ChronoSAGE: Diversifying Topic Modeling Chronologically

Abstract. In this paper, we propose a new chronological modeling of topics latent in documents. We apply sparse additive generative models (SAGE) [5] in a manner so that we diversify topic modeling results chronologically by using document timestamps. We call our approach *ChronoSAGE*. SAGE can represent each word probability by exponential of the sum of multiple parameters representing various facets of documents. Therefore, we prepare three types of parameter to utilize document timestamps: the parameters for each topic, those for each timestamp, and those for each pair of topic and timestamp. Consequently, word tokens are generated not only in a topic-specific manner, but also in a time-specific manner. We first compare ChronoSAGE and vanilla SAGE with LDA in terms of pointwise mutual information (PMI) [10] to show the practical effectiveness of SAGE-type approaches. We then give examples of time-differentiated latent topics obtained by ChronoSAGE to show the usefulness of our chronological topic modeling. As another contribution, we also provide an approximated inference that makes the implementation far easier.

1 Introduction

Topic modeling approach prevails in the field of text mining research, because it provides a clear and compact representation of a wide variety of topics, which are latent and intertwined in large document sets. Latent Dirichlet allocation (LDA) [4] represents each latent topic as a probability distribution over words and extracts a predefined number, say K, of topics from a given document set. Each such distribution gives a large probability to the words whose meaning is closely related to a particular subject. Consequently, LDA provides a summarizing view of the document set as word lists, each expressing a particular subject in a human readable way (cf. Figure 8 in [4]).

However, many recent applications of text mining require utilizing document metadata effectively to make topic modeling results more persuasive. Especially, *spatio-temporal* metadata of documents are mainly considered due to their importance in social media texts, newswire documents, academic articles, etc. In this paper, we focus on document timestamps and utilize them in topic modeling by making per-topic probability distributions over words dependent on timestamps. While we follow a similar line to existing proposals [3, 14, 11], we make our approach based on sparse additive generative models (SAGE) [5]. We adopt the multifaceted SAGE introduced in Section 5 of [5] to diversify topic modeling results *chronologically*. We call our approach *ChronoSAGE* and call the LDA-type SAGE in its simplest form, given in Section 4 of [5], as *vanilla SAGE*.

ChronoSAGE has three types of parameter for defining word probabilities: the parameters for each topic, those for each timestamp, and those for each pair of topic and timestamp. The parameters of the first type give omni-temporal word probabilities for each topic. Those of the second type are introduced to find the words trivially dependent on timestamps, e.g. "Sunday", "May", "2001", etc, which are not informative for chronological topic modeling. Those of the third type give time-differentiated word probabilities for each topic, which are the most important for our application. Consequently, ChronoSAGE outputs as many human readable word lists as timestamps for each topic. That is, when the number of timestamps is T, ChronoSAGE outputs TK word lists, each corresponding to a different per-topic and time-dependent distribution.

It may be considered as a disadvantage of ChronoSAGE that the number of parameters representing word probabilities is large, which is equal to $KW + TW + TKW$ when the number of different words is W. Vanilla SAGE and LDA only require KW parameters, because they only give omni-temporal word probabilities. However, this is not a disadvantage, because text mining in recent days is required to analyze document sets where the number of documents, say D, is larger than W in order of magnitude. Since W does not increase so rapidly as D, the number of parameters representing per-document topic probabilities, which amounts to DK, is more critical. Further, it may be argued that the inherent relationship between timestamps should be considered by making the parameters at timestamp t dependent on those at $t-1$ as in [3, 14, 11]. In this paper, we take a different approach and assume that the probability of word w in topic k at timestamp t is derived from the omni-temporal probability of word w in topic k, not from the corresponding probability at timestamp $t-1$.

This paper provides another contribution aside from the utilization of document timestamps in SAGE. We provide a new approximated inference applicable to any version of SAGE. This revised inference uses Newton-Raphson method only in the single variable case, as well as avoiding using Hessian matrices. Consequently, we need no calls of quasi-Newton method like L-BFGS and can make implementation far easier. In sum, the contributions of this paper are 1) to provide a novel application of SAGE standing on its own merit and 2) to devise a new inference for SAGE that makes parameter updates easier to implement.

In the evaluation experiment, we first compare ChronoSAGE and vanilla SAGE with LDA with respect to their basic competence in topic modeling. While perplexity is widely used as an evaluation measure for topic models, we adopt an external evaluation measure, called pointwise mutual information (PMI) [10], to achieve a realistic evaluation. PMI is calculated by using the entire English Wikipedia that was downloaded on June 6, 2013 and contains 7,298,899 entries.

Table 1. Definition of symbols.

$x_{di} \in \mathcal{W}$	the word observed as the ith token of document d
$z_{di} \in \mathcal{K}$	the latent topic to which the ith token of document d is assigned
$y_d \in \mathcal{T}$	the observed timestamp of document d
n_d	the number of word tokens appearing in document d
n_w	the frequency of word w in the entire document set
m_w	the background parameter for word w
$\eta^{(1)}_{kw}$	the parameter for word w with respect to topic k
$\eta^{(2)}_{tw}$	the parameter for word w with respect to timestamp t
$\eta^{(3)}_{tkw}$	the parameter for word w with respect to the pair of timestamp t and topic k
ϕ_{tkw}	the probability that w expresses topic k in the document having timestamp t.
$\tau^{(1)}_{kw}$	the variance of the Gaussian distribution generating $\eta^{(1)}_{kw}$
$\tau^{(2)}_{tw}$	the variance of the Gaussian distribution generating $\eta^{(2)}_{tw}$
$\tau^{(3)}_{tkw}$	the variance of the Gaussian distribution generating $\eta^{(3)}_{tkw}$
θ_{dk}	the probability that a word token in document d represents topic k

The size of our reference corpus is enough to make our evaluation reliable. The result will show that both ChronoSAGE and vanilla SAGE can give a better PMI than LDA. Further, we present time-dependent word lists extracted by ChronoSAGE and discuss them from a qualitative view point. The discussion illustrates that ChronoSAGE can diversify topic modeling results chronologically without harming the basic competence vanilla SAGE has as a topic model.

