九州大学学術情報リポジトリ
Kyushu University Institutional Repository

A Bayesian nonparametric topic model for repeated measured data : An application to prescription data

Okui, Tasuku
Medical Information Center, Kyusyu University Hospital

https://hdl.handle.net/2324/4150683

出版情報：Behaviormetrika, pp.1-12, 2020-07-07. The Behaviormetric Society of
バージョン：
権利関係：

KYUSHU UNIVERSITY

Noname manuscript No.
(will be inserted by the editor)

A Bayesian nonparametric topic model for repeated measured data

An application to prescription data

Tasuku Okui

Received: date / Accepted: date

Abstract Topic models are currently used in many fields, particularly for marketing or medical science data analysis, often where an individual subject is repeatedly measured. A topic tracking model(TTM) that can consider the persistency of topics of individual subjects has been already proposed. Although the TTM estimates several parameters for each timepoint through online learning, offline learning should be utilized for analyses of preexisting data sets. Additionally, when a topic model is used, the number of topics should be decided in advance. However, deciding an appropriate number of topics is often difficult. Therefore, we propose a TTM with offline learning and a Bayesian nonparametric TTM (BNPTTM) for time series data sets where data from individual subjects are repeated measures. The performance of the proposed topic model is evaluated using an actual prescription data set. Our results suggest that the TTM with offline learning has better predictive ability than the existing TTM, and the BNPTTM can deduce the number of topics from a given data set.

Keywords Repeated measures data · Bayesian nonparametric model · Prescription data · Latent Dirichlet allocation model

1 Introduction

Topic models are machine learning models, which can extract latent topics based on clusters of similar words from documents (Blei et al. 2003). Although topic models are mainly used for natural language processing, they are useful for analyzing several different kinds of data, e.g., consumer data in marketing or prescription data in medicine (Zafari and Ekin 2019; Iwata et al. 2009). When a topic model is applied

Tasuku Okui
Medical Information Center, Kyusyu University Hospital, Fukuoka city, Japan
Tel.: +81-092-642-5881
Fax: +81-092-642-5889
E-mail: task10300@gmail.com

to prescription data, it can summarize the prescription patterns of a physician or elucidate the relation of prescribed medicines and diagnosed diseases (Zafari and Ekin 2019; Park et al. 2017). Additionally, when several medicines are included in a large data set, a topic model can be useful for summarizing it.

When topic models are applied to marketing or prescription data, traits of individual subjects are often measured repeatedly. When a trait is repeatedly measured, the correlation of data for each subject should be considered. The topic tracking model (TTM), a topic model that uses online learning and can consider change of subject interest and content change of topics over time, has been proposed (Iwata et al. 2009). The TTM was originally proposed for extracting purchase item topics based on the change of interest of consumers, and it was designed for data that were obtained sequentially. When static, preexisting data sets are available, which is often the case for repeatedly measured data, it is desirable to consider a model that uses offline learning.

When a topic model is used, the number of topics should be decided a priori; however, it is often difficult to decide an appropriate number of topics. If the number of topics is high, many topics are generated with relatively even proportions. To decide the number of topics computationally based on a given data set, a Bayesian nonparametric model is appropriate. A Bayesian nonparametric model is a method for reducing the parametric assumption of a Bayesian model (Müller and Fernando 2004), and a hierarchical Dirichlet process is often used for topic modeling (Teh et al. 2006). By using the hierarchical Dirichlet process, only topics that are minimally essential for data analyses can be extracted. The TTM also needs to be extended to a Bayesian nonparametric model to decide the number of topics from the data.

With this research, we propose a TTM with offline learning and a Bayesian nonparametric TTM (BNPTTM) with offline learning and evaluate their performance using a prescription data set.

2 Method

In this section, we explain the existing TTM and the proposed TTM with offline learning. Moreover, we show the hierarchical Dirichlet process (HDP) and the proposed BNPTTM with offline learning.

