Decision Support Systems 70 (2015) 1–14

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Detecting short-term cyclical topic dynamics in the user-generated content and news

Hsin-Min Lu [*]

Department of Information Management, National Taiwan University, Taipei 106, Taiwan

ARTICLE INFO

Article history:
Received 6 June 2014
Received in revised form 22 October 2014
Accepted 30 November 2014
Available online 6 December 2014

Keywords:
Topic models
Gibbs sampling
Temporal dynamics
Context dependent
Cyclical dynamics

ABSTRACT

With the maturation of the Internet and the mobile technology, Internet users are now able to produce and consume text data in different contexts. Linking the context to the text data can provide valuable information regarding users' activities and preferences, which are useful for decision support tasks such as market segmentation and product recommendation. To this end, previous studies have proposed to incorporate into topic models contextual information such as authors' identities and timestamps. Despite recent efforts to incorporate contextual information, few studies have focused on the short-term cyclical topic dynamics that connect the changes in topic occurrences to the time of day, the day of the week, and the day of the month. Short-term cyclical topic dynamics can both characterize the typical contexts to which a user is exposed at different occasions and identify user habits in specific contexts. Both abilities are essential for decision support tasks that are context dependent. To address this challenge, we present the Probit-Dirichlet hybrid allocation (PDHA) topic model, which incorporates a document's temporal features to capture a topic's short-term cyclical dynamics. A document's temporal features enter the topic model through the regression covariates of a multinomial-Probit-like structure that influences the prior topic distribution of individual tokens. By incorporating temporal features for monthly, weekly, and daily cyclical dynamics, PDHA is able to capture interesting short-term cyclical patterns that characterize topic dynamics. We developed an augmented Gibbs sampling algorithm for the non-Dirichlet-conjugate setting in PDHA. We then demonstrated the utility of PDHA using text collections from user generated content, newswires, and newspapers. Our experiments show that PDHA achieves higher hold-out likelihood values compared to baseline models, including latent Dirichlet allocation (LDA) and Dirichlet-multinomial regression (DMR). The temporal features for short-term cyclical dynamics and the novel model structure of PDHA both contribute to this performance advantage. The results suggest that PDHA is an attractive approach for decision support tasks involving text mining.

1. Introduction

Making informed decisions in our current fast-changing environment often demands the timely and comprehensive analysis of large amounts of text data. Researchers have attempted to address this challenge by developing text mining approaches such as topic models [1]. Topic models, including latent Dirichlet allocation (LDA) and its variations [2,3], have been applied to discover coherent topics, analyze trends in academic publications [4], and examine user-generated content from different sources [5].

Additional contextual information that signals the unobservable structures beneath the observed textual data can help a model extract better latent topics. Thus, an emerging research direction in topic models is to incorporate contextual information such as authors' identities and timestamps to extract latent topics that reveal the influence of changing contexts. For example, incorporating the authors' identities into a topic model [6] can improve performance because an author's specialties can reveal additional information regarding the latent topics of the document.

Incorporating contextual information into topic models also provides a direct route for topic models to support decision making. As a generative probabilistic model, the learned topic model is essentially the joint distribution of contextual information and textual data. By computing the conditional distributions of contextual information given observed textual data, a topic model can provide crucial information that supports decision-making employing contextual information. For example, by incorporating additional information of microblog users who shared news articles online, a model is able to recommend news articles to other microblog users who may have similar interests [7].

Despite recent progressions, few studies have focused on the short-term cyclical topic dynamics that connect changing topic occurrences to the time of day, the day of the week, and the day of the month. Content generated by social media and mobile platforms often reveal strong short-term cyclical dynamics because users' day-to-day routines heavily influence the contexts of use, which contribute to the variations of topic

[*] Tel.:+886 2 33661184.
E-mail address: luim@ntu.edu.tw.

http://dx.doi.org/10.1016/j.dss.2014.11.006
0167-9236/© 2014 Elsevier B.V. All rights reserved.

occurrence. By including short-term cyclical dynamics in a topic model, we are able to better characterize the cyclical dynamics that reflect users' activities, habits, and preferences [8], three factors that can improve decision support tasks such as market segmentation [9] and product recommendation [10].

To fill this gap, we introduce a new family of topic models that can discover short-term cyclical patterns from a document collection. The proposed Probit-Dirichlet hybrid allocation (PDHA) provides a general framework with which to link discrete and continuous document-specific exogenous temporal features to topic distributions. PDHA includes features for daily, weekly, and monthly cyclical patterns as a way of capturing short-term dynamics. In addition to the topic–token distributions and document–topic mixes provided by a typical topic model, PDHA learns the coefficients of these temporal features through a multinomial Probit-like structure; these coefficients can reveal the occurrences of topic changes within a day, week, or month. Unlike the topic over time (TOT) [11] and dynamic topic model (DTM) [12], which focus more on the long-term evolution of topics, PDHA can model short-term cyclical variations that may be harder to capture using other flavors of topic models. Moreover, our PDHA model includes random variables for document-specific topic tendencies. These random variables allow each document to deviate from the mean tendency specified by the temporal features while preserving a common theme for each document.

In the subsequent sections, we first review previously proposed time-dependent topic models. We then present the PDHA model and discuss in detail the Gibbs sampling algorithm. Afterward, we present experimental results that incorporate daily, weekly, and monthly cyclical patterns. We conclude with a short discussion of future research directions.

2. Literature review

Topic models [1,4] are a family of algorithms aimed at discovering latent structures in large document collections. Based on the assumption that observed tokens are governed by latent topics, topic models define a generative process that first generates the mixture of topics in a document and then select observed tokens conditioned on latent topics. The data generating process provides a rich structure that is capable of capturing meaningful latent topical structures in documents. Compared to their predecessors, such as the probabilistic latent semantic indexing (pLSI), topic models do not have the over-fitting problem and outperform pLSI in terms of perplexity [1].

The original topic models are often referred to as the latent Dirichlet allocation (LDA) because they adopt the conjugate prior for multinomial distribution, the Dirichlet distribution, to simplify computation. The idea of capturing short-term cyclical dynamics is related to the research stream that incorporates additional time-dependent information to improve LDA models. We review selected time-dependent topic models in this section. We refer readers to Blei [13] for a general introduction of topic models.

2.1. Time-dependent topic models

We start with an overview of the LDA model and then extend it to time-dependent topic models. For a document collection that contains D documents indexed by integers 1, 2, …, D, LDA assumes that the N_d tokens in the document d, $w_d = (w_{d1},\ w_{d2},\ \dots,\ w_{dN_d})$, were generated by first drawing the topic mix $\theta_d \sim Dir(\alpha)$, where $Dir(\alpha)$ is a Dirichlet distribution with symmetric concentration parameter α. The topic mix θ_d is a vector of length J, where J is the total number of topics in a document collection. Each element of θ_d is the probability of selecting the corresponding topic for a position in document d. All elements of θ_d sum to one.

The second step is to determine the topic for a token at position i, $1 \le i \le N_d$, by drawing $z_{di} \sim Multinomial(\theta_d)$. This process assumes that given the topic mix θ_d, the latent topic for each token in document d is independent of one another. Finally, a token at position i is determined by drawing from the corresponding topic–token distribution $w_{di} \sim Multinomial\left(\phi_{z_{di}}\right)$, where $\phi_{z_{di}}$ is a vector determining the probability that a token may appear given z_{di}, the topic at position i of document d. The length of each ϕ_j is the vocabulary size W for $j = 0, 1, \dots, J-1$. The model assumes that each ϕ_j is generated from a Dirichlet distribution with a symmetric concentration parameter β.

The generative process can be represented using the plate notation shown in Fig. 1. The shaded circle indicates observed variables, and the open circles indicate latent variables and parameters. Starting from the upper left, Panel (A) in Fig. 1 provides a summary for the data-generating process described above.

In the subsequent discussion, variables such as z_{di} and θ_d should be regarded as latent variables because the number of these variables grow with the size of the dataset [14]. Other variables, including α, β, and $\phi_{z_{di}}$ are regarded as parameters. The joint distribution of observed tokens, latent topic variables and other parameters conditional on α and β is given by:

$$p(\theta, \phi, Z, w|\alpha, \beta) = \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{j=0}^{J-1} p\left(\phi_j|\beta\right) \prod_{i=1}^{N_d} p(z_{di}|\theta_d) p\left(w_{di}|\phi_{z_{di}}\right), \quad (1)$$

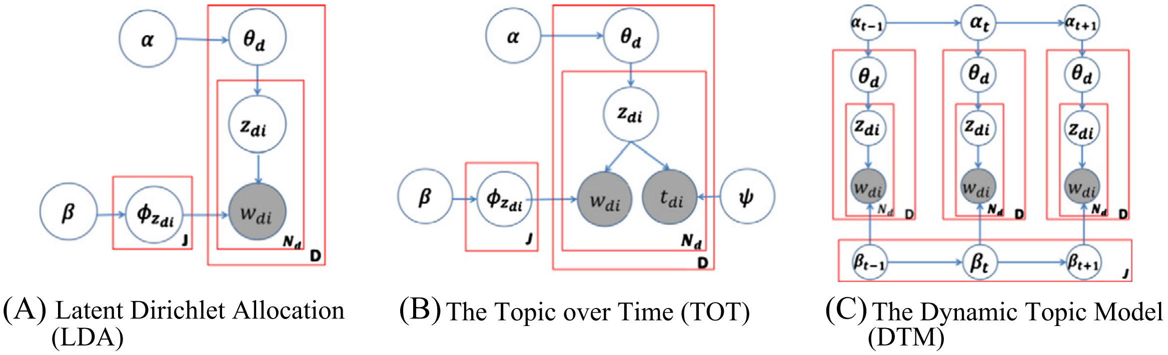
where $\theta = (\theta_1, \theta_2, \dots, \theta_D)$, $\phi = (\phi_1, \phi_2, \dots, \phi_J)$, $w = (w_1, w_2, \dots, w_D)$, $Z = (z_1, z_2, \dots, z_D)$, and $z_d = \left(z_{d1}, z_{d2}, \dots, z_{dN_d}\right)$. One challenge presented by topic models is to design efficient and effective algorithms for estimating θ, ϕ, and Z given w, α, and β. We will review model estimation methods later.

The LDA model does not explicitly include temporal features. However, simple post-processing can be used to determine time trends. As demonstrated by Griffiths and Steyvers [4], the estimated θ_d for individual documents can be averaged by year to identify the trending topics across the sample period. This post-processing approach, however, is unable to take advantage of the potential time-dependent clusters naturally occurring in datasets.