The rest of the paper is organized as follows. Section 2 describes the model structure of ChronoSAGE and its variational inference. Section 3 presents the results of evaluation experiment. Section 4 reviews existing approaches. Section 5 concludes the paper with a summary and a discussion on future work.

2 ChronoSAGE

ChronoSAGE is an application of the multifaceted SAGE for utilizing document timestamps. However, our description of ChronoSAGE is more than a simple repetition of that given in [5]. ChronoSAGE has its own application-dependent characteristics and further is equipped with a new approximated inference.

2.1 Generative description

In this paper, we identify documents, words, topics, and timestamps with its index number. $\mathcal{D} = \{1, \dots, D\}$ is the set of documents, $\mathcal{W} = \{1, \dots, W\}$ is the set of different words, $\mathcal{K} = \{1, \dots, K\}$ is the set of latent topics, and $\mathcal{T} = \{1, \dots, T\}$ is the set of document timestamps. Table 1 contains the definition of symbols.

ChronoSAGE generates documents as follows.

– With respect to each word $w \in \mathcal{W}$, draw parameters $\tau^{(1)}_{kw}$ for each $k \in \mathcal{K}$, $\tau^{(2)}_{tw}$ for each $t \in \mathcal{T}$, and $\tau^{(3)}_{tkw}$ for each pair $(t, k) \in \mathcal{T} \times \mathcal{K}$ from the improper Jeffrey's prior distribution $p(\tau) \propto 1/\tau$. These are variance parameters. We adopt the Jeffrey's prior to reduce the number of free parameters.

– With respect to each $w \in \mathcal{W}$, draw parameters $\eta^{(1)}_{kw}$, $\eta^{(2)}_{tw}$, and $\eta^{(3)}_{tkw}$ as follows:

  - For each k, draw $\eta^{(1)}_{kw}$ from the zero-mean Gaussian distribution $\mathcal{N}(0, \tau^{(1)}_{kw})$.
  - For each t, draw $\eta^{(2)}_{tw}$ from the zero-mean Gaussian $\mathcal{N}(0, \tau^{(2)}_{tw})$.
  - For each pair (t, k), draw $\eta^{(3)}_{tkw}$ from the zero-mean Gaussian $\mathcal{N}(0, \tau^{(3)}_{tkw})$.
- Obtain the probability ϕ_{tkw} that word w is used to express topic k in the documents having timestamp t as follows:

$$\phi_{tkw} \equiv \frac{\exp(m_w + \eta^{(1)}_{kw} + \eta^{(2)}_{tw} + \eta^{(3)}_{tkw})}{\sum_v \exp(m_v + \eta^{(1)}_{kv} + \eta^{(2)}_{tv} + \eta^{(3)}_{tkv})} \,. \tag{1}$$

- For each document $d \in \mathcal{D}$, draw a multinomial parameter $\boldsymbol{\theta}_d = (\theta_{d1}, \ldots, \theta_{dK})$ from the symmetric Dirichlet prior Dirichlet(α). Further,
  - For the ith word token of document d, draw a latent topic z_{di} from the multinomial distribution Multi$(\boldsymbol{\theta}_d)$ and draw a word x_{di} from the multinomial Multi$(\boldsymbol{\phi}_{y_d z_{di}})$ as the ith token of document d.

The important feature of ChronoSAGE is that each word probability ϕ_{tkw} is obtained by combining the four parameters: m_w, $\eta^{(1)}_{kw}$, $\eta^{(2)}_{tw}$, and $\eta^{(3)}_{tkw}$. In vanilla SAGE, we set $\phi_{kw} \propto \exp(m_w + \eta_{kw})$ and use no time-specific word probabilities. The role played by each parameter in ChronoSAGE is explained below.

- m_w is equivalent to the log of the background probability of word w, because it depends neither on topic k nor on timestamp t. While m_w is treated as a constant in [5], we update m_w in the inference as will be shown later.
- $\eta^{(1)}_{kw}$ represents the dependency of the probability of word w on topic k. This parameter reflects one of the key ideas in topic modeling, i.e., an idea that different word probability distributions correspond to different topics.
- $\eta^{(2)}_{tw}$ represents the dependency of the probability of word w on timestamp t. This parameter is introduced to find the words showing a non-informative time-dependency. Table 2 gives an example of top 10 words sorted by their $\eta^{(2)}_{tw}$ for each timestamp t. This example is obtained from a topic modeling result of ChronoSAGE for TDT4, which is one among the three document sets used in our experiment. In TDT4 document set, we give the same timestamp to the documents belonging to the same range of seven days (e.g. from December 14 to 20, 2000). Table 2 shows that, for almost all timestamps, top seven words are the dates falling in the corresponding range of seven days. $\eta^{(2)}_{tw}$ is introduced to remove this type of trivial dependency on timestamps.
- $\eta^{(3)}_{tkw}$ is the most important, because we devise ChronoSAGE to diversify topic modeling results chronologically. $\eta^{(3)}_{tkw}$ represents a time-dependent displacement from $\eta^{(1)}_{kw}$. This type of parameter tells how topic-specific word probabilities are diversified according to document timestamps. We will later give examples of word lists sorted by $\eta^{(3)}_{tkw}$, where the words having trivial time-dependency do not appear owing to the introduction of $\eta^{(2)}_{tw}$.