2.1 Topic tracking model

The existing TTM (Iwata et al. 2009) is illustrated in Figure 1. $z_{t,d,n}$ represent the $n^{th}(n=1,\ldots,N_{t,d})$ prescription count of a subject $d(d=1,\ldots,D_t)$ at time $t(t=1,\ldots,T)$, $w_{t,d,n}$ represent a prescribed drug for the nth prescription count of a subject d at time t, $\boldsymbol{\pi}_{t,d}$ represent the topic distribution of each subject d at t^{th} time, and $\boldsymbol{\phi}_{t,k}$ represent the drug distribution of topic $k(k=1,\ldots,K)$ at time t. $\alpha_{t,d}$ represent the degree of association between $\boldsymbol{\pi}_{t-1,d}$ and $\boldsymbol{\pi}_{t,d}$, and $\eta_{t,k}$ represent the degree of association between $\boldsymbol{\phi}_{t-1,k}$ and $\boldsymbol{\phi}_{t,k}$.

$\boldsymbol{\phi}_{t,k}$ are vectors that are $W\times1$, and $\boldsymbol{\pi}_{t,d}$ are vectors that are $K\times1$. $\alpha_{t,d}$ and $\eta_{t,k}$ represent scalars.

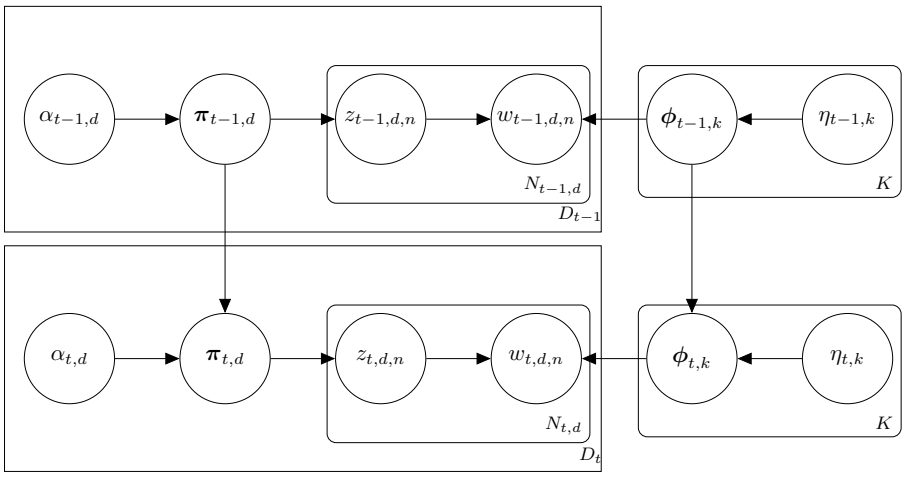

Fig. 1: **Graphical model of TTM (Iwata et al. 2009)**

Let D_t be the number of subjects at time t, K be the number of topics, and $N_{t,d}$ be the number of prescriptions for a subject d at time t. The frames in Figure 1 indicate that the numbers of parameters and variables in a given frame are the same as the values shown in the bottom right corner of each frame. In other words, the number of parameters for $z_{t,d,n}$ and $w_{t,d,n}$ is $N_{t,d}$ for a subject d at time t; the number of parameters for $\boldsymbol{\pi}_{t,d}$ and $\alpha_{t,d}$ is D_t for time t; and the number of parameters for $\eta_{t,k}$ and $\boldsymbol{\phi}_{t,k}$ is K for time t.

The data generation process of the existing TTM is as follows.

1. For each topic k at time t,
   (a) Draw $\boldsymbol{\phi}_{t,k} \sim Dirichlet(\eta_{t,k}\hat{\boldsymbol{\phi}}_{t-1,k})$
2. For each subject d at time t,
   (a) Let $\boldsymbol{\pi}_{t,d} \sim Dirichlet(\alpha_{t,d}\hat{\boldsymbol{\pi}}_{t-1,d})$
   (b) For each nth prescription count,
      i. Draw $z_{t,d,n} \sim Multinomial(1, \boldsymbol{\pi}_{t,d})$
      ii. Draw $w_{t,d,n} \sim Multinomial(1, \boldsymbol{\phi}_{t,z_{t,d,n}})$