Two types of time-dependence structures, upstream and downstream, can incorporate temporal features (see Fig. 2) [3]. The upstream structure allows the temporal features (e.g., timestamps) to influence the topic–mix distribution of a document, thereby determining the latent topics and tokens in a document. The downstream structure, in contrast, generates both tokens and timestamps conditioned on a latent topic. We first introduce TOT, a downstream model, followed by two upstream models, temporal collection (TC) and DTM.

The TOT model (see Panel (B) of Fig. 1) associates the document timestamp to every token in the document. It assumes that the topic mix of a document determines the latent topic at each position, which subsequently determines the observed tokens and the timestamp [11]. This model has a downstream structure because the topic mix (θ_d) determines the distribution of observed tokens (w_{di}) and timestamps (t_{di}) [3]. The additional timestamp variables in TOT allow the discovery of time-sensitive topics. One example is discovering topics over 21 decades of U.S. Presidential State-of-the-Union Addresses. The LDA model combines statements about the Mexican-American War (1846–1848) with those about World War I. The result is in contrast with topics discovered by TOT. TOT is able to localize statements about the Mexican-American War [11] and considers statements about World War I as belonging to a different topic because of the time gap between the two wars.

The temporal collection (TC) model [5] is based on similar ideas but instead adopts an upstream structure. The timestamp variable t enters the topic model under the assumption that the parameters of the prior distribution of topic mix, α, are a function of t. As a result, t influences the topic mix of document d, θ_d, the latent topics, and the observed tokens. TC adopts the gamma distribution to model time-dependent topic occurrence.



(A) Latent Dirichlet Allocation (LDA)

(B) The Topic over Time (TOT)

(C) The Dynamic Topic Model (DTM)

Fig. 1. Selected time-dependent topic models.

A somewhat different approach is adopted in DTM [12]. DTM assumes that both topic occurrence and topic-token distributions (i.e., topic meanings) can change over time. In this way, DTM is different from TOT and TC, which have fixed topic-token distributions but changing topic occurrences. One implication of changing topic-token distribution is that the meaning of a given topic can change substantially over a long time period. While this characteristic may be useful in some situations, it can lead to DTM accidentally piecing together unrelated topics across time. More studies are needed to determine whether an evolving topic-token distribution is a reasonable assumption.

2.2. Inference for time-dependent topic models

The insights the topic models provide come from the estimated latent variables and parameters given the observed documents. As an illustrative example, consider the LDA model discussed above. The key problem is to compute the joint posterior probability distribution of latent variables and parameters given w (the observed documents):

$$p(\theta, \phi, Z|w, \alpha, \beta) = \frac{p(\theta, \phi, z, w|\alpha, \beta)}{p(w|\alpha, \beta)}, \tag{2}$$

where $p(\theta, \phi, Z, w|\alpha, \beta)$ is defined in Eq. (1) and $p(w|\alpha, \beta)$ can be computed from $p(\theta, \phi, z, w|\alpha, \beta)$ by integrating out θ, ϕ, and Z. The TOT model has a similar structure but contains additional parameters from the beta distributions; these additional parameters generate the normalized timestamps. The TC model needs to include additional parameters and latent variables that control the time-dependent topic occurrence. The DTM model has a dynamic structure that further complicates the estimation problem. Each period t has vectors α_t and β_t · that need to be estimated.

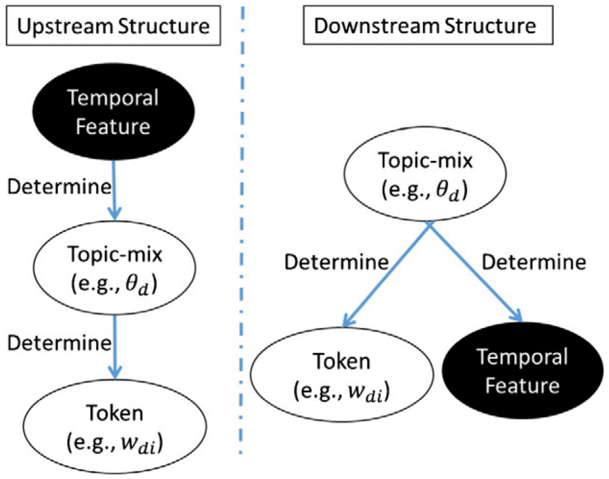


Fig. 2. Upstream and downstream time-dependent structures.

The inference methods for topic models fall roughly into two types: sampling-based algorithms and variational algorithms. A sampling-based algorithm constructs a Markov chain whose limiting distribution is the joint posterior of latent variables and parameters [15,16]. This type of algorithm can approximate the joint posterior to an arbitrary precision given unlimited computing resources. In practice, a fixed number of iterations will collect enough samples for subsequent inference tasks.

A variational algorithm [17] is a deterministic approach that searches for the best solution in a restricted family of probability distributions that is "simpler" compared to the joint posterior of latent variables and parameters. The nature of a variational algorithm is optimization. Noteworthy is that the solutions found by variational algorithms live in the restricted family of the probability distribution. The solution is, in general, not the mode of the joint posterior of latent variables and parameters. Whether the solutions found by variational algorithms are adequate is an empirical question.

One interesting question concerns the relative performance of the estimation approaches. Previous studies have shown that collapsed Gibbs sampling outperforms variational Bayes in terms of perplexity [4,18]. However, collapsed Gibbs sampling can require a longer running time to ensure convergence. The stochastic EM algorithm allows for the adjustment of parameters that are fixed in collapsed Gibbs sampling; this can have a positive impact on model performance [3].

2.3. Comparison of time-dependent topic models

To allow for a better understanding of the current status of time-dependent topic models, I discuss the following important characteristics of time-dependent topic models: time range, time-dependent structure, topic–token evolution, continuous or discrete time, time-dependent topic occurrence, and model estimation methods. Table A.1 in Appendix A provides a summary of time-dependent topic models based on these characteristics.

Time range characterizes the rough time interval from which the dynamic is considered. LDA relies on post-processing to capture the dynamic and, thus, does not have a target time range. DTM needs to first divide a dataset into discrete time intervals and aims at long-term topic dynamics that cover decades of documents. TOT also focuses on long-term topic clusters that can span months or years. TC targets medium-term changes across days or months. Short-term dynamics, ranging from hours to days, do not receive much attention in this research stream.

Noteworthy is that the discussion on time range is based on the model characteristics and the experiments conducted using these models. These time ranges are often the typical use cases intended in the original design. Adopting a model to different time ranges is possible and might be worth further investigation.

The time-dependent structure provides a channel for topic models to incorporate time-sensitive topic occurrences. As discussed in the previous subsection, this can be achieved through an upstream or downstream structure. The TOT model has a downstream structure while DTM and TC adopt an upstream structure. The upstream structure is consistent with the intuition that timestamps "influence" topics instead of the other way around. It also provides a clean analytical structure for model inference.

Previous studies adopted two types of granularity in the time-dependent structure: document-level and token-level. Document-level granularity assumes that every token in a document has the same temporal features while token-level granularity allows different feature values for tokens in a document. The dependence structure of the DTM and TC models is at document-level, while the timestamp in TOT is at token-level. Token-level dependence has the potential to provide more detailed information but can require additional computing resources.

Topic–token distributions determine the meanings of topics. Most time-dependent topic models, including TOT and TC, adopt fixed topic–token distributions. The meaning of a topic, as a result, is the same across time. The time-dependent topic occurrence determines the chance of encountering a topic at a given time. A topic might not be associated with a document if the topic is no longer active. Evolving topic–token distribution, on the other hand, allows the meaning of a topic to change over time. DTM adopts an evolving topic–token and topic occurrence setting.

A time-dependent topic model can adopt discrete or continuous time. A continuous time model allows the timestamp to enter the model directly without the need to be discretized. Both TOT and TC are continuous time models while DTM is a discrete time model. DTM, as a result, needs to determine the time interval for discretization before model inference. The continuous time dynamic topic model (cDTM) addresses this issue by relaxing the time-dependent structure of topic–token distributions [19]. The cDTM model allows for different time intervals between observations. Longer time intervals are associated with larger variances.

Time-dependent topic occurrences allow the occurrence rate of a topic to change over time. TOT adopts the beta distribution to model the normalized timestamp so that time-dependent clusters can be better captured. TC adopts the gamma distribution for a similar purpose. DTM model adopts the random walk process for changing topic occurrence. The cDTM model does not include this type of time-dependency structure but instead focuses on changing topic–token distributions.

Previous studies on time-dependent topic models have mostly focused on long-term (months-to-years) to median-term (days-to-months) dynamics; few studies have focused on short-term dynamics (hours-to-days). Both TOT and TC use single-modal (beta and gamma) distributions to model time-dependent topic occurrence. DTM adopts a random walk process for the same purpose. These approaches are unsuitable for modeling cyclical topic occurrence patterns. This study addresses the gap in the previous studies and proposes a novel topic model that can capture short-term cyclical topic dynamics in large document collections.

3. Probit-Dirichlet hybrid allocation (PDHA)

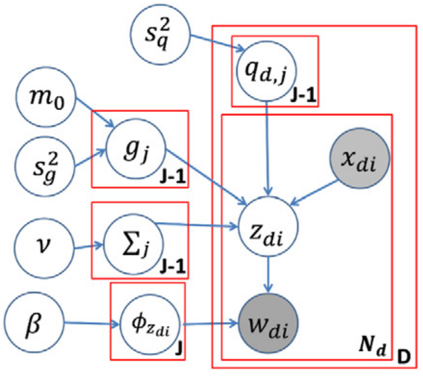
The basic idea of PDHA is to adopt a multinomial Probit-like structure so that temporal features can influence the latent topic of each token. PDHA has a token-level upstream structure, which is different from other time-dependent topic models. Table A.1 in Appendix A provides a summary of the differences between PDHA and existing time-dependent topic models.

This section presents the general structure of PDHA, followed by the temporal features incorporated for short-term cyclical patterns. We then discuss augmented Gibbs sampling for model inference. To streamline the discussion, it is assumed that all tokens in a document have the same temporal features. Extending this to the case of having different feature values for each token is straightforward.