We believe that the discussion above will prove the uniqueness of our approach. While we introduce no modification into the model structure of the multifaceted SAGE proposed in [5], our application of it for chronological topic analysis stands on its own application-dependent merit.

2.2 Variational lower bound

Based on the generative description of ChronoSAGE in Section 2.1, we obtain the following full joint distribution:

$$
\begin{aligned}
&p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\eta}^{(1)}, \boldsymbol{\eta}^{(2)}, \boldsymbol{\eta}^{(3)}, \boldsymbol{\tau}^{(1)}, \boldsymbol{\tau}^{(2)}, \boldsymbol{\tau}^{(3)}, \boldsymbol{\theta} | \boldsymbol{m}, \boldsymbol{\alpha}) \\
&\propto \prod_{k,w} \Big[\frac{1}{\tau^{(1)}_{kw}} \cdot \frac{\exp\big\{-(\eta^{(1)}_{kw})^2/(2\tau^{(1)}_{kw})\big\}}{\sqrt{2\pi\tau^{(1)}_{kw}}} \Big] \cdot \prod_{t,w} \Big[\frac{1}{\tau^{(2)}_{tw}} \cdot \frac{\exp\big\{-(\eta^{(2)}_{tw})^2/(2\tau^{(2)}_{tw})\big\}}{\sqrt{2\pi\tau^{(2)}_{tw}}} \Big] \\
&\cdot \prod_{t,k,w} \Big[\frac{1}{\tau^{(3)}_{tkw}} \cdot \frac{\exp\big\{-(\eta^{(3)}_{tkw})^2/(2\tau^{(3)}_{tkw})\big\}}{\sqrt{2\pi\tau^{(3)}_{tkw}}} \Big] \cdot \prod_d \frac{\mathit{\Gamma}(K\alpha)}{\mathit{\Gamma}(\alpha)^K} \prod_k \theta_{dk}^{\alpha-1} \\
&\cdot \prod_{d=1}^{D} \prod_{i=1}^{n_d} \prod_{k=1}^{K} \Big\{ \theta_{dk} \cdot \frac{\exp(m_{x_{di}} + \eta^{(1)}_{kx_{di}} + \eta^{(2)}_{y_d x_{di}} + \eta^{(3)}_{y_d k x_{di}})}{\sum_v \exp(m_v + \eta^{(1)}_{kv} + \eta^{(2)}_{y_d v} + \eta^{(3)}_{y_d k v})} \Big\}^{\delta(z_{di}=k)} , \qquad (2)
\end{aligned}
$$

where $\delta(\cdot)$ is 1 if the condition in the parentheses is true and is 0 otherwise.

With respect to $\boldsymbol{\eta}^{(1)}$, $\boldsymbol{\eta}^{(2)}$, and $\boldsymbol{\eta}^{(3)}$, we optimize them directly. With respect to the other parameters, we obtain their posteriors by a variational inference. Let $q(\boldsymbol{z}, \boldsymbol{\tau}^{(1)}, \boldsymbol{\tau}^{(2)}, \boldsymbol{\tau}^{(3)}, \boldsymbol{\theta})$ denote a variational posterior. We assume that this posterior is factorized as $q(\boldsymbol{z})q(\boldsymbol{\tau}^{(1)})q(\boldsymbol{\tau}^{(2)})q(\boldsymbol{\tau}^{(3)})q(\boldsymbol{\theta})$ and obtain a lower bound of the log of the marginal probability distribution $p(\boldsymbol{x}, \boldsymbol{\eta}^{(1)}, \boldsymbol{\eta}^{(2)}, \boldsymbol{\eta}^{(3)} | \boldsymbol{m}, \alpha)$ by using Jensen's inequality as follows:

$$
\begin{aligned}
&\ln p(\boldsymbol{x}, \boldsymbol{\eta}^{(1)}, \boldsymbol{\eta}^{(2)}, \boldsymbol{\eta}^{(3)} | \boldsymbol{m}, \alpha) \\
&\geq \int q(\boldsymbol{\tau}^{(1)}) \ln p(\boldsymbol{\eta}^{(1)} | \boldsymbol{\tau}^{(1)}) d\boldsymbol{\tau}^{(1)} - D[q(\boldsymbol{\tau}^{(1)}) \parallel p(\boldsymbol{\tau}^{(1)})] \\
&\quad + \int q(\boldsymbol{\tau}^{(2)}) \ln p(\boldsymbol{\eta}^{(2)} | \boldsymbol{\tau}^{(2)}) d\boldsymbol{\tau}^{(2)} - D[q(\boldsymbol{\tau}^{(2)}) \parallel p(\boldsymbol{\tau}^{(2)})] \\
&\quad + \int q(\boldsymbol{\tau}^{(3)}) \ln p(\boldsymbol{\eta}^{(3)} | \boldsymbol{\tau}^{(3)}) d\boldsymbol{\tau}^{(3)} - D[q(\boldsymbol{\tau}^{(3)}) \parallel p(\boldsymbol{\tau}^{(3)})] \\
&\quad + \int \sum_{\boldsymbol{z}} q(\boldsymbol{\theta}) q(\boldsymbol{z}) \ln p(\boldsymbol{z} | \boldsymbol{\theta}) d\boldsymbol{\theta} + \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \ln p(\boldsymbol{x} | \boldsymbol{m}, \boldsymbol{z}, \boldsymbol{\eta}^{(1)}, \boldsymbol{\eta}^{(2)}, \boldsymbol{\eta}^{(3)}) \\
&\quad + \int q(\boldsymbol{\theta}) \ln p(\boldsymbol{\theta} | \boldsymbol{\alpha}) d\boldsymbol{\theta} - \int q(\boldsymbol{\theta}) \ln q(\boldsymbol{\theta}) d\boldsymbol{\theta} - \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \ln q(\boldsymbol{z}) \ , \qquad (3)
\end{aligned}
$$