In the above process, $\boldsymbol{\phi}_{t,z_{t,d,n}}$ denotes the drug distribution of topic $z_{t,d,n}$ at time t. $\hat{\boldsymbol{\phi}}_{t-1,k}$ and $\hat{\boldsymbol{\pi}}_{t-1,d}$ denote the estimates of $\boldsymbol{\phi}_{t-1,k}$ and $\boldsymbol{\pi}_{t-1,d}$, respectively. $\boldsymbol{\phi}_{t,k}$ and $\boldsymbol{\pi}_{t,d}$ are drawn based on the values of $\hat{\boldsymbol{\phi}}_{t-1,k}$ and $\hat{\boldsymbol{\pi}}_{t-1,d}$, respectively. Moreover, similar to the original topic model (Blei et al. 2003), the model assumes that each prescription drug, $w_{t,d,n}$, is drawn based on topic $z_{t,d,n}$ and each topic $z_{t,d,n}$ is drawn based on topic distribution $\boldsymbol{\pi}_{t,d}$.

The parameters are estimated by collapsed Gibbs sampling (Iwata et al. 2009). They are sequentially estimated in each timepoint t by using online learning. In other words, the parameters of the timepoint t are estimated after estimating those of the timepoint $t-1$.

The existing TTM (Iwata et al. 2009) is effective when data are sequentially obtained in real time. However, when data to be analyzed are already obtained, we should consider a model that uses offline learning. Therefore, we propose a TTM with offline learning. The data generation process of the proposed method is not largely different from that of the existing method; however, in the proposed method, $\boldsymbol{\phi}_{t,k}$ and $\boldsymbol{\pi}_{t,d}$ are drawn based on the values of $\boldsymbol{\phi}_{t-1,k}$ and $\boldsymbol{\pi}_{t-1,d}$, respectively, instead of $\hat{\boldsymbol{\phi}}_{t-1,k}$ and $\hat{\boldsymbol{\pi}}_{t-1,d}$, respectively. Therefore, in the proposed method, the parameters of all the timepoints are estimated together in each iteration of Gibbs sampling. The parameters can be estimated by collapsed Gibbs sampling just like they were estimated in the existing TTM.

2.2 Bayesian nonparametric topic tracking model

In this section, we first explain the HDP that is used to construct the BNPTTM. Then, we discuss the proposed BNPTTM with offline learning that uses HDP.

2.2.1 Hierarchical Dirichlet process

The building block of HDP is that the random base measure G_0 for a Dirichlet process $G_j \sim DP(\xi, G_0)$ is itself a draw from a Dirichlet process $G_0 \sim DP(\gamma, H)$ (Teh et al. 2006).

$$\begin{gathered} G_0 \mid \gamma, H \sim DP(\gamma, H), \\ G_j \mid \xi, G_0 \sim DP(\xi, G_0) \\ for j \in J \end{gathered}$$

G_j are group (j) specific random measures, and H is the base measure of G_0. J is the index set.

The proposed BNPTTM uses the HDP based on the stick-breaking process, and we show a stick-breaking construction for the HDP. The random base measure G_0 can be represented in the following equation.

$$\begin{gathered} G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k^{**}}, \\ \delta_{\theta_k^{**}} \mid H \sim H. \end{gathered}$$

θ_k^{**} represent independent random variables distributed according to G_0, where $\delta_{\theta_k^{**}}$ is a probability measure concentrated at θ_k^{**}. The random variables β_k are generated based on the stick-breaking process.

$$\begin{gathered} v_k \mid \gamma \sim Beta(1, \gamma), \\ \beta_k = v_k \prod_{l=1}^{k-1} (1 - v_l) \\ for\ k = 1, \ldots, \infty \end{gathered}$$

γ is the hyperparameter of the first level Dirichlet process. The stick-breaking process postulates an infinite number of latent classes to a model, and by using this method, only the proportion of latent classes that are needed to represent the data set increase significantly.