3.1. The general structure of PDHA

The data-generating process (DGP) of PDHA is similar to that of the original LDA model. The main difference is how the temporal features affect topic distributions. In the following discussion, the J latent topics are indexed from 0 to $J-1$. The topic index starts from zero instead of one to facilitate the subsequent discussion of Probit-based topic generation defined by Eqs. (3) and (4). Fig. 3 plots the PDHA model in plate notation. The tokens in a document (w_{di}) and their temporal feature vectors (x_{di}) are observable in PDHA. Other latent variables, including the latent topic z_{di}, are unobservable and require estimation. PDHA assumes that a document is associated with a document-specific topic tendency vector $q_d=(q_{d,1}, q_{d,2}, \ldots, q_{dJ-1})$. Each element $q_{d,j}$ (j = 1, 2, …, J-1) is normally distributed with variance s_q^2 and mean zero. The other source of influence comes from the temporal feature vector x_{di} and its weight g_j (j = 1, 2, …, J-1). For a token at position i of document d, the latent topic z_{di} is generated by drawing from a multinomial distribution with parameters that are a function of q_d, x_{di}, g_j, and Σ_j. The token can then be generated by drawing from a multinomial distribution with the probability vector $\phi_{z_{di}}$.

The key idea of PDHA is that the sub-problem of generating z_{di} given temporal features resembles a classification problem (e.g., [20]). The main difference is that the outcome variable z_{di} in this classification problem is unobservable in the larger model defined by PDHA. As a result, it is difficult to extend a classification model directly into a topic model that incorporates token-level temporal features. However, because Gibbs sampling allows for approximating a larger model by drawing from the posteriors of smaller sub-problems, PDHA can be



Document-specific variables for cyclical patterns:

$$x_{di}^T=(1, x_{di,1}^T, x_{di,2}^T, x_{di,3}^T);$$

Daily: $x_{di,1}^T=(\cos\left(2\pi\frac{hd_d}{24}\right), \sin\left(2\pi\frac{hd_d}{24}\right))$

hd_d: time of day in hour, starting from midnight

Weekly: $x_{di,2}^T=(w1_d, w3_d, w4_d, w5_d, w6_d, w7_d)$

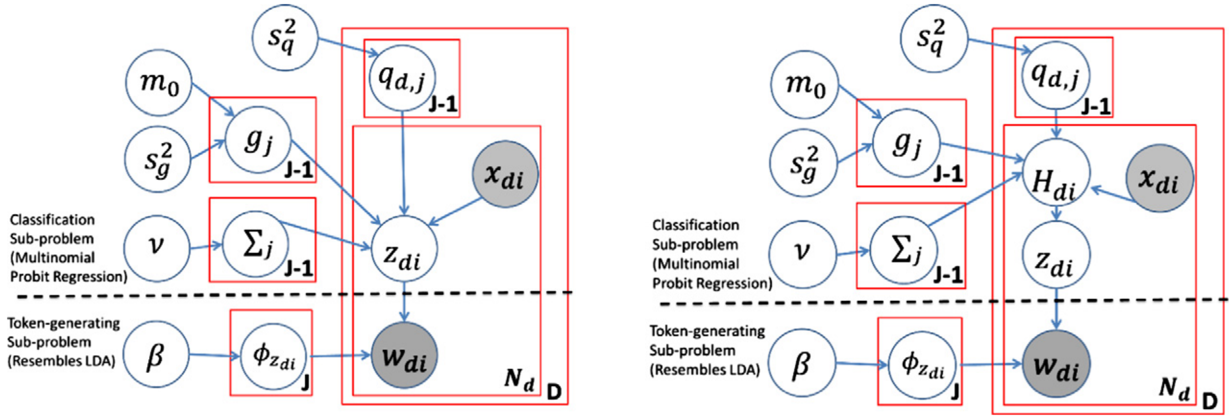
$\mathrm{wa_d}$: day-of-the-week dummy, a = 1,3,4,5,6,7.

Monthly: $x_{di,3}^T=\left(\cos\left(2\pi\frac{md_d}{ml_d}\right), \sin\left(2\pi\frac{md_d}{ml_d}\right)\right)$

md_d: day of the month

ml_d: number of day in the month

Fig. 3. Probit-Dirichlet hybrid allocation (PDHA).



(A) PDHA as a classification problem coupled with a token-generating problem

(B) PDHA augmented with additional latent variable for topic generation

Fig. 4. Two sub-problems of PDHA.

considered a classification sub-problem coupled with a token-generating structure if the latent topic z_{di} can be treated as observable in the classification sub-problem during model inference using Gibbs sampling. This line of reasoning leads to the adoption of multinomial Probit regression in the classification sub-problem of PDHA due to the development of Gibbs sampling-based inference for multinomial Probit regression (Imai and Dyk [21], and Albert and Chib [22]).

Panel (A) of Fig. 4 plots the classification sub-problem and token-generating sub-problem in PDHA. The upper-half is a multinomial Probit regression model while the lower-half resembles LDA. Following the specification of multinomial Probit regression, the probability of selecting a latent topic for z_{di} is defined by an additional latent vector $H_{di} \equiv (H_{di,1}, H_{di,2}, \ldots, H_{di,J-1})$:

$$H_{di,j} = q_{d,j} + x_{di}^T g_j + e_{di,j}; \cdot j = 1, 2, \ldots, J{-}1, \tag{3}$$

where $e_{di,j} \sim N(0, \Sigma_j)$ is white noise. The latent topic z_{di} is then determined by inspecting the relative value of H_{di} and assigning the topic by:

$$z_{di} = Y(H_{di}) = \begin{cases} 0, & if \quad \max(H_{di}) \le 0 \\ j, & if \quad \max(H_{di}) = H_{di,j} > 0, \end{cases} \tag{4}$$

The augmented vector H_{di} bridges the discrete topic assignment z_{di} and other document-specific variables. As defined by $Y(\cdot)$ in Eq. (4), the J-1 elements in H_{di} compete for topic assignment, and the one with the highest positive value "wins." The first topic (that of index 0) is assigned if all of the elements in H_{di} are negative. The vector g_j determines the effect of x_{di} on topic j. The first element in g_j (i.e., $g_{j,1}$) is the corpora-wide proportion of topic j. Other things being equal, a higher $g_{j,1}$ leads to a higher proportion of tokens associated with topic j in a document collection. The document-specific $q_{d,j}$ influences the proportion of topic j in tokens from document d.

Note that the token-level white noise $e_{di,j}$ introduces the variation of topics within a document given the document-level mean topic tendency defined by $q_{d,j} + x_{di}'g_j$ when all tokens in a document have the same x_{di}. The latent topics of tokens in the same document can differ because each token is associated with a different white noise. Following the specifications of Imai and Dyk [21] for multivariate probit regression, the first white noise has unit variance ($\mathrm{Var}(e_{di,1}) = 1$), which makes the model identifiable.

3.2. Temporal features for short-term cyclical dynamics

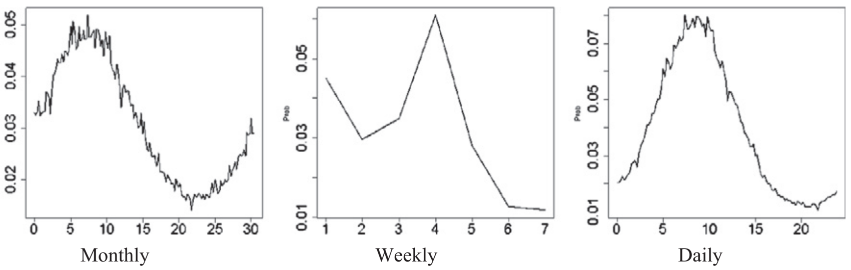
There are two common approaches to model cyclical patterns. The first approach is to use dummy variables (i.e., indication variables). For example, 6 dummy variables can describe different topic occurrence rates across the 7 days in a week. The other approach is to adopt the Serfling model [23]. The basic idea of the Serfling model is to use appropriate sine and cosine functions to capture the cyclical patterns of selected frequencies. Consider a document d with time of day hd_d (in hours; starting from midnight), $0 \le hd_d < 24$. Then $a_1 \cos\left(\frac{2\pi\, hd_d}{24}\right) + b_1 \sin\left(\frac{2\pi\, hd_d}{24}\right)$ can be used to capture the daily cyclical pattern. The coefficients a_1 and b_1 are estimated from training data. Similarly, monthly cyclical patterns can be captured by $a_2 \cos\left(2\pi \frac{md_d}{ml_d}\right) + b_2 \sin\left(2\pi \frac{md_d}{ml_d}\right)$, where md_d is the day of the month of document d, and ml_d is the total number of days in the corresponding month.

For a document d that contains timestamps up to minutes, the temporal features for daily, weekly, and monthly cyclical patterns are

$$x_{di}^T = \left(1, \cos\left(2\pi \frac{hd_d}{24}\right), \sin\left(2\pi \frac{hd_d}{24}\right), w1_d, w3_d, w4_d, w5_d, w6_d, w7_d, \cos\left(2\pi \frac{md_d}{ml_d}\right), \sin\left(2\pi \frac{md_d}{ml_d}\right)\right). \tag{5}$$

Table 1
Research testbeds.

Dataset	# of doc.	# of tokens	# of unique tokens	Temporal features	Remarks
WMT	24,995	825,653	37,315	Day of the week, day of the month, and time of day	Postings on the Yahoo Finance Wal-Mart message board from 1/1/2007 midnight to 5/1/2007 midnight (Eastern Time)
NYT	4224	1,056,717	59,830	Day of the week and day of the month	10% random sample of NYT articles from 1/1/2008 to 6/30/2008
RTS	11,771	775,553	26,898	Day of the week, day of the month, and time of day	Reuters-21578



(A)

Regression Parameters:

Variable	Est. Value	t-value[1]
const.	-1.434***	-2260.4
w7	-0.106***	-13.4
w1	0.064***	7.1
w3	0.038***	6.1
w4	0.143***	20.6
w5	-0.007	-0.9
w6	-0.109***	-17.1
Mcos†	0.011***	2.6
Msin†	0.196***	35.4
Hcos‡	-0.209***	-24.3
Hsin‡	0.226***	36.8

†For monthly cyclical patterns.
‡For daily cyclical patterns.
[1] Computed using time-series corrected standard error.
***, **, and * indicate significant at the 99%, 95%, and 90% confidence levels.