where $D[q(\boldsymbol{\tau}) \parallel p(\boldsymbol{\tau})] \equiv \int q(\boldsymbol{\tau}) \ln q(\boldsymbol{\tau}) d\boldsymbol{\tau} - \int q(\boldsymbol{\tau}) \ln p(\boldsymbol{\tau}) d\boldsymbol{\tau}$, i.e., Kullback-Leibler divergence. We further assume the followings for the posteriors:

- $q(\boldsymbol{\tau}^{(1)})$ is factorized as $\prod_{k,w} q(\tau^{(1)}_{kw})$. Each $q(\tau^{(1)}_{kw})$ is a Gamma distribution $\mathrm{Gamma}(a^{(1)}_{kw}, b^{(1)}_{kw})$. We assume the same for $q(\boldsymbol{\tau}^{(2)})$ and $q(\boldsymbol{\tau}^{(3)})$.
- $q(\boldsymbol{\theta})$ is factorized as $\prod_d q(\boldsymbol{\theta}_d)$. Each $q(\boldsymbol{\theta}_d)$ is a Dirichlet distribution $\mathrm{Dirichlet}(\boldsymbol{\zeta}_d)$.
- $q(\boldsymbol{z})$ is factorized as $\prod_{d,i} q(z_{di})$. z_{di} is drawn from a multinomial distribution $\mathrm{Multi}(\boldsymbol{\lambda}_{di})$, where λ_{dik} is a variational probability that $z_{di} = k$ holds.

Table 2. Top 10 words sorted by their η_{tw} for each t in case of TDT4.

$t = 0$	edt paralymp lebanon 32nd wild-card u.s china russia join carter's
$t = 1$	kippur 10-13 lebanon china palestinian text join iran parti dynamit
$t = 2$	10-14 10-16 10-18 10-15 10-19 10-17 10-20 sharm lebanon edt
$t = 3$	10-24 10-23 10-22 10-25 10-21 10-26 10-27 lebanon china octob
$t = 4$	10-29 10-28 10-31 10-30 11-3 leipzig lebanon stump join 11-1
$t = 5$	11-10 11-8 11-9 11-6 11-7 11-5 convuls 11-4 russia clinton
$t = 6$	11-17 11-16 11-11 11-14 11-15 11-12 11-13 anchorag china russia
$t = 7$	11-18 11-19 11-24 11-22 11-23 11-20 11-21 930-vote china taint
$t = 8$	11-25 11-27 11-28 11-26 11-30 11-29 seclus join novemb bush
$t = 9$	12-8 12-6 12-5 12-7 12-3 537-vote 12-4 russia parti novemb
$t = 10$	12-12 12-15 12-14 12-10 12-13 12-11 12-9 decemb join novemb
$t = 11$	12-17 12-18 12-21 12-20 12-19 12-22 12-16 ronni veneman china
$t = 12$	12-24 12-28 12-29 12-23 12-27 12-26 12-25 alcoa jiri holidai
$t = 13$	309 tabasco 2001 1-5 vy 12-0 free-agent lighten 31st 1-4
$t = 14$	presid-elect's 1-12 1-8 1-11 1-9 1-10 1-7 70-year-old u.s tycoon
$t = 15$	1-14 1-13 1-19 1-18 1-17 1-16 1-15 rosa 560 lyle
$t = 16$	1-21 1-26 1-25 1-22 1-20 1-23 1-24 hanun faizabad taba
$t = 17$	1-28 1-31 1-30 1-27 1-29 dawosi bhuj fasa greenspan's 960

2.3 Parameter updates

We denote the right hand side of Eq. (3) as $\mathcal{L}$ for short. By maximizing $\mathcal{L}$ with respect to each parameter, we obtain a formula for updating the parameter.

The terms related to topic assignments $\boldsymbol{z}$ in $\mathcal{L}$ can be rewritten as follows:

$$\begin{aligned} L_{\boldsymbol{z}} &= \sum_{d,i,k} \lambda_{dik} \{ \Psi(\zeta_{dk}) - \Psi(\sum_k \zeta_{dk}) \} \\ &+ \sum_{d,i,k} \lambda_{dik} \ln \frac{\exp(m_{x_{di}} + \eta^{(1)}_{kx_{di}} + \eta^{(2)}_{y_d x_{di}} + \eta^{(3)}_{y_d k x_{di}})}{\sum_w \exp(m_w + \eta^{(1)}_{kw} + \eta^{(2)}_{y_d w} + \eta^{(3)}_{y_d k w})} - \sum_{d,i,k} \lambda_{dik} \ln \lambda_{dik} \ , \end{aligned} \tag{4}$$

where $\Psi(\cdot)$ is digamma function. λ_{dik} is a variational posterior probability that the ith word token in document d is assigned to topic k. ζ_{dk} is a variational Dirichlet posterior parameter of topic k in document d. By maximizing $L_{\boldsymbol{z}}$ with respect to λ_{dik}, we obtain the following update for λ_{dik}:

$$\lambda_{dik} \propto \exp\Big\{ \Psi(\zeta_{dk}) - \Psi\big(\sum_k \zeta_{dk}\big) \Big\} \cdot \frac{\exp(m_{x_{di}} + \eta^{(1)}_{kx_{di}} + \eta^{(2)}_{y_d x_{di}} + \eta^{(3)}_{y_d k x_{di}})}{\sum_w \exp(m_w + \eta^{(1)}_{kw} + \eta^{(2)}_{y_d w} + \eta^{(3)}_{y_d k w})} \ . \tag{5}$$