A set of group (j) specific random measures G_j are also represented by the following equation.

$$G_j = \sum_{k=1}^{\infty} \pi_{j,k} \delta_{\theta_k^{**}}$$

Parameters of each group $\pi_{j,k}$ are also generated based on beta distributions for the stick-breaking process (Teh et al. 2006).

$$\pi'_{j,k} \sim Beta(\xi\beta_k, \xi(1 - \sum_{l=1}^{k} \beta_l))$$

$$\pi_{j,k} = \pi'_{j,k} \prod_{l=1}^{k-1} (1 - \pi'_{j,l})$$

$$for\ k = 1, \ldots, \infty.$$

$\boldsymbol{\pi}_j(\boldsymbol{\pi}_j = (\pi_{j,1}, ..., \pi_{j,K})^T)$ and $\boldsymbol{\pi}'_j(\boldsymbol{\pi}'_j = (\pi'_{j,1}, ..., \pi'_{j,K})^T)$ are group specific vectors, and ξ is a hyperparameter for beta distribution of the second level Dirichlet process.

2.2.2 The proposed Bayesian nonparametric topic tracking model

In this section, we propose a novel BNPTTM with offline learning using HDP. We included a dependent Dirichlet process method that combines the stick-breaking process with covariates of a subject for the model (Dunson and Park 2008; Hossain et al. 2013).

The proposed model is illustrated in Figure 2. We extended the proposed TTM by offline learning, and the parameters for constructing HDP were added to the model. The parameters γ, ξ, and β_k are scalars and they act as previously stated. The number of parameters $\beta_{t,k}$, $\boldsymbol{\phi}_{t,k}$ are ∞ for time t.

The data generation process of the proposed method is as follows:

1. For each topic k,
   (a) Draw $v_k \sim Beta(1,\gamma)$
   (b) Let $\beta_k = v_k \prod_{i=1}^{k-1}(1 - v_i)$
   (c) For each time t,
      i. Draw $\boldsymbol{\phi}_{t,k} \sim Dirichlet(\eta_{t,k}\boldsymbol{\phi}_{t-1,k})$
2. For each subject d at time t,
   (a) For each kth topic,
      i. Draw $\pi'_{t,d,k} \sim Beta(\xi\beta_k, \xi(1 - \sum_{l=1}^{k} \beta_l))$
      ii. Let $\pi_{t,d,k} = \pi'_{t,d,k}\text{expit}(\alpha_{t,d}\pi_{t-1,d,k}) \prod_{l=1}^{k-1}(1 - \pi'_{t,d,l}\text{expit}(\alpha_{t,d}\pi_{t-1,d,l}))$
   (b) For each nth prescription count,
      i. Draw $z_{t,d,n} \sim Multinomial(1,\boldsymbol{\pi}_{t,d})$
      ii. Draw $w_{t,d,n} \sim Multinomial(1,\boldsymbol{\phi}_{z_{t,d,n}})$