(B)

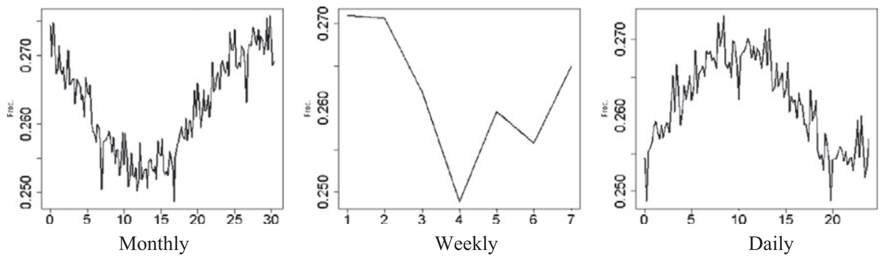
Top Keyword Probability:

Word	Prob.	2.5% Percentile	97.5% Percentile
sales	0.0365	0.0348	0.0383
year	0.0312	0.0298	0.0329
growth	0.0211	0.0204	0.0220
percent	0.0193	0.0177	0.0206
stores	0.0176	0.0156	0.0190
billion	0.0161	0.0148	0.0180
years	0.0130	0.0114	0.0146
store	0.0129	0.0111	0.0142
stock	0.0117	0.0105	0.0131
company	0.0092	0.0074	0.0109
earnings	0.0082	0.0072	0.0089
target	0.0081	0.0067	0.0095

Number of significant tokens in this topic: 284

(C)

Fig. 5. Topic “earnings” from WMT.



(A)

Regression Parameters:

Variable	Est. Value	t-value[1]
const.	-1.438***	-903.3
w7	-0.011*	-1.9
w1	0.001	0.2
w3	-0.004	-0.8
w4	-0.045***	-7.3
w5	-0.052***	-7.0
w6	-0.048***	-8.1
Mcos†	0.052***	12.4
Msin†	-0.013***	-3.4
Hcos‡	-0.029***	-5.1
Hsin‡	0.049***	12.1

†For monthly cyclical patterns.
‡For daily cyclical patterns.
[1] Computed using time-series corrected standard error.
***, **, and * indicate significant at the 99%, 95%, and 90% confidence levels.

(B)

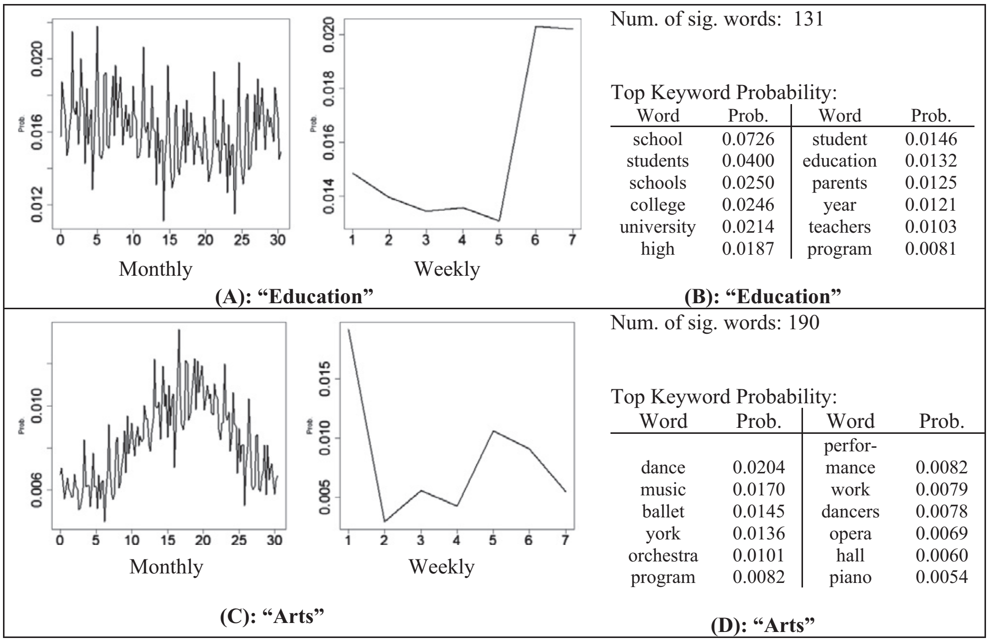
Top Keyword Probability:

Word	Prob.	2.5% Percentile	97.5% Percentile
people	0.0200	0.0180	0.0219
time	0.0130	0.0104	0.0145
business	0.0113	0.0092	0.0133
make	0.0105	0.0098	0.0115
work	0.0092	0.0053	0.0147
pay	0.0081	0.0049	0.0119
good	0.0076	0.0055	0.0090
back	0.0066	0.0039	0.0081
company	0.0066	0.0030	0.0119
job	0.0063	0.0041	0.0080
long	0.0056	0.0048	0.0064
working	0.0050	0.0046	0.0056

Number of significant tokens in this topic: 238

(C)

Fig. 6. Topic “employee relationship” from WMT.



Num. of sig. words: 131

Top Keyword Probability:

Word	Prob.	Word	Prob.
school	0.0726	student	0.0146
students	0.0400	education	0.0132
schools	0.0250	parents	0.0125
college	0.0246	year	0.0121
university	0.0214	teachers	0.0103
high	0.0187	program	0.0081

(B): "Education"

Num. of sig. words: 190

Top Keyword Probability:

Word	Prob.	Word	Prob.
dance	0.0204	perfor-mance	0.0082
music	0.0170	work	0.0079
ballet	0.0145	dancers	0.0078
york	0.0136	opera	0.0069
orchestra	0.0101	hall	0.0060
program	0.0082	piano	0.0054

(D): "Arts"

Fig. 7. Topics from NYT.

If the document's publication time is specific only up to the date, then the daily variables are removed:

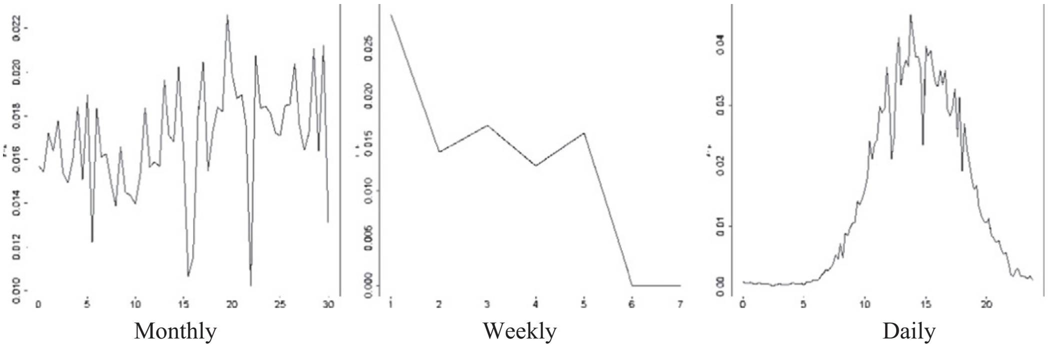
$$x_{di}^{T}=\left(1,w1_d,w3_d,w4_d,w5_d,w6_d,w7_d,\ \cos\left(2\pi\frac{md_d}{ml_d}\right),\ \sin\left(2\pi\frac{md_d}{ml_d}\right)\right). \quad (6)$$

Cyclical patterns of frequencies higher than days or lower than months can also be included using the Serfling model representation. For example, the cyclical patterns for a half-day (12 h) can be captured by adding $\left(\cos\left(2\pi\frac{hd_d}{12}\right),\ \sin\left(2\pi\frac{hd_d}{12}\right)\right)$ to the feature vector. Second, it is possible to automatically select the best combination of frequencies using standard model selection techniques, such as model evidence [14]. This study considers the case of cyclical patterns with fixed frequencies.

3.3. Model inference using augmented Gibbs sampling

We adopted Gibbs sampling for model inference because it facilitates the extension and combination of inference algorithms. Specifically, we extend augmented Gibbs sampling for multinomial Probit regression developed by Imai and Dyk [21] and Albert and Chib [22] for use in PDHA inference. Gibbs sampling constructs a Markov chain that converges to the posterior of latent variables and coefficients [16,24]. Both our inference algorithm and the one for multinomial Probit regression adopt additional augmented variables (e.g., H_{di} in PDHA) to facilitate the sampling process. We refer to these algorithms as augmented Gibbs sampling algorithms to indicate the use of additional augmented variables.

All variables with open circles in Panel (A) of Fig. 4 are parameters or latent variables and require estimation. Our approach follows the collapsed Gibbs sampling for LDA developed by Griffiths and Steyvers

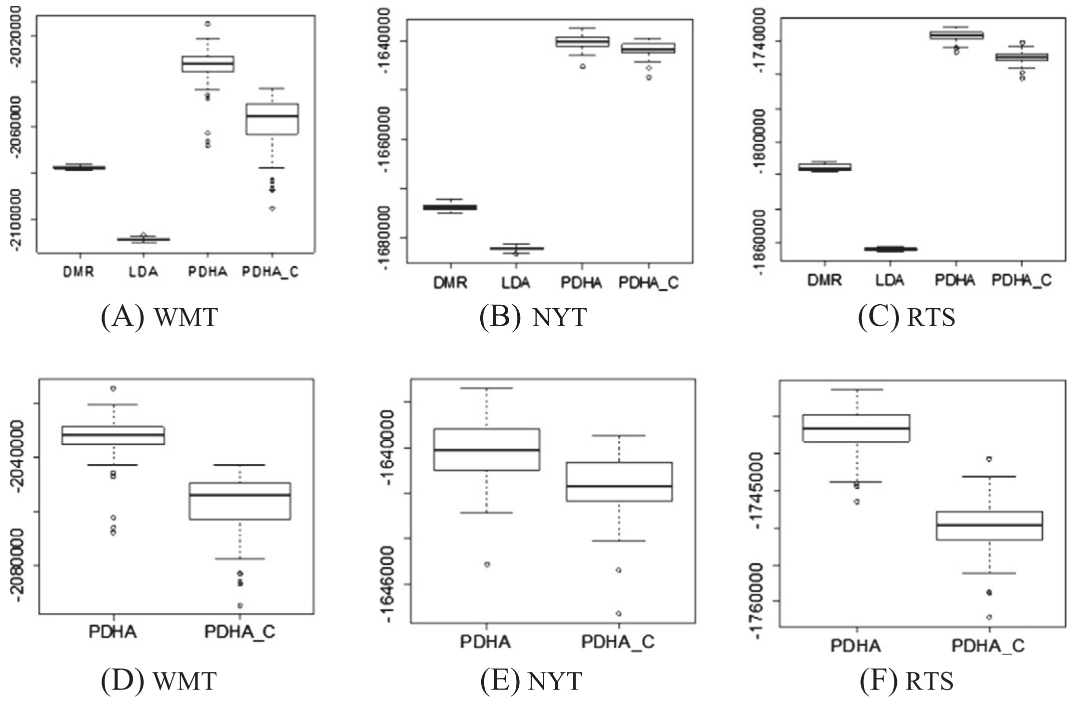


Number of significant tokens in this topic: 134

Top Keyword Probability:

Word	Prob.	Word	Prob.	Word	Prob.
company	0.0330	sale	0.0165	letter	0.0137
agreement	0.0327	mln	0.0161	acquisition	0.0133
dlrs	0.0206	corp	0.0140	subject	0.0133
general	0.0167	signed	0.0139	transaction	0.0128

Fig. 8. Topic "merger and acquisition" from RTS.