We rewrite the terms related to per-document topic distributions $\boldsymbol{\theta}$ in $\mathcal{L}$ as:

$$\begin{aligned} L_{\boldsymbol{\theta}} &= \sum_{d,i,k} \lambda_{dik} \Big\{ \Psi(\zeta_{dk}) - \Psi(\sum_k \zeta_{dk}) \Big\} + \sum_{d,k} (\alpha_k - \zeta_{dk}) \Big\{ \Psi(\zeta_{dk}) - \Psi(\sum_k \zeta_{dk}) \Big\} \\ &+ \sum_d \log \Gamma(\sum_k \alpha_k) - \sum_{d,k} \log \Gamma(\alpha_k) - \sum_d \log \Gamma(\sum_k \zeta_{dk}) + \sum_{d,k} \log \Gamma(\zeta_{dk}). \end{aligned} \tag{6}$$

By solving $\partial L_{\boldsymbol{\theta}} / \partial \zeta_{dk} = 0$, we obtain an update as $\zeta_{dk} = \alpha_k + n_{dk}$, where we define $n_{dk} \equiv \sum_i \lambda_{dik}$. The derivation is the same as that of the vanilla LDA [4].

The terms related to variance parameters $\boldsymbol{\tau}^{(1)}$ in $\mathcal{L}$ can be rewritten as:

$$L(\boldsymbol{\tau}^{(1)}) = \sum_{k,w} \Big\{ - \frac{b^{(1)}_{kw}(\eta^{(1)}_{kw})^2}{2(a^{(1)}_{kw} - 1)} - \Big(a^{(1)}_{kw} + \frac{1}{2}\Big)\Psi(a^{(1)}_{kw}) + \frac{\ln b^{(1)}_{kw}}{2} + a^{(1)}_{kw} + \ln \Gamma(a^{(1)}_{kw})\Big\} + const. \tag{7}$$

By solving $\partial L(\boldsymbol{\tau}^{(1)})/\partial \tau_{kw} = 0$, we obtain the following updates for $a^{(1)}_{kw}$ and $b^{(1)}_{kw}$:

$$a^{(1)}_{kw} \leftarrow a^{(1)}_{kw} + \frac{2b^{(1)}_{kw}(\eta^{(1)}_{kw})^2(a^{(1)}_{kw}-1)^{-2} - \big(a^{(1)}_{kw} + \frac{1}{2}\big)\Psi'(a^{(1)}_{kw}) + 1}{b^{(1)}_{kw}(\eta^{(1)}_{kw})^2(a^{(1)}_{kw}-1)^{-3} + \Psi'(a^{(1)}_{kw}) + \big(a^{(1)}_{kw} + \frac{1}{2}\big)\Psi''(a^{(1)}_{kw})} \tag{8}$$

$$b^{(1)}_{kw} \leftarrow (a^{(1)}_{kw} - 1)(\eta^{(1)}_{kw})^{-2} \tag{9}$$

Similar updates can also be obtained for $\boldsymbol{\tau}^{(2)}$ and $\boldsymbol{\tau}^{(3)}$. Eq. (9) will be useful in the discussion below when we eliminate $a^{(1)}_{kw}$ and $b^{(1)}_{kw}$ from parameter updates.

We estimate $\boldsymbol{\eta}^{(1)}$, $\boldsymbol{\eta}^{(2)}$, and $\boldsymbol{\eta}^{(3)}$ by maximizing $\mathcal{L}$ directly. A function to be maximized with respect to $\boldsymbol{\eta}^{(1)}$ is:

$$L(\boldsymbol{\eta}^{(1)}) = \sum_{k,w} n_{kw}\eta^{(1)}_{kw} - \sum_{t,k} n_{tk} \ln \Big\{ \sum_w \exp(m_w + \eta^{(1)}_{kw} + \eta^{(2)}_{tw} + \eta^{(3)}_{tkw})\Big\} - \frac{b^{(1)}_{kw}(\eta^{(1)}_{kw})^2}{2(a^{(1)}_{kw} - 1)} + const., \tag{10}$$

where we define $n_{kw} \equiv \sum_d \sum_{\{i:x_{di}=w\}} \lambda_{dik}$ and $n_{tk} \equiv \sum_{\{d:y_d=t\}} \sum_i \lambda_{dik}$.

Here we propose a new approximation to avoid inverting Hessian matrices (cf. Section 3.1 of [5]) and thus to make variational inferences far easier to implement.