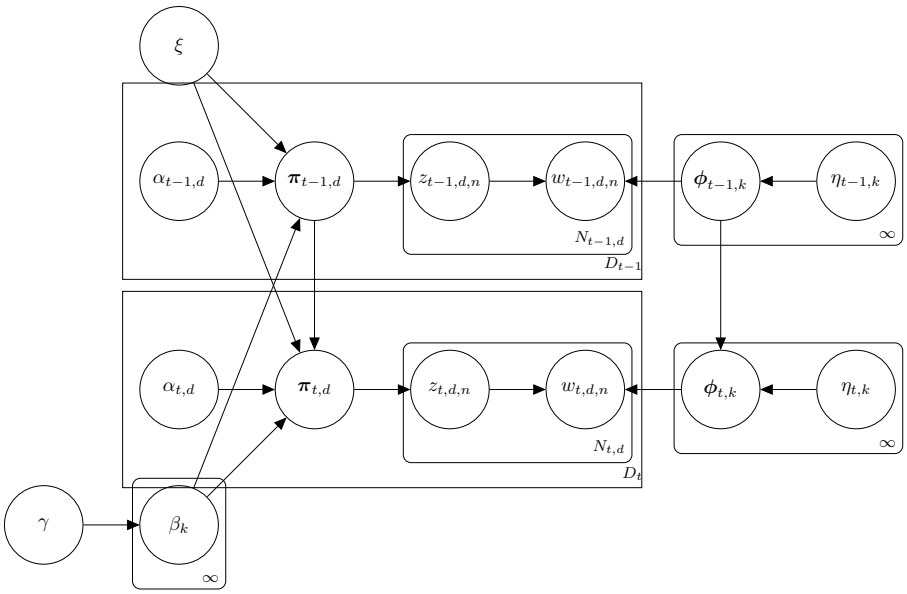

Fig. 2: **Graphical model of the proposed Byesian nonparametric model**

In the above process, expit signifies sigmoid function. A method of a dependent Dirichlet process that combines the stick-breaking process and covariates of a subject (Dunson and Park 2008) was used to combine parameters from the stick-breaking process, β_k, with subject specific $\alpha_{t,d}\pi_{t-1,d,k}$. $\alpha_{t,d}\pi_{t-1,d,k}$ was scaled to range from 0 to 1 by sigmoid function and was multiplied by $\pi'_{t,d,k}$, which were drawn based on β_k. With this method, topic distributions of specific times are generated from HDP while taking into account the topic distributions of the previous timepoints.

The proposed method postulates a model that uses offline leaning, which is appropriate for when data of all timepoints are already acquired. The proposed method can also be formulated to operate as an online-learning model for data that is sequentially acquired. Additionally, the proposed method includes a filtering method, i.e., a method that estimates parameters of a specific time based on previous time points. If all timepoints are already collected, estimating parameters with a smoothing method can be done. In which case, topic distributions and medication (word) distributions are generated based on not only those at previous timepoints, but also those at subsequent timepoints. This method is referred to as the smoothing method, while the method shown in Figure 2 shall be referred to as the filtering method.

The smoothing method is shown below. The data generation process of the smoothing method is as follows:

1. For each topic k,
   (a) Draw $v_k \sim Beta(1,\gamma)$
   (b) Let $\beta_k = v_k \prod_{i=1}^{k-1}(1 - v_i)$
   (c) For each time t,

i. Draw $\boldsymbol{\phi}_{1,t,k} \sim Dirichlet(\eta_{1,t,k}\boldsymbol{\phi}_{1,t-1,k})$
ii. Draw $\boldsymbol{\phi}_{2,t,k} \sim Dirichlet(\eta_{2,t,k}\boldsymbol{\phi}_{2,t+1,k})$
iii. Let $\boldsymbol{\phi}_{t,k} = \boldsymbol{\phi}_{1,t,k} \odot \boldsymbol{\phi}_{2,t,k}/(\boldsymbol{\phi}_{1,t,k}{}'\boldsymbol{\phi}_{2,t,k})$

2. For each subject d at time t,
   (a) For each kth topic,
      i. Draw $\pi'_{t,d,k} \sim Beta(\xi\beta_k, \xi(1 - \sum_{l=1}^{k}\beta_l))$
      ii. Let $\pi_{1,t,d,k} = \pi'_{t,d,k}\mathrm{expit}(\alpha_{1,t,d}\pi_{1,t-1,d,k})\prod_{l=1}^{k-1}(1 - \pi'_{t,d,l}\mathrm{expit}(\alpha_{1,t,d}\pi_{1,t-1,d,l}))$
      iii. Let $\pi_{2,t,d,k} = \pi'_{t,d,k}\mathrm{expit}(\alpha_{2,t,d}\pi_{2,t+1,d,k})\prod_{l=1}^{k-1}(1 - \pi'_{t,d,l}\mathrm{expit}(\alpha_{2,t,d}\pi_{2,t+1,d,l}))$
   (b) Let $\boldsymbol{\pi}_{t,d} = \boldsymbol{\pi}_{1,t,d} \odot \boldsymbol{\pi}_{2,t,d}/(\boldsymbol{\pi}_{1,t,d}{}^T\boldsymbol{\pi}_{2,t,d})$
   (c) For each nth prescription count,
      i. Draw $z_{t,d,n} \sim Multinomial(1, \boldsymbol{\pi}_{t,d})$
      ii. Draw $w_{t,d,n} \sim Multinomial(1, \boldsymbol{\phi}_{z_{t,d,n}})$

$\boldsymbol{\phi}_{1,t,k}(\boldsymbol{\phi}_{1,t,k} = (\phi_{1,t,k,1}, ..., \phi_{1,t,k,W})^T)$ represent medication distribution of a topic k at time t estimated from the previous time, and $\boldsymbol{\phi}_{2,t,k}(\boldsymbol{\phi}_{2,t,k} = (\phi_{2,t,k,1}, ..., \phi_{2,t,k,W})^T)$ represent medication distribution of topic k at time t estimated from the subsequent time. $\eta_{1,t,k}$ represent the parameter of degree of association between $\boldsymbol{\phi}_{1,t-1,k}$ and $\boldsymbol{\phi}_{1,t,k}$, and $\eta_{2,t,k}$ represent the parameter of degree of association between $\boldsymbol{\phi}_{2,t+1,k}$ and $\boldsymbol{\phi}_{1,t,k}$.