(A) WMT (B) NYT (C) RTS

(D) WMT (E) NYT (F) RTS

Fig. 9. Likelihood of WMT, NYT, and RTS. The upper panels are boxplots of the likelihood of the DMR, LDA, PDHA, and PDHA without temporal features (PDHA_C). The lower panels show the comparison of only PDHA and PDHA_C.

[4] and integrates out $\phi_{z_{di}}$. The remaining parameters and latent variables are divided into two groups. The first group contains the latent topic $Z = \{z_{di}\}$, and the second group contains the parameters and latent variables related to the multinomial Probit regression sub-problem, including slopes $G = \{g_j\}$, document-specific topic tendency $Q = \{q_{d,j}\}$, and variance Σ_j. The tokens in all documents $w = \{w_{di}\}$ and their temporal features $X = \left(x_{11}^T, x_{12}^T, \ldots, x_{1N_1}^T, x_{21}^T, x_{22}^T, \ldots, x_{2N_2}^T, \ldots, x_{D1}^T, x_{D2}^T, \ldots, x_{DN_D}^T\right)^T$ are observable variables in PDHA. Augmented Gibbs sampling approximates the joint posterior $p(Z, Q, G, \Sigma | w, X, \cdot)$ by iteratively updating the latent topic Z and regression coefficients via the steps:

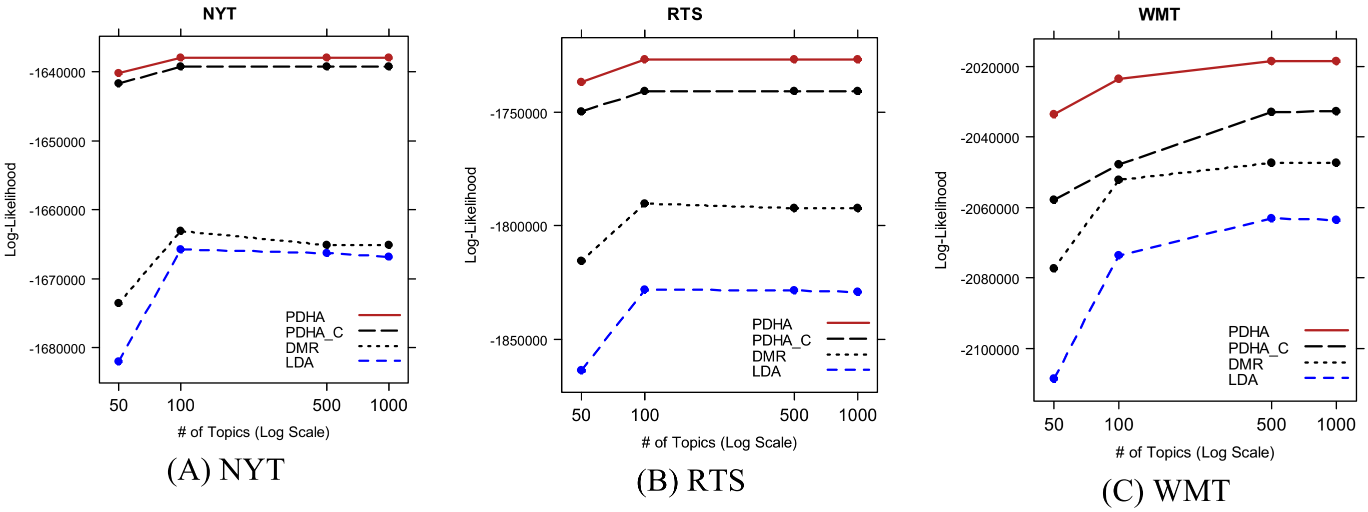
1. Sample latent topic Z from $p(Z | w, X, Q, G, \Sigma, \cdot)$.
2. Sample regression coefficients Q, G, Σ from $p(Q, G, \Sigma | w, X, Z, \cdot)$.

Note that Step 2 can be achieved by the augmented Gibbs sampling algorithms for multinomial Probit regression because, given the latent topic Z, the conditional posterior $p(Q, G, \Sigma | w, X, Z, \cdot)$ has the same structure as that of the multinomial Probit regression. The subsequent discussion focuses on developing the sampling procedure in Steps 1 and 2.

3.3.1. Sampling latent topic Z

Similar to the collapsed Gibbs sampling scheme of the LDA model [4], our method updates z_{di} in a sequential manner. The posterior z_{di} conditional on other variables is

$$\begin{aligned} &p(z_{di} = j | z_{-di}, w_{di}, w_{-di}, X, Q, G, \Sigma) \\ &\quad \propto p(w_{di} | z_{di} = j, z_{-di}, w_{-di}) p(z_{di} = j | Q, G, \Sigma, \mathrm{X}) \\ &\quad = \frac{n_{-di,j}^{(w_{di})} + \beta}{n_{-di,j}^{(\cdot)} + W\beta} p(z_{di} = j | q_d, G, \Sigma, \mathrm{x_d}), \end{aligned} \tag{7}$$



(A) NYT (B) RTS (C) WMT

Fig. 10. Likelihoods of LDA, DMR, PDHA, and PDHA_C with 50, 100, 500, and 1000 latent topics. Panels (A), (B), and (C) plot the results using NYT, RTS, and WMT datasets.

where $n_{-di,j}^{(\cdot)}$ is the number of assignments to topic j, excluding the assignment at position i of document d; $n_{-di,j}^{(w_{di})}$ is the instance token-type w_{di} assigned to topic j, excluding the instance at position i in document d; W is the number of unique tokens in the corpus. The token position index is suppressed for temporal features x_d because all tokens in a document have the same value.

The first term in Eq. (7) is derived by integrating out ϕ_j and can be readily computed based on token–topic assignments. The second term, however, involves intractable integrals. We propose using a simulation method to evaluate this term.

Let $\hat{\theta}_{d|x_d,j}$ denote the estimated $p(z_{di} = j|q_d, G, \Sigma, \mathrm{x_d})$, the probability of assigning a token to topic j given x_d and regression parameters. The idea is to compute $\hat{\theta}_{d|x_d,j}$ by generating $e_{di,j} \sim N(0, \Sigma_j)$ and computing $H_{di,j}$ using Eq. (3). The topic assignment then can be determined using Eq. (4). The process is repeated G times to compute $\hat{\theta}_{d|x_d,j}$.

Direct implementation of this approach introduces independent simulation error into $\hat{\theta}_{d|x_d,j}$, which can be undesirable in this setting. Consider the case of two documents, d_1 and d_2, having exactly the same temporal features and document-specific topic tendency ($q_{d_1} = q_{d_2}$). Then $p(z_{d_1 i} = j|q_{d_1}, G, \Sigma, \mathrm{x_{d_1}})$ and $p(z_{d_2 i} = j|q_{d_2}, G, \Sigma, \mathrm{x_{d_2}})$ should be the same for each topic j. However, because of the simulation error, the numerical procedure usually results in $\hat{\theta}_{d_1|x_{d_1},j} \neq \hat{\theta}_{d_2|x_{d_2},j}$. To address this problem, we adopt the common random variable method [25] and cache $C(J-1)$ draws of a random variable from the standard normal distribution. The probability estimation $\hat{\theta}_{d|x_d,j}$ for each document is computed based on the same set of random variables. This approach provides the internal consistency for the estimated probability and ensures that $\hat{\theta}_{d_1|x_{d_1},j} = \hat{\theta}_{d_2|x_{d_2},j}$s in the above example. The procedure is summarized in Algorithm 1.

Algorithm 1. Estimate $p(z_{di} = j|q_d, G, \Sigma, \mathrm{x_d})$ via simulation

```
Inputs: C(J-1) random variables drawn from N(0,1) ($\epsilon_{c,j} \sim N(0,1)$, c= $1,2,\ldots \mathrm{C}, j = 1,2,\ldots,J-1$);
temporal feature $x_d$; document-specific topic tendency $q_d$; regression coefficients $g, \Sigma$.
Outputs: Numerically estimated $p(z_{di} = j|q_d, g, \Sigma, \mathrm{x_d})$ (i.e., $\hat{\theta}_{d|x_d,j}$) for each document.

Set $\hat{\theta}_{d|x_d,j}$=0 for $d = 1,2,\ldots,D$ and $j = 0,2,\ldots,J-1$.
for each $d = 1,2,\ldots,D$ do
    for each $c = 1,2,\ldots,\mathrm{C}$ do
        for each $j = 1,2,\ldots,J-1$ do
            $H_{d,j}^{(\mathrm{c})} = q_{d,j} + x_d^T g_j + \sqrt{\Sigma_{jj}}\epsilon_{g,j}$;
        end

        $z = \begin{cases} 0, \text{ if } \max\left(H_d^{(\mathrm{c})}\right) \le 0 \\ j, \text{ if } \max\left(H_d^{(\mathrm{c})}\right) = H_{d,j}^{(\mathrm{c})} > 0; \end{cases}$
        $\hat{\theta}_{d|x_d,z} = \hat{\theta}_{d|x_d,z} + 1$;
    end

    for each $j = 0,2,\ldots,J-1$ do
        $\hat{\theta}_{d|x_d,j} = \hat{\theta}_{d|x_d,j}/\mathrm{C}$;
    end
end
```

The parameter C should be large enough to enable accurate probability estimation. Preliminary experiments suggest that for a model with a moderate number of topics (e.g., 50), $C \approx 1500$ is a reasonable choice. The time complexity of Algorithm 1 is O(DJC). Note that few changes are needed even if each token has different temporal features. The main change would be to loop over individual tokens, instead of documents and compute a simulated $H_{di,j}$ for each token.

3.3.2. Sampling regression coefficients

We applied the marginal data augmentation approach for multinomial Probit regression [21,26] to draw regression coefficients from $p(Q, G, \Sigma|w, X, Z, \cdot)$. We present the outline of the algorithm and the analysis of time complexity.

The basic idea is to include additional augmented variables to facilitate model inferences. Panel (B) of Fig. 4 plots the model with augmented variable H_{di}. We adopted another augmented variable a ($a > 0$; not visible in Panel (B) of Fig. 4) to address the technical difficulties caused by the identification constraint $\Sigma_1 = 1$ (the first diagonal element in Σ). The variable a scales the original model to an equivalent one with $\Sigma_1 = a^2$. The regression coefficients are updated through the transformed model. The updated variables are then scaled back to the original model. Fig. B.1 in Appendix B summarizes the major steps for drawing the regression coefficients.