Our approximation is based on the observation that the log function satisfies $\ln x \leq \frac{x}{\xi} - 1 + \ln \xi$ for any $\xi > 0$. This observation is also used by [2] in a different situation. By introducing an auxiliary variable ξ_{tk} for each pair of timestamp t and latent topic k, we obtain a lower bound of Eq. (10) as follows:

$$L(\boldsymbol{\eta}^{(1)}) \geq \sum_{k,w} n_{kw}\eta^{(1)}_{kw} - \sum_{t,k} n_{tk} \frac{\sum_w \exp(m_w + \eta^{(1)}_{kw} + \eta^{(2)}_{tw} + \eta^{(3)}_{tkw})}{\xi_{tk}} - \sum_{t,k} n_{tk} \ln \xi_{tk} - \sum_{k,w} \frac{(\eta^{(1)}_{kw})^2}{2(a^{(1)}_{kw} - 1)b^{(1)}_{kw}} + const. \tag{11}$$

We denote this lower bound as $l(\boldsymbol{\eta}^{(1)})$ and maximize it in place of $L(\boldsymbol{\eta}^{(1)})$. The first and the second derivatives of $l(\boldsymbol{\eta}^{(1)})$ with respect to $\eta^{(1)}_{kw}$ are obtained as:

$$\frac{\partial l(\boldsymbol{\eta}^{(1)})}{\partial \eta^{(1)}_{kw}} = n_{kw} - e^{\eta^{(1)}_{kw}} \sum_t \frac{n_{tk} \exp(m_w + \eta^{(2)}_{tw} + \eta^{(3)}_{tkw})}{\xi_{tk}} - \frac{\eta^{(1)}_{kw}}{(a^{(1)}_{kw} - 1)b^{(1)}_{kw}} \ ,$$

$$\frac{\partial^2 l(\boldsymbol{\eta}^{(1)})}{\partial {\eta^{(1)}_{kw}}^2} = -e^{\eta^{(1)}_{kw}} \sum_t \frac{n_{tk} \exp(m_w + \eta^{(2)}_{tw} + \eta^{(3)}_{tkw})}{\xi_{tk}} - \frac{1}{(a^{(1)}_{kw} - 1)b^{(1)}_{kw}} \ . \tag{12}$$

Table 3. Specifications of the three document sets used in the experiment.

	D: # documents	W: # words	T: # timestamps	average document length
DBLP	2,093,913	10,694	22	5.2
NSF	128,181	19,066	13	95.9
TDT4	96,246	15,153	18	156.6

Further, we use Eq. (9) to eliminate $a_{kw}^{(1)}$ and $b_{kw}^{(1)}$ and obtain a Newton-Raphson update of $\eta_{kw}^{(1)}$ as follows:

$$\eta_{kw}^{(1)} \leftarrow \eta_{kw}^{(1)} + \frac{\{n_{kw} - C_{kw} e^{\eta_{kw}^{(1)}}\}(\eta_{kw}^{(1)})^2 - \eta_{kw}^{(1)}}{C_{kw} e^{\eta_{kw}^{(1)}}(\eta_{kw}^{(1)})^2 + 1} \ , \tag{13}$$

where we define $C_{kw} \equiv \sum_t n_{tk} \exp(m_w + \eta_{tw}^{(2)} + \eta_{tkw}^{(3)})/\xi_{tk}$. A similar simple update can be obtained also for $\eta_{tw}^{(2)}$ and $\eta_{tkw}^{(3)}$. While m_w is kept as a constant in the original paper of SAGE [5], we update m_w by maximizing $\mathcal{L}$. The relevant terms in $\mathcal{L}$ can be rewritten as:

$$L(\boldsymbol{m}) \geq \sum_w n_w m_w - \sum_{t,k} n_{tk} \frac{\sum_w \exp(m_w + \eta_{kw}^{(1)} + \eta_{tw}^{(2)} + \eta_{tkw}^{(3)})}{\xi_{tk}} \ . \tag{14}$$

Let $l(\boldsymbol{m})$ denote the right hand side of Eq. (14). By solving $\partial l(\boldsymbol{m})/\partial m_k = 0$, we obtain an update of m_w as:

$$m_w \leftarrow \ln \frac{n_w}{\sum_{t,k} n_{tk} \exp(\eta_{kw}^{(1)} + \eta_{tw}^{(2)} + \eta_{tkw}^{(3)})/\xi_{tk}} \ . \tag{15}$$

We update m_w also for vanilla SAGE by $m_w \leftarrow \ln \frac{n_w}{\sum_k n_k \exp(\eta_{kw}^{(1)})/\xi_k}$ in the experiment. By differentiating the lower bound achieved by our new approximation with respect to ξ_{tk}, we obtain the following update: $\xi_{tk} \leftarrow \sum_w \exp\left(m_w + \eta_{kw}^{(1)} + \eta_{tw}^{(2)} + \eta_{tkw}^{(3)}\right)$. For the Dirichlet hyperparameter α, we used a fixed value $50/K$, because its optimization gave no substantial difference in evaluation results.

3 Evaluation experiment

We perform an evaluation of ChronoSAGE in two phases. Firstly, we compare ChronoSAGE and vanilla SAGE with LDA. This comparison will reveal that ChronoSAGE has almost the same topic modeling competence with vanilla SAGE and that ChronoSAGE and vanilla SAGE are superior to LDA. Secondly, we give examples of timestamped word lists extracted by ChronoSAGE and discuss them from a qualitative viewpoint. This discussion will reveal that ChronoSAGE successfully diversify topic modeling results chronologically. Before giving the results of evaluation, we describe experiment settings in detail.

We used three document sets, called DBLP, NSF, and TDT4, whose specifications are summarized in Table 3. DBLP is a set of paper titles in DBLP computer science bibliography, available at its Web site[1]. We used a version of `dblp.xml` downloaded on June 11, 2013. We removed all records whose publication year was 2013, because the number of such records was small. We regarded paper title as document and publication year as document timestamp. NSF is a set of research awards abstracts available at the UCI machine learning repository[2]. Also in this document set, we regarded publication year as timestamp. TDT4 is a corpus for the TDT4 topic detection and tracking evaluation by LDC[3]. In TDT4, we gave the same timestamp to the documents belonging to the same chronological range of seven days (e.g. from December 14 to 20, 2000)[4]. We preprocessed each document set by a series of standard procedures. However, stemming was not applied to DBLP, because paper titles were short, and therefore word forms were thought to play an important role.