Similarly, $\boldsymbol{\pi}_{1,t,d}(\boldsymbol{\pi}_{1,t,d} = (\pi_{1,t,d,1}, ..., \pi_{1,t,d,K})^T)$ represent topic distribution of each subject d at time t estimated from the previous time, and $\boldsymbol{\pi}_{2,t,d}(\boldsymbol{\pi}_{2,t,d} = (\pi_{2,t,d,1}, ..., \pi_{2,t,d,K})^T)$ represent topic distribution of each subject d at time t estimated for the subsequent time. $\alpha_{1,t,d}$ represent the parameter of degree of association between $\boldsymbol{\pi}_{1,t-1,d}$ and $\boldsymbol{\pi}_{1,t,d}$, and $\alpha_{2,t,d}$ represent the parameter of degree of association between $\boldsymbol{\pi}_{2,t+1,d}$ and $\boldsymbol{\pi}_{2,t,d}$. $\boldsymbol{\pi}_{1,t,d}$ and $\boldsymbol{\pi}_{2,t,d}$ are vectors of $K\times1$, and $\boldsymbol{\phi}_{1,t,k}$ and $\boldsymbol{\phi}_{2,t,k}$ are vectors of $W\times1$. $\eta_{1,t,k}$, $\eta_{2,t,k}$, $\alpha_{1,t,d}$, and $\alpha_{2,t,d}$ are scalars.

In this model, $\boldsymbol{\phi}_{t,k}$ are estimated by normalizing the product of $\boldsymbol{\phi}_{1,t,k}$ and $\boldsymbol{\phi}_{2,t,k}$, and $\boldsymbol{\pi}_{t,d}$ are estimated by normalizing the product of $\boldsymbol{\pi}_{1,t,d}$ and $\boldsymbol{\pi}_{2,t,d}$. Hereinafter, we denote the Bayesian nonparametric model with filtering as "BNPTTM by filtering method," and the Bayesian nonparametric model with smoothing as "BNPTTM by smoothing method."

3 Numerical implementation

The numerical computing method for topic models are various, and an expectation-maximization (EM) algorithm and Gibbs sampling are frequently used. We used Stan for estimating parameters (Stan Development Team 2018). Stan optimizes each parameter using the Hamiltonian Monte Carlo method for maximizing likelihood. We used R3.5.1 (R Core Team 2017) for the analysis, and Stan was conducted with R-package rstan. The parameters of the existing TTM and the proposed TTM with offline learning were estimated by collapsed Gibbs sampling. The codes for the methods were written by using an in-house pipeline written with the R package Rcpp (Eddelbuettel and Francois 2011). Perplexity was also calculated by Rcpp.

4 Performance evaluation

We evaluated the performance of the proposed methods with real prescription data using perplexity, which is often used as one of performance evaluation methods for a topic model. Perplexity evaluates the predictive ability of a model to new data. We subsampled half of the sum of prescriptions for each subject and designated the subsampled data set as the training data set and the other as the experimental test data set. We trained our model using the training data set, and calculated perplexity with the test data set.

$$\text{Perplexity} = \exp\{-\frac{\sum_{t=1}^{T}\sum_{d=1}^{D_t}\sum_{n=1}^{N_{t,d}}\log p(w_{t,d,n})}{\sum_{t=1}^{T}\sum_{d=1}^{D_t}N_{t,d}}\},$$
$$p(w_{t,d,n}) = \hat{\pi}_{t,d,z_{t,d,n}}\hat{\phi}_{t,z_{t,d,n},w_{t,d,n}}$$

where $\hat{\pi}_{t,d,z_{t,d,n}}$ denote topic distribution estimated in the training data set, and $\hat{\phi}_{t,z_{t,d,n},w_{t,d,n}}$ denote medication distribution estimated in the training data set. The lower the perplexity is, the higher the predictive ability for the test data set is. Perplexity can also be calculated by using different subjects from the training data, which is like the method used by Park et al. (2017). However, in this study, we simply used the standard method for calculating perplexity (Blei et al. 2003).

