if we set the number of topics to be fairly small. However, note that the bottom panel in Fig. 1 contains the best perplexity among the three panels in this figure. Therefore, it is better to choose the revised inference when we would like to make topic models fully show their predictive power. In contrast, if we are under a situation where perplexity hardly matters for some reasons, the original inference may find relevance.

[8] Needless to say, if we do not conduct the inference for CTM, we just have a result of CGS for LDA and cannot consider correlations among topics.

[9] When we conduct the revised inference without regularization, we consistently observed worse perplexities, though we do not present the corresponding results here.

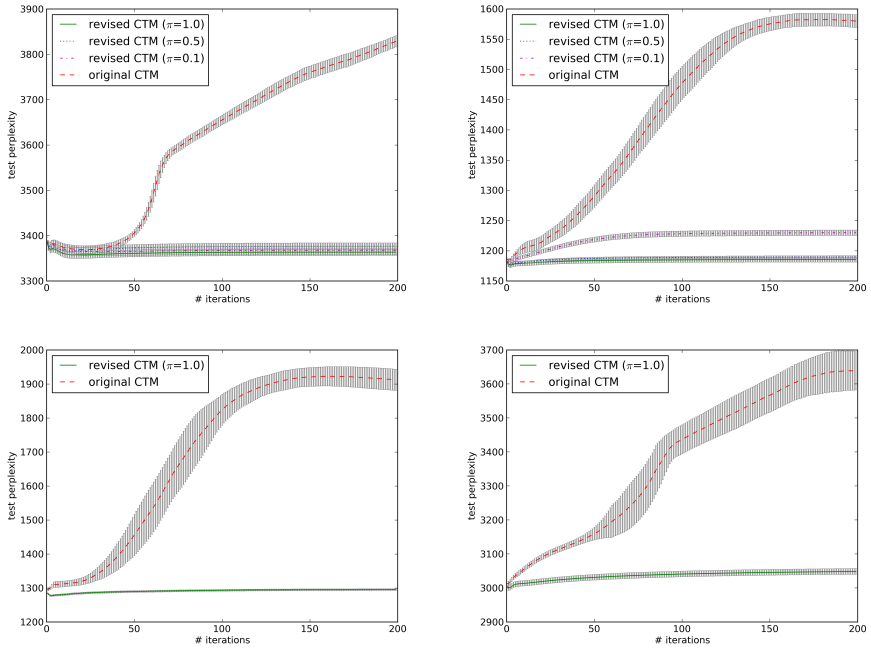

Fig. 2. Comparing the revised inference with the original one on MOVREV (top left), TDT4 (top roght), NSF (bottom left), and MEDLINE (bottom right) document sets.

We obtained similar results also for the other document sets in Table 1. Fig. 2 contains all results. As in case of the bottom panel of Fig. 1, we ran 1,000 iterations of CGS for LDA and then started 200 iterations of the revised or the original inference for CTM. For MOVREV and TDT4 document sets, we tested the following four settings: $\pi = 1.0, 0.5, 0.1$, and 0.0. The regularization is applied only to the revised inference with $\rho = 1.0$ in Eq. (7). As the top left panel shows, for MOVREV document set, the three settings $\pi = 1.0, 0.5$, and 0.1 gave almost the same perplexity, which is slightly better than the perplexity achieved by CGS for LDA. In contrast, the original inference led to a far larger perplexity. For TDT4, we obtained almost the same perplexity when $\pi = 1.0$ and 0.5, and the two other cases resulted in worse perplexities, as the top left panel depicts. The bottom left and the bottom right panels present the results for NSF and MEDLINE document sets, respectively. For these relatively large document sets, we only provide the results for the two extreme cases, i.e., the case $\pi = 1.0$ and the original inference. Obviously, the revised inference achieved a better perplexity by a large margin.

In the end, we add comments on implementation. With respect to matrix inversion, we used `dgetrf_()` and `dgetri_()` in CLAPACK. As a byproduct of `dgetrf_()`, we can calculate the determinant of the covariance matrix. When we ran the original inference, the absolute value of the determinant went to a value

close to zero in case the number of topics was relatively large. Consequently, $\mathcal{L}$, i.e., the lower bound of the log evidence, decreased by a large amount in earlier iterations, though $\mathcal{L}$ needed to be maximized. While we knew that there was an implementation by the authors of the original paper [3], it consumed main memory by a considerable factor when compared with our implementation and thus could not load any document sets in a reasonable time. Therefore, we evaluated the original inference by setting π to 0.0 in our own implementation.

5 Conclusion

In this paper, we provide a revised inference for CTM. We modify diagonal elements of the covariance matrix lest it be almost singular and regularize variational mean parameters lest they deviate from their initial values. However, as long as the model is described by using a $K \times K$ covariance matrix, it seems difficult to completely avoid the problem of matrix singularity, because K needs to be larger for a topic analysis over larger document sets. Therefore, recent proposals [11, 4] may find their relevance. However, CTM has an advantage in its simplicity and efficiency of inference. We tried to make such CTM to be utilized in a wider range of situations. An important future work is to compare our revised inference with the original one in terms of other evaluation measures [9].

References

1. Asuncion, A., Welling, M., Smyth, P., Teh, Y. W.: On smoothing and inference for topic models. in Proc. of UAI'09, pp. 27–34 (2009)
2. Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent Dirichlet allocation. JMLR, 3, pp. 993–1022 (2003)
3. Blei, D., Lafferty, J.: Correlated topic models. NIPS 18, pp. 147–154 (2006)
4. Chen, C., Ding, N., Buntine, W.: Dependent hierarchical normalized random measures for dynamic topic modeling. in Proc. of ICML'12 (2012)
5. Chen, Y., Wiesel, A., Eldar, Y.C., Hero III, A.O.: Shrinkage algorithms for MMSE covariance estimation. IEEE Trans. on Sign. Proc., 58, 10, pp. 5016–5029 (2010)
6. Griffiths, T. L., Steyvers, M.: Finding scientific topics. PNAS, 101, Suppl 1, pp. 5288–5235 (2004)
7. Liu D.C., Nocedal, J.: On the limited memory method for large scale optimization. Math. Programming B, 45, 3, pp. 503–528 (1989)
8. Minka, T.P.: Estimating a Dirichlet distribution. `http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/` (2000)
9. Newman, D., Karimi, S., Cavedon, L.: External Evaluation of Topic Models. in Proc. of ADCS'09, pp.11–18 (2009)
10. Nocedal, J.: Updating quasi-Newton matrices with limited storage. Math. Comp., 35, pp. 773–782 (1980)
11. Paisley, J., Wang, C., Blei, D.: The discrete infinite logistic normal distribution for mixed-membership modeling. in Proc. of AISTATS'11, pp. 74–82 (2011)
12. Salakhutdinov, R., Hinton, G.: Replicated Softmax: an Undirected Topic Model. NIPS 23, pp. 1607–1614 (2009)
13. Wan, L., Zhu, L., Fergus, R.: A Hybrid Neural Network-Latent Topic Model. in Proc. of AISTATS'12, pp. 1287–1294 (2012)