We ran the variational inference presented in Section 2 on each document set. The inference for vanilla SAGE was achieved by ignoring time-dependent parameters in the inference for ChronoSAGE. Before staring an instance of the inference, we conducted 500 iterations of collapsed Gibbs sampling (CGS) for LDA [6] and initialized $\eta^{(1)}_{kw}$s based on the topic assignment result as $\ln p(w|k) - m_w$, where $p(w|k)$ is the probability of word w within topic k. $\eta^{(2)}_{tw}$s and $\eta^{(3)}_{tkw}$s were initialized to 1. After 500 iterations of CGS, we ran 100 iterations of the variational inference. We confirmed that this number of iterations was enough by inspecting the change in the variational lower bound. With respect to K, we tested the following two settings: $K = 100$ and $K = 300$. For each of the compared approaches, i.e, LDA, vanilla SAGE, and ChronoSAGE, and for each setting of K, we ran the variational inference ten times starting from a random initialization of topic assignments in CGS. Consequently, we obtained ten topic modeling results for each compared approach and for each setting of K.

3.1 Comparison using external measure

While perplexity is often used for an evaluation of topic models, we adopted an external measure, called pointwise mutual information (PMI) [10], for a more realistic evaluation. We did not use coherence measure [9], because this measure is likely to give a worse result for a larger number of topics, as we can observe in Figure 6 of [1], and thus makes the comparison between different K difficult. We used the entire English Wikipedia, which was downloaded on June 6, 2013 and contains 7,298,899 entries, as the reference corpus for PMI.

The evaluation was done as follows. We selected top 10 words $(w_1, \ldots, w_{10})$ sorted by $\eta^{(1)}_{kw}$ for each k and calculated PMI for all pairs of words as $\mathrm{PMI}(w_i, w_j) = \ln \frac{p(w_i, w_j)}{p(w_i)p(w_j)}$, for $i, j \in \{1, \ldots, 10\}$. The probability $p(w_i)$ is defined as R_i/R,

[1] http://dblp.uni-trier.de/xml/

[2] http://archive.ics.uci.edu/ml/

[3] http://projects.ldc.upenn.edu/TDT4/

[4] We make the first range contain from December 1 to 6, 2000 and the last one contain from January 27 to 31, 2001 so that the sizes of these two ranges, placed at both ends of the whole period, are as equal as possible.

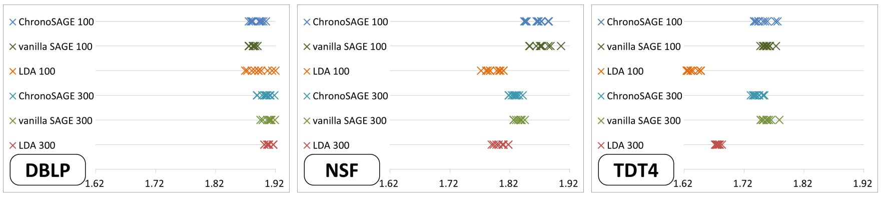

Fig. 1. Comparing ChronoSAGE with vanilla SAGE and LDA in PMI on DBLP (left), NSF (center), and TDT4 (right).

where R_i is the number of documents containing w_i in the reference corpus, and R is the size of the reference corpus. The co-occurrence probability $p(w_i, w_j)$ is defined as R_{ij}/R, where R_{ij} is the number of documents containing both w_i and w_j in the reference corpus. We compared the three approaches by the median of all calculated PMIs. A larger median is better.

Fig. 1 summarizes the evaluation. Ten medians obtained from the ten different instances of the inference, each starting from a random initialization of topic assignments in CGS, are plotted for each approach and for each K. The horizontal axis represents the magnitude of PMI. As Fig. 1 shows, ChronoSAGE gave almost the same medians as vanilla SAGE. Further, both methods worked better than LDA for both NSF and TDT4 and at least gave a result comparable with LDA for DBLP. Therefore, it can be concluded that SAGE-type topic modeling is a better choice than LDA in terms of PMI.

3.2 Timestamped word lists

Next, we give an example of timestamped word lists obtained by ChronoSAGE in Fig. 2. We obtained this example from one among the ten results ChronoSAGE gave for DBLP when $K = 300$. The two panels in Fig. 2 correspond to two among 300 topics. The left and the right panel give word lists seemingly related to mobile communications and to video coding, respectively. On the top of each panel, top 15 words are enumerated based on $\eta^{(1)}_{kw}$. These words represent the omni-temporal content of the corresponding latent topic. The size of an ellipse behind each word indicates the magnitude of $\eta^{(1)}_{kw}$. Below these top 15 words, we present top 10 words for each timestamp based on $\eta^{(3)}_{tkw}$. The size of a circle behind each timestamp indicates the largest $\eta^{(3)}_{tkw}$ for each t.