Although perplexity is a general method to evaluate topic models, it cannot evaluate whether the methods can correctly deduce the number of topics from a data set. Rather, perplexity may become lower when the number of topics is affected by the prespecified number of topics. In addition to perplexity, we aggregated the number of topics with the largest or the second largest proportions for any subject and the mean of proportions of these topics. We focused on the number of these topics because topics with high rank proportions for any subject are important when interpreting the topic distribution. If the number of these topics is not significantly changed by the prespecified number of topics and values stay small, it can be concluded that the number of topics is accurately determined by the data set. We prespecified the upper limit of the number of topics from a range of 10 to 50 by increments of 10, and investigated how the evaluation measure results changed at the various increments.

In this study, we used previously-acquired data about kampo medicines, a traditional type of Japanese herbal medicine, prescribed in a university hospital from 2008 to 2018. We aggregated the data for each clinical department by month and created count data sets of kampo medicine prescriptions. We used monthly data that had a sum of prescriptions larger than 10, and used only medicines whose sum of prescriptions were larger than 100 across departments. Additionally, we only included clinical departments with data for all years. Kampo medicines were categorized based on the drug classification system, KEGG: Kyoto Encyclopedia of Genes and Genomes (Kyoto Encyclopedia of Genes and Genomes).

Table 1 shows the summary statistics for the data in this study. Although the number of different medications is 90, the mean of the types of medications for each observation is 20.72. Therefore, compared with total number of medications, the number of medications each clinical department prescribes in one month is relatively restricted.

Table 1: The attributes of data

Items	Values
The number of observations*	275
The number of clinical departments*	25
The number of medications*	90
The number of time for each clinical department data†	11(0.00)
The types of medications for each observation†	20.72(16.52)
The number of prescriptions for each observation†	923.39(1284.47)

* Cases
† Mean(Standard deviations)

Table 2: The result of perplexity

	Number of topics				
Method	10	20	30	40	50
The existing TTM*	16.55	15.51	15.29	14.84	14.95
The TTM by offline learning	14.94	14.74	14.48	14.46	14.44
BNPTTM by filtering method	22.94	24.17	21.87	21.60	23.78
BNPTTM by smoothing method	21.14	20.83	20.46	21.23	19.74

* The existing TTM, which are estimated by online learning

Table 3: The number of topics with high rank proportions for any subject and the mean of the proportions

		The number of topics				
The topics*	Method	10	20	30	40	50
The largest topics	online†	9(0.57)	18(0.57)	19(0.53)	27(0.58)	32(0.52)
	offline‡	10(0.611)	16(0.51)	21(0.48)	24(0.46)	23(0.46)
	filtering§	5(0.55)	5(0.51)	7(0.53)	6(0.54)	5(0.58)
	smoothing¶	6(0.80)	4(0.64)	4(0.69)	7(0.71)	6(0.68)
The largest or the second largest topics	online†	10(0.83)	19(0.83)	27(0.78)	34(0.82)	44(0.76)
	offline‡	10(0.83)	19(0.72)	26(0.68)	32(0.65)	34(0.64)
	filtering§	7(0.77)	7(0.72)	9(0.75)	6(0.76)	7(0.77)
	smoothing¶	8(0.95)	6(0.88)	7(0.91)	8(0.91)	9(0.89)

* The topics to be taken into account for aggregation
† The existing TTM, which are estimated by online learning
‡ The proposed TTM by offline learning
§ The BNPTTM by filtering method
¶ The BNPTTM by smoothing method

The result of the perplexity calculation is shown in Table 2. The perplexity values of the proposed TTM with offline learning were lower than those of the existing TTMirrespective of the prespecified number of topics. In addition, the values of perplexity of the BNPTTM by smoothing method were lower than those of the BNPTTM by filtering method. Furthermore, the values of perplexity of the TTMs that are not Bayesian nonparametric models were lower than those of the BNPTTMs. Table 3 shows the number of topics with the largest or second-largest proportions

for any subject and the mean proportions. The means were approximately 0.5~0.7 for the largest topics and around 0.7~0.9 for the combined largest and second-largest topics. A topic with the largest or second-largest proportion for an individual subject accounted for a large percentage in all topics. In our analysis, the numbers were relatively stable for the BNPTTMs compared with those of the TTMs that are not Bayesian nonparametric models regardless of the prespecified number of topics With respect to the TTMs that are not Bayesian nonparametric models, the numbers of the high rank topics had a tendency to increase with an increase in the pre-specified number of topics. Although the size of the numbers of the high rank topics of the BNPTTM by filtering method and the BNPTTM by smoothing method differed with respect to the prespecified number of topics, the mean proportions of the high rank topics of the smoothing method were higher than those of the filtering method.

5 Discussion & Conclusion

We proposed a TTM with offline learning and a BNPTTM with offline learning for time series data sets where traits of subjects are repeatedly measured. When perplexity was calculated, the proposed TTM with offline learning showed better predictive performance compared with the existing TTM. By using information of all the timepoints to estimate the parameters together, the predictive ability of the model seems to improve. Therefore, if data to be analyzed are already earned, using the proposed TTM with offline learning will be meaningful.

The predictive ability of the BNPTTM by smoothing method was better than that of the BNPTTM by filtering method. A possible explanation for this finding is that with the smoothing method, the parameters of each specific timepoint are not only drawn from the previous timepoint but also from the subsequent timepoint allowing the parameters to be more accurately estimated than those obtained with the filtering method. In addition, the proportion of topics of each subject is summarized into fewer topics for the smoothing method than those for the filtering method. The topic model is a kind of dimension reduction method, and the point of using it is that multiple variables are summarized into a smaller number of variables (topics). If the number of topics with large proportions is lower, we are able to focus on fewer topics. Therefore, using the BNPTTM by smoothing method might be better when using a Bayesian nonparametric TTM. .

However, predictive ability worsened by using the Bayesian nonparametric models. Although the largest possible value for perplexity is infinite, if we assign the same or randomly generated probability to each topic distribution and medication distribution, the perplexity becomes around 90. This is the same value as the total number of the prescribed drugs in this study. The lowest perplexity value would be one. Therefore, using a simple calculation, the perplexity of BNPTTM by smoothing can predict $((90-1)-(21.14-1)) \times 100/(90-1) = 77.37\%$ of the test data, when we use 10 as the value of the pre-specified number of topics. In contrast, the perplexity of the offline method can predict $((90-1)-(14.92-1)) \times 100/(90-1) = 84.36\%$ of the test data. It is difficult to determine whether this is a serious difference or not, but when using BNPTTMs, the predictive performance worsened.

The reason for the result of perplexity might be related to the results in Table 3. Although the number of topics with high rank proportions of the non-Bayesian nonparametric models was relatively large, the BNPTTMs restricted the number of topics with high proportions to fewer values. As a result, the overall fitting of the model might have worsened compared with that of the non-Bayesian nonparametric models. However, our results suggest that the number of topics for the proposed Bayesian nonparametric models is accurately deduced from the data set because the number of topics has relatively small values regardless of an increase in the prespecified number of topics. With respect to the non-Bayesian nonparametric models, the numbers of the high rank topics had a tendency to increase with an increase in the prespecified number of topics. In addition, the values of perplexity had a tendency to decrease with an increase in the prespecified number of topics, and selecting an appropriate number of topics from the data was difficult. Therefore, the Bayesian nonparametric method contributed to extracting minimally essential topics from the data. The results of this study indicate that when analyzing repeated measure data, deducing the number of topics from the data by the Bayesian nonparametric models and fitting the proposed TTM by offline learning, whose prespecified number of topics is the deduced one, might be better.

The limitation of this study is that we evaluated the methods with only one prescription data set. In the future, it will be interesting to see the results of this proposed method on other data sets as well as other types of data, e.g., consumer data for marketing or text data. The evaluation and summaries from analyzing these data sets may also provide novel insights into underlying trends.

6 Acknowledgements

We would like to thank the referees for their thorough review of the manuscript and appropriate