Running a single sweep for PDHA inference includes computing $\hat{\theta}_{d|x_d,j}$ via Algorithm 1, updating Z, and other parameters (see Fig. B.1 in Appendix B for details). The overall time complexity is $O\left(D\overline{N}_d K^2 J + JK^\gamma + J^\gamma\right)$, where $\overline{N}_d$ is the average document length and the exponent γ is associated with the computational cost of matrix inversion. Inverting a matrix via Gauss–Jordan elimination has a time complexity of $O(K^3)$ for a K-by-K square matrix. However, matrix inversion can run at the same time complexity as matrix multiplication using the block-wise inversion method [27], and matrix multiplication using the Coppersmith–Winograd algorithm [28] has a time complexity of $O(K^{2.376})$. As a result, $\gamma = 2.376$ would be a reasonable choice for the current analysis.

The first term of the overall time complexity is contributed by the inner product of the temporal feature matrix X when computing the posterior mean of $\tilde{g}_j^*$. The second term is contributed by inverting the precision matrix of $\tilde{g}_j^*$. The last term comes from the inversion of Σ (in Step 1 of Fig. B.1 in Appendix B). Note that the computational costs of updating Z ($D\overline{N}_d J$) and Algorithm 1 (DJC) are dominated by the cost of sampling regression coefficients.

The PDHA inference method grows linearly with respect to the size of the corpus ($D\overline{N}_d$). This means that our approach is scalable if the number of topics and the lengths of the temporal features are fixed. When the length of temporal features increases, the running time will eventually increase at the speed of K^γ. Note that the first term ($D\overline{N}_d K^2 J$) typically has a larger constant compared to the second term (JK^γ). As a result, a growth rate close to K^2 will be observed for smaller K (e.g., $K < 100$). A similar effect applies to the growth rate of the number of topics. The asymptotic growth rate is J^γ but, in practice, a near-linear growth rate is observed for smaller J.

The main focus of the current study is to evaluate PDHA model in an archival setting. Applying a learned PDHA model in a streaming environment is possible. To do so, the latent topics of unseen documents need to be estimated based on a learned model. The time complexity for this is $O\left(D'\overline{N}'_d J\right)$, where D' is the number of unseen documents and $\overline{N}'_d$ is the average token length of these documents.

3.4. Analysis of sampling results

The augmented Gibbs sampling approach can be run for L sweeps to collect the sampling results. The first B burn-in sweeps are discarded to minimize the impact of initial values. After the burn-in sweeps, every L_T sweep is recorded for subsequent analysis. The practice of storing only every L_T sweep is called "thinning." Thinning is used to reduce the autocorrelation between the recorded sweeps introduced by Gibbs sampling. We refer readers to Gelfand [29] for a more detailed introduction to thinning and other related concepts.

To simplify notation, we re-index the collected sweeps from 1 to L_S. For example, for $L = 2000$, $B = 1000$, and $L_T = 20$, the augmented Gibbs sampling runs for 2000 sweeps and the first 1000 sweeps are discarded. The latent variables and parameters from 1020, 1040, …, and 2000 sweeps are recorded and re-indexed from 1 to 50.

The latent topics from sweep r, $Z^{(r)}$, can help with the estimation of the probability of observing token w, conditional on topic j:

$$\phi_{j,w}^{(r)} = \frac{n_{j,w}^{(r)} + \beta}{n_{j,\cdot}^{(r)} + W\beta}.$$

where $n_{j,\cdot}^{(r)}$ is the count of topic j in $Z^{(r)}$, and $n_{j,w}^{(r)}$ is the count for which token w is associated with topic j. Given the recorded L_s sweep, the posterior mean of $\phi_{j,w}$ is $\hat{\phi}_{j,w} = \frac{1}{L_s}\sum_{r=1}^{L_s}\phi_{j,w}^{(r)}$. The 95% confidence interval of $\phi_{j,w}$ is the interval between the 2.5 and the 97.5 percentiles of $\{\phi_{j,w}^{(r)}\}$, $r = 1, 2, \ldots, L_s$. The confidence interval of $\phi_{j,w}$ provides a convenient way to identify significant keywords in a topic. The conjugate prior to $\phi_{j,w}$ has a mean of 1/W. A $\phi_{j,w}$ with a confidence interval larger than 1/W suggests that token w is indeed associated with topic j and is therefore referred to as a significant keyword for the underlying topic. The number of significant keywords in a topic provides a useful reference for the vocabulary size of the topic.

The posterior distribution of G and Σ can be analyzed using the recorded sweeps. In addition to computing posterior mean and confidence interval, we compute the p-value of each element in G using the t-value computed based on a time-series standard error. The time-series standard error considers the potential autocorrelation in the recorded sweeps and is usually more conservative than the standard error computed ignoring the effects of autocorrelation.

Finally, the cyclical patterns of a topic can be computed using Algorithm 1. For example, to compute the dynamics of daily cyclical patterns of topic j, the input temporal features are $x^T = \left(1, \cos\left(2\pi\frac{h}{24}\right), \sin\left(2\pi\frac{h}{24}\right), \overline{w1}, \overline{w3}, \overline{w4}, \overline{w5}, \overline{w6}, \overline{w7}, \overline{cos\left(2\pi\frac{md}{ml}\right)}, \overline{\sin\left(2\pi\frac{md}{ml}\right)}\right)$, where $\overline{w1}, \overline{w3}, \overline{w4}, \overline{w5}, \overline{w6}, \overline{w7}, \overline{\cos\left(2\pi\frac{md}{ml}\right)}$, and $\overline{\sin\left(2\pi\frac{md}{ml}\right)}$ are the sample means in X and h is the selected time of day in hours. The mean regression coefficients, $\hat{g}_j$ and $\hat{\Sigma}_{jj}$ are also needed. Note that we are interested in the cyclical patterns of a topic instead of those of a document. Thus, we use the average document-specific topic tendency for each topic. Using the common random variable approach outlined in Algorithm 1, we can compute the conditional probability of a topic at different h.

3.5. Baseline model

This study adopts two baseline models for comparison with the proposed PDHA model. The first baseline is the LDA that did not consider the temporal features included in PDHA. The LDA represents the performance of a widely-used topic model.

The second baseline is the DMR model that includes exactly the same set of temporal features as in the proposed PDHA model. DMR extends the LDA and allows temporal features to influence the prior distribution of the topic distribution of a document. DMR is similar to the TC model but allows the time-dependent features to capture cyclical topic dynamics. We excluded TOT as a baseline in this study mainly because the TOT model normalizes the timestamps to a value between zero and one as a way of accommodating the assumption that timestamps are generated from the beta distribution. The shape of beta distribution is controlled by two parameters and is not capable of tracking cyclical patterns.

3.6. Model evaluation

A common approach to evaluate a topic model is computing the log likelihood of a testing set given the learned coefficients. Let $G = \{g^{(1)}, g^{(2)}, \ldots, g^{(L_s)}\}$, $q = \{q^{(1)}, q^{(2)}, \ldots, q^{(L_s)}\}$, and $\Omega = \{\Sigma^{(1)}, \Sigma^{(2)}, \ldots, \Sigma^{(L_s)}\}$ denote the parameters and latent variables recorded from L_S sweeps. For a testing document d with temporal feature x_d and a collection of tokens w_d, the testing likelihood $p(w_d|x_d, G, Q, \Omega)$ is computed using the importance sampling approach [30].

This approach mainly involves three steps. First, sample $\theta_{d|x_d}^{(r)}$ is created based on the record sweep r, $r = 1, 2, \ldots, L_s$. This step follows the data-generating process of PDHA with a few exceptions. The document-specific topic tendencies q_j are no longer drawn from the prior distribution $N(0, s_q^2)$ but from $N\left(\hat{m}_{q,j}, \hat{s}_{q,j}^2\right)$, where $\hat{m}_{q,j}$ and $\hat{s}_{q,j}^2$ are the mean and variance of the vector $(q_{1,j}^{(r)}, q_{1,j}^{(r)}, \ldots, q_{1,j}^{(r)}, q_{2,j}^{(r)}, q_{2,j}^{(r)}, \ldots, q_{2,j}^{(r)}, \ldots, q_{D,j}^{(r)}, q_{D,j}^{(r)}, \ldots, q_{D,j}^{(r)})$, with each $q_{d,j}^{(r)}$ repeating N_d times. Combined with $g_j^{(r)}$ and $\Sigma^{(r)}$, $\theta_{d|x_d}^{(r)}$ can be readily computed using Algorithm 1.

The second step is to compute the probability of a token w in document d, $\hat{w}_{di}$. Because $\theta_{d|x_d}^{(r)}$ is vector of topic probability in document d, $\overline{w}_{di} = \frac{1}{L_s}\sum_{r=1}^{L_s}\sum_{j=1}^{J}\phi_{j,w_{di}}^{(r)}\theta_{d|x_d,j}^{(r)}$. The first summation is over all possible topics that can contribute to the occurrence of the token w_{di}; the second summary is over all recorded sweeps. Finally, $\log p(w_d|x_d, G, Q, \Omega)$ sums up all $\log\overline{w}_{di}$ because each token is independently distributed given q_d, g, Σ, and x_d:

$$\log p(w_d|x_d, G, Q, \Omega) = \sum_{i=1}^{N_d} \log\hat{w}_{di}.$$

A similar process can be used to compute the testing likelihood of the LDA, DMR, and DTM models.

4. Experimental results

We evaluated PDHA on three datasets: the Yahoo Finance Wal-Mart message board (WMT), the New York Times (NYT), and Reuters-21578 (RTS). The three datasets cover different types of text content. WMT is user generated content; NYT and RTS are

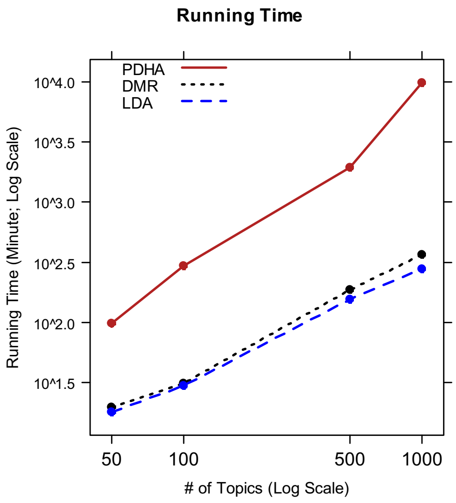


Fig. 11. Running time comparison (using the WMT dataset).

news and newswire articles, respectively. The variety among the three datasets allows the study to better characterize the performance of the proposed PDHA. Table 1 presents a summary of the three datasets. We included different temporal features based on availability. The WMT dataset is composed of postings on Yahoo Finance's Wal-Mart message board. Because of the presence of timestamps (Eastern Time; precise to the second) for each posting, we included variables for daily, weekly, and monthly cyclical patterns, as in Eq. (5). The NYT dataset is a 10% random sample of NYT articles published between 1/1/2008 and 6/30/2008. The model for NYT only contains variables for weekly and monthly cyclical patterns as in Eq. (6). The RTS dataset includes timestamps precise to the second. Thus, we included variables for daily, weekly, and monthly cyclical patterns, as in Eq. (5). We normalized regressors to have a zero mean and unit variance.

We estimated all models with Gibbs sampling of 2000 sweeps. To minimize the impact of initial values, we discarded the first 1000 sweeps. We recorded the sampled parameters every 20 sweeps to reduce autocorrelation and computed estimation results based on the recorded sweeps. We set the first element in m_0, the prior mean of intercept $g_{j,1}$, to give all topics an equal prior probability if all other coefficients in Eq. (3) were zero. The prior variances s_g^2 and s_q^2 were $2000/\bar{l}$ and $100/\bar{l}$, where $\bar{l}$ was the average token length of documents in a corpus. We present the estimation results of PDHA with 50 latent topics in Section 4.1. In Section 4.2, we present the likelihood comparisons and running time comparisons of four different topic amounts (50, 100, 500, and 1000).

4.1. Topic cyclical patterns in WMT, NYT, and RTS

This section presents selected topic cyclical patterns estimated from WMT, NYT, and RTS. Fig. 5 summarizes the "Earnings" topic from WMT. The topic name was assigned manually based on the top keywords. Panel (A) plots the monthly, weekly, and daily cyclical patterns computed based on the recorded Gibbs sampling sweeps. This topic is clearly more popular during the first half of a month. Thursday has the highest probability for this topic, while weekends have a much lower probability. The time-of-day pattern shows that it reaches a peak around the same time as the opening of the stock market: 9:30 am. The occurrence rate drops quickly after noon and reaches its trough around midnight. The probability differences between the peak and trough are quite large. The probability density during peak hours is 0.08 while the probability drops to 0.01 during the trough.

Panel (B) lists the estimated regression coefficients used to compute probability in Panel (A). The t-value listed in the last column is computed using a time-series corrected standard error to account for the autocorrelation between recorded sweeps. Coefficients for monthly and daily cyclical patterns are significant, consistent with the patterns in Panel (A). Most day-of-the-week coefficients are also significant.

Panel (C) lists the top keywords in this topic. All top keywords have 95% confidence intervals far above the expected prior probability of a keyword ($0.000026 \approx 1/37315$), suggesting strong evidence for them. The top keywords such as "sales," "year," "growth," "percent," "stores," and "billion" typically appear in postings discussing revenues, sales, store expansion, etc. This type of new information is directly related to stock price. Thus, the observance of a close relationship between the "Earnings" topic and stock trading hours is reasonable.

Fig. 6 plots the "Employee Relationship" topic from WMT. Like the "Earnings" topic from the same dataset, this topic shows distinct monthly, weekly, and daily cyclical patterns. The monthly pattern in Panel (A) shows that this topic is more popular during the beginning and ending of a month. Thursday has the lowest probability of all the days in a week. This topic is denser from around 7 am-to-3 pm. The probability difference between peak and trough hours, however, is smaller compared to that of the "Earnings" topic. One possible reason for this is that though trading activities drive discussions of earnings, no similar driving force exists for conversations on employee relationships. The phenomenon of a lower probability on Thursdays might be related to the payday schedule of Wal-Mart. Panel (C) lists the important keywords in this topic. Top keywords such as "people," "time," "business," "make," "work," and "pay" in this topic often appear in postings about schedules, hourly wage, and the lifestyles of Wal-Mart employees.

Fig. 7 summarizes two selected topics from NYT. The estimated regression coefficients and confidence intervals for top keywords are suppressed to save space. Panel (A) plots the monthly and weekly patterns of the topic "Education." This topic is without a clear monthly peak or trough. Weekly dynamics, on the other hand, clearly show that weekends have a much higher probability compared to weekdays. The weekly pattern might reflect the choices of NYT's editors. The top keywords in Panel (B), including "school," "students," "schools," "college," and "university," typically appear in articles about educational issues, such as teacher pay, charter schools, principal hiring, and safety issues in schools. The "Arts" topic presented in Panel (C) has a clear monthly pattern that peaks around the 20th day of a month. Because the topic often appears in articles that comment on and introduce musical and theatrical events, which are often on weekends, the topic's higher probabilities on Monday and Friday are not surprising. The top keywords in Panel (D), including "dance," "music," "ballet," "york," and "orchestra," are consistent with the common knowledge of news articles about this topic.

Fig. 8 summarizes the "Merger and Acquisition" (M&A) topic from RTS. Clear day-of-the-week and time-of-day patterns are present. Articles about M&A often appear on Monday and seldom appear during the weekends. The daily pattern shows a peak around 2 pm. The day-of-the-month plot does not reveal a clear cyclical pattern. The top keywords in this topic, including "company," "agreement," "dlrs," and "sale," typically appear in articles that announce corporate mergers and unit sale events.

The selected estimation results presented in this section clearly show that PDHA is able to capture interesting monthly, weekly, and daily cyclical patterns in document collections. The additional information required for the proposed model to work is the timestamp for each document, which is usually available from the text from the mass media and user-generated content. We report the performance of the proposed approach and baselines in the following section.

4.2. Likelihood comparisons

This section reports the log-likelihood comparison of PDHA with other baseline models. We consider models with higher testing log-likelihoods to excel at generalizing the estimated models to unseen data. We designed the experiments to answer two research questions: (1) Does PDHA perform better compared to other baseline topic modeling approaches? (2) How do the temporal features included in PDHA contribute to the performance difference? To answer these two research questions, we compared PDHA to LDA, DMR, and a simplified PDHA model, PDHA_C, that excluded temporal features from PDHA. The log-likelihood difference between LDA and PDHA can be interpreted as the joint effect of the model difference and temporal features. Since both DMR and PDHA have access to exactly the same set of temporal features, the log-likelihood difference stems from model difference only. The underlying models for PDHA and PDHA_C are the same. The log-likelihood difference, as a result, stems from temporal features.

For each testbed, we allocated 70% of the documents for training and reserved the rest for testing. We set the smoothing parameters of all models to $\alpha = \frac{50}{T}, \beta = 0.01$, following a previous study [6]. We ran Gibbs sampling for 2000 sweeps, and used the importance sampling approach to compute the log-likelihood from the last 20 recorded sweeps (with thinning $L_T = 20$). For each model, we repeated the estimation 40 times using different seeds.

Fig. 9 presents boxplots of the log-likelihoods for the setting using 50 latent topics. Panels (A) to (C) plot the log-likelihood values for the DMR, LDA, PDHA, and PDHA_C models estimated using WMT, NYT, and RTS. PDHA models, on average, achieved the highest log-likelihood values of the three testbeds. The LDA models, on the other hand, resulted in the lowest log-likelihood values. Note that the range for PDHA does not overlap with the range of the DMR or LDA, suggesting that the performance difference is statistically significant. ANOVA tests reject the null hypothesis that the log-likelihood values of the four models are the same (p-value < 0.01). The t-tests suggest that PDHA is significantly better than DMR, LDA, and PDHA_C.

The other research question regards the contribution of the temporal features to the increase in performance. Observing the relative performances between PDHA and PDHA_C can shed light on this question. The difference between these two models can be considered as the contribution of temporal features. Panels (D) to (F) plot the log-likelihood values of these two models. The log-likelihood difference is quite large for the WMT and RTS datasets and is smaller for that of NYT. One possible reason for this is that NYT articles lack intra-day variables; in addition, the weekly and monthly cyclical patterns are less informative compared to the fine-grained timestamps in WMT and RTS. The mode of PDHA is higher than that of PDHA_C; this holds true for all three testbeds. The t-tests comparing PDHA and PDHA_C are all significant at the 99% confidence level, suggesting that temporal features alone contribute positively to the performance of topic modeling.

The log-likelihood difference between the DMR and LDA can also be interpreted as the contribution of temporal features. Panels (A) to (C) clearly indicate that DMR delivers a higher log-likelihood than LDA. The performance difference is significant at the 99% confidence level, and is usually larger than the gap between PDHA and PDHA_C. One reason for this is the fixed prior parameter α for document–topic distribution constraining LDA's ability to adapt to the training data. DMR allows constant terms for each topic to be estimated based on data, a fact that provides an advantage in addition to the temporal features. This result is, in general, consistent with studies that investigate the effect of the prior distribution on LDA [31].

As discussed above, the difference between DMR and PDHA is due to the novel model structure of PDHA. PDHA performs significantly better than DMR (p-value < 0.01), suggesting that the upstream token-level structure performs better than the upstream document-level structure in DMR. Despite the performance gain of PDHA, Panels (A) to (C) reveal a disadvantage of PDHA. The testing log-likelihoods of the LDA and DMR models have smaller variances than those of the PDHA models. One possible reason is that the intermediate parameters of our LDA and DMR models are collapsed, which allow for more efficient estimation. The PDHA models are also estimated via Gibbs sampling, but only with $\phi_{z_{di}}$ integrated out. Thus, we observe a higher variation of the log-likelihood values for the PDHA and PDHA_C models.

To understand the influence of the number of latent topics, we performed tests using varying latent topic numbers: 50, 100, 500, and 1000. We report the results in Fig. 10. Increasing the latent topic number from 50 to 100 improved the performance of PDHA and other baseline models. The improvement of PDHA is smaller compared to that of LDA and DMR. We see this effect consistently across the three testbeds. Further increasing the latent topic number from 100 to 500 has different effects on the three testbeds. All models showed higher likelihood values for the WMT testbed while the likelihood values remains flat for the NYT and RTS datasets. The likelihood values only changed slightly when the number of latent topics increased to 1000 from 500.

In addition to the discussion on computational complexity in Section 3.3.2, we have also included the running time comparisons of PDHA, LDA, and DMR. We conducted all experiments on a PC with an Intel i5 CPU and 16GB of RAM. Fig. 11 summarizes the running times of PDHA, DMR, and LDA using different numbers of latent topics. For a model with 50 latent topics, PDHA, DMR, and LDA took 98.9 min, 19.8 min, and 18 min to complete, respectively. PDHA took longer to complete but provided additional information regarding the short-term cyclical patterns.

The running time for any of the three models is roughly linear with respect to the number of topics when the number is small (e.g., $J \leq 500$). However, it took much longer for PDHA to finish when the number of topics was large (e.g., $J = 1000$). In fact, PDHA took about a week to complete when the number of topics was 1000. In contrast, LDA took less than five hours to finish estimating a model with 1000 topics. The additional computational cost of incorporating short-term cyclical dynamics is clearly high when using a large amount of latent topics (e.g., 1000). However, the additional running time can be justified when we need to include a moderate amount of latent topics (e.g., 50 or 100) because PDHA provides richer information, better summarizing the dynamics of latent topics. These results are generally consistent with the theoretical discussion in Section 3.3.2.

5. Conclusions

This paper presents a PDHA model that includes temporal features into topic models. The model adopts a multinomial Probit regression structure to incorporate temporal features. Temporal features that model monthly, weekly, and daily cyclical patterns allow PDHA to capture short-term cyclical patterns that naturally occur in text from user-generated content, newswire, and newspapers.

To facilitate the efficient estimation of PDHA, we developed an augmented Gibbs sampling algorithm that iteratively updates latent topic variables and regression coefficients. Experimental results show that PDHA outperformed LDA and DMR in terms of hold-out log-likelihood. Both the temporal features and the more flexible upstream token-level model structure contribute to the improved performance of PDHA. Likelihood comparisons show that daily and weekly cyclical patterns are more important in WMT and RTS compared to NYT, suggesting that short-term cyclical dynamics may be more important in some datasets.

Extending the PDHA model to include other meta-data variables, such as authors and citations, is possible. To further improve performance, we are also working on including token-level features such as the WordNet senses and the position of a word in a document. Another interesting application is to apply PDHA to context-aware recommender systems. PDHA can be used to recommend different products based on holiday seasons, day-of-the-week, time-of-day, and other relevant context variables.

The other extension is to consider “latent topics” that may have ordinal relationships. One application might be mining product reviews that naturally fit into this application. A structure similar to ordinary Probit regression could be adopted for this purpose. Small modifications to the inference method would be needed for such applications.

Acknowledgments

This work was supported in part by the Ministry of Science and Technology of Taiwan under grants 100-2410-H-002-025-MY3 and 103-2410-H-002-110-MY3.

Appendix A. Summary of selected time-dependent topic models

Table A.1
Summary of selected time-dependent topic models.

	Latent Dirichlet (LDA)	Dynamic topic model (DTM)	Topic over time (TOT)	Temporal collection (TC)	Probit-Dirichlet hybrid allocation (PDHA)
Main reference	Griffiths and Steyvers [4] and Blei et al. [1]	Blei and Laferty [12]	Wang and McCallum [11]	Hong et al. [5]	This study
Model estimation method	Collapsed Gibbs sampling, variational Bayes	Variational Bayes	Stochastic EM	Stochastic EM	Augmented Gibbs sampling
Topic–token distribution (topic meaning)	Fixed	Markovian (random walk)	Fixed	Fixed	Fixed
Time range	Not considered	Long-term (years)	Long-term (months-to-years)	Medium-term (days-to-months)	Short-term (hours-to-days)
Continuous or discrete time	Not considered	Discrete	Continuous	Continuous	Continuous
Time-dependence structure	Not considered	Upstream at document-level	Downstream at token-level	Upstream at document-level	Upstream at token-level
Time-dependent topic occurrence	Not considered	Dirichlet prior with random walk	Beta distribution for normalized time variable	Gamma distribution	Multinomial Probit regression with Serfling model

Appendix B. Major steps for drawing regression coefficients

```
Inputs: Updated latent topic $Z$ and regression coefficients from the previous sweep:
        $Q^{(old)}$, $G^{(old)}$, $\Sigma^{(old)}$, $H^{(old)}$
Outputs: Updated regression coefficients: $Q^{(new)}$, $G^{(new)}$, $\Sigma^{(new)}$, $H^{(new)}$

  1) Draw $a^{*2}$ from $trace(\Sigma^{-1^{(old)}})/\chi^2_{(J-1)^2}$.
  2) Draw $\widetilde{H}^*_{di,j}$ by first drawing $H^*_{di,j}$ from the truncated normal distribution using the
     original model and scaling the result by $\widetilde{H}^*_{di,j} = a^* H^*_{di,j}$.
  3) Draw scaled regression coefficients $\tilde{g}^*_j$ from the multivariate normal distribution
     $(j = 1, 2, \dots, J-1)$.
  4) Draw scaled document-specific topic tendencies $\tilde{q}^*_{d,j}$ from the normal distribution
     $(d = 1, 2, \dots, D$ and $j = 1, 2, \dots, J-1)$.
  5) Draw another scale variable $a^{**2}$ from the inverse Chi-squared distribution that
     depends on the newly-updated regression coefficients and document-specific topic
     tendencies.
  6) Compute $Q^{(new)}$, and $G^{(new)}$ by $q^{(new)}_{d,j} = \tilde{q}^*_{d,j}/a^{**}$, $g^{(new)}_j = \tilde{g}^*_j/a^{**}$.
  7) Draw $\Sigma^{(new)}$ by first sampling $\tilde{\Sigma}^*$ from the inverse Chi-squared distribution based
     on the scaled model and set $\Sigma^{(new)} = \tilde{\Sigma}^*/\tilde{\Sigma}^*_{11}$
  8) Compute $H^{(new)}$ by $H^{(new)}_{di,j} = \frac{\widetilde{H}^*_{di,j}}{\sqrt{\tilde{\Sigma}^*_{11}}}$.
```

Fig. B.1. Major steps for drawing regression coefficients.

References

[1] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.

[2] D.M. Blei, J.D. Lafferty, Correlated topic models, Neural Information Processing Systems (NIPS), 2006.

[3] D. Mimno, A. McCallum, Topic models conditioned on arbitrary features with Dirichlet-multinomial regression, Uncertainty in Artificial Intelligence (UAI), 2008.

[4] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proceedings of the National Academy of Sciences of the United States of America 101 (2004) 5228–5235.

[5] L. Hong, B. Dom, S. Gurumurthy, K. Tsioutsiouliklis, A time-dependent topic model for multiple text streams, Presented at the Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, California, USA, 2011.

[6] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, M. Steyvers, Learning author-topic models from text corpora, ACM Transactions on Information Systems 28 (2010) 1–38.

[7] K. El-Arini, M. Xu, E.B. Fox, C. Guestrin, Representing Documents Through Their Readers, Presented at the Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, 2013.

[8] H. Ma, H. Cao, Q. Yang, E. Chen, J. Tian, A habit mining approach for discovering similar mobile users, Presented at the Proceedings of the 21st International Conference on World Wide Web, Lyon, France, 2012.

[9] X. Wu, J. Yan, N. Liu, S. Yan, Y. Chen, Z. Chen, Probabilistic latent semantic user segmentation for behavioral targeted advertising, Presented at the Proceedings of the Third International Workshop on Data Mining and Audience Intelligence for Advertising, Paris, France, 2009.

[10] M. Deshpande, G. Karypis, Item-based Top-N recommendation algorithms, ACM Transactions on Information Systems 22 (2004) 143–177.

[11] X. Wang, A. McCallum, Topics over time: a non-Markov continuous-time model of topical trends, Presented at the KDD, Philadelphia, Pennsylvania, USA, 2006.

[12] D.M. Blei, J.D. Laferty, Dynamic topic models, *International Conference on Machine Learning*, Pittsburgh, PA, 2006.

[13] D.M. Blei, Introduction to probabilistic topic models, Communications of the ACM, 2012 (forthcoming).

[14] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[15] J. Besag, Spatial interaction and the statistical analysis of lattice systems, Journal of the Royal Statistical Society. Series B (Methodological) 36 (1974) 192–236.

[16] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, IEEE Transactions on Pattern Analysis and Machine Intelligence 6 (1984) 721–741.

[17] M. Jordan, Z. Ghahramani, T. Jaakkola, L. Saul, Introduction to variational methods for graphical methods, Machine Learning 37 (1999) 183–233.

[18] Y.W. Teh, D. Newman, M. Welling, A collapsed variational bayesian inference algorithm for latent Dirichlet allocation, NIPS, 2006.

[19] C. Wang, D. Blei, D. Heckerman, Continuous time dynamic topic models, Uncertainty in Artificial Intelligence (UAI), 2008.
[20] H.-Y. Lin, Efficient classifiers for multi-class classification problems, Decision Support Systems 53 (2012) 473–481.
[21] K. Imai, D.A.v. Dyk, A Bayesian analysis of the multinomial probit model using marginal data augmentation, Journal of Econometrics 124 (2005) 311–334.
[22] J.H. Albert, S. Chib, Bayesian analysis of binary and polychotomous response data, Journal of the American Statistical Association 88 (1993) 669–679.
[23] R.E. Serfling, Methods for current statistical analysis of excess pneumonia–influenza deaths, Public Health Reports 78 (1963) 494–506.
[24] K.S. Chan, C.J. Geyer, Discussion: Markov chains for exploring posterior distributions, The Annals of Statistics 22 (1994) 1747–1758.
[25] J.M. Hammersley, D.C. Handscomb, Monte Carlo Methods, Halsted, New York, 1964.
[26] X.-L. Meng, D.A.V. Dyk, Seeking efficient data augmentation schemes via conditional and marginal augmentation, Biometrika 86 (1999) 301–320.
[27] T.H. Cormen, C.E. Leiserson, R.L. Rivest, Introduction to Algorithms, MIT Electrical Engineering and Computer Science1990.
[28] D. Coppersmith, S. Winograd, Matrix multiplication via arithmetic progressions, Journal of Symbolic Computation 9 (1990) 251–280.
[29] A.E. Gelfand, Gibbs sampling, Journal of the American Statistical Association 95 (2000) 1300–1304.
[30] H.M. Wallach, I. Murray, R. Salakhutdinov, D. Mimno, Evaluation methods for topic models, Proceedings of the 26th International Conference on Machine Learning (ICML), 2009.
[31] H.M. Wallach, D. Mimno, A. McCallum, Rethinking LDA: why priors matter, Neural Information Processing Systems (NIPS), 2009.

Hsin-Min Lu received the bachelor's degree in business administration and MA degree in economics from the National Taiwan University, and the PhD degree in information systems from the University of Arizona. He is an Assistant Professor in the Department of Information Management at the National Taiwan University. His papers have appeared in IEEE Transactions of Knowledge and Data Engineering, IEEE Intelligent Systems, Journal of Biomedical Informatics, International Journal of Medical Informatics, Journal of Forecasting, Decision Support Systems, and Review of Accounting Studies. His research interests include data mining, text mining, and health informatics.