On the left panel in Fig. 2, we can read out a clear trend transition. For example, the word *GSM*, mainly related to 2G networks, appears in the word lists of earlier years. The word *GPRS* comes after it and appears in the lists of 2001 and 2002. The word *LTE* appears only in the lists of recent years. While we do not explicitly model the inherent relationships between timestamps, we can observe such a clear trend. The right panel provides an interesting observation. For example, the word *HDTV* cannot be found in the word lists after 1995. This may be because HDTV had already become a part of consumer technologies at that time. *MPEG-2* and *MPEG-4* are found in the lists of late 90's and

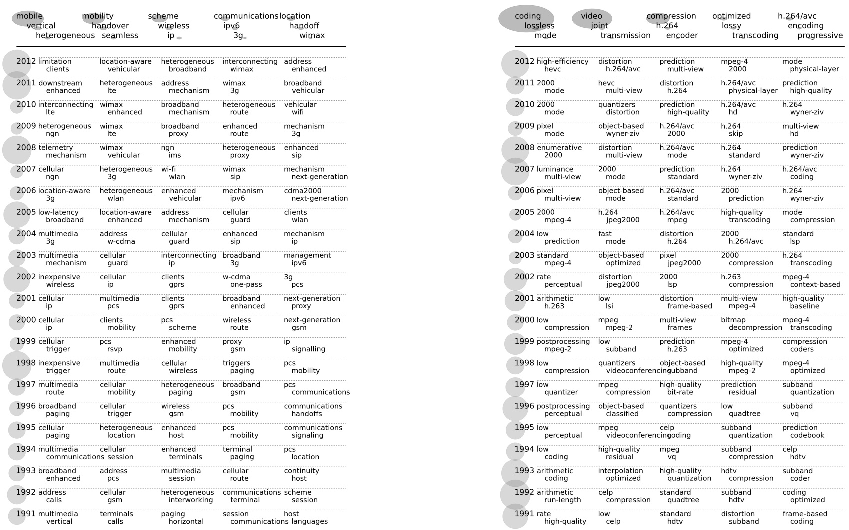

Fig. 2. Timestamped word lists extracted by ChronoSAGE from DBLP. The left and the right panels correspond to different latent topics seemingly related to mobile communications and to video coding, respectively.

early 2000's. *H.264* comes after them, and *HEVC* appears in the lists of very recent years. It can be concluded that ChronoSAGE extracts clear trends by diversifying topic modeling chronologically with document timestamps.

4 Existing approaches

Among existing approaches, the structural topic model (STM) [13] is closest to ChronoSAGE in its use of the multifaceted SAGE. The authors make word probabilities proportional to an exponential of the sum of the four parameters (cf. Eq. (9) in [13]) so that word use within a topic varies by multiple factors. This is also an application of the multifaceted SAGE and is similar to ours in this sense. However, the authors consider covariates, e.g. gender or political ideology, as the factors diversifying word probabilities. On the other hand, we use document timestamps to diversify word probabilities and clarify an application-dependent merit of ChronoSAGE through our experiment.

The dynamic topic model (DTM) [3] has time-dependent word probabilities that can be written as $\phi_{tkw} \equiv \exp(\eta^{(3)}_{tkw}) / \sum_v \exp(\eta^{(3)}_{tkv})$ by using our symbols. However, as is discussed in Section 2.1, it is important for us to remove a trivial time-dependency from each word probability ϕ_{tkw} by introducing a parameter $\eta^{(2)}_{tw}$ that is dependent only on timestamp t and not on any latent topics. This technical aspect differentiates ChronoSAGE from DTM.

Factorial LDA [8] has a similar flavor to SAGE, because an exponential of the sum of multiple parameters is used to vary word probabilities. However, the exponential is used to describe not word probabilities themselves, but hyperparameters of Dirichlet prior distributions that generate word probability distributions. Consequently, the inference requires the multivariate gradient ascent for optimizing the parameters. In contrast, our approximated inference only uses the Newton-Raphson method in the single variable case and makes the implementation easier.

5 Conclusions

In this paper, we proposed ChronoSAGE, a novel application of the multifaceted SAGE standing on its own merit. As the results of evaluation experiment revealed, ChronoSAGE has the same competence with vanilla SAGE in topic modeling and, however, can extract informative timestamped word lists, which cannot be obtained by vanilla SAGE. Further, we devised a new approximated inference using the Newton-Raphson method only in the single variable case. Our important future work is to explicitly model the inherent dependency among the timestamps by e.g. using Gaussian processes [12].

References

1. Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y.-C., Zhu., M.: A practical algorithm for topic modeling with provable guarantees. ICML (2013)
2. Blei, D. M., Lafferty, J. D.: Correlated topic models. NIPS (2005)
3. Blei, D. M., Lafferty, J. D.: Dynamic topic models. ICML, pp. 113–120 (2006)
4. Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent Dirichlet allocation. JMLR 3, pp. 993–1022 (2003)
5. Eisenstein, J., Ahmed A., Xing, E. P.: Sparse additive generative models of text. ICML, pp. 1041–1048 (2011)
6. Griffiths, T. L., Steyvers, M.: Finding scientific topics. PNAS, 101, Suppl 1, pp. 5288–5235 (2004)
7. Hoffman, M. D., Blei, D. M., Bach, F. R.: Online learning for latent Dirichlet allocation. NIPS, pp. 856–864 (2010)
8. Paul, M. J., Dredze, M.: Factorial LDA: sparse multi-dimensional models of text. NIPS (2012)
9. Mimno, D., Wallach, H., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. EMNLP (2011)
10. Newman, D., Karimi, S., Cavedon, L.: External evaluation of topic models. ADCS, pp. 11-18 (2009)
11. O'Connor, B., Stewart, B. M., Smith, N. A.: Learning to extract international relations from political context. ACL (2013)
12. Rasmussen, C. E.: Gaussian Processes for Machine Learning. MIT Press (2006)
13. Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S., Albertson, B., Rand, D.: Structural topic models for open-ended survey responses. American Journal of Political Science. to appear. (2013)
14. Wang, C., Blei, D. M., Heckerman, D.: Continuous time dynamic topic models. UAI, pp. 579–586 (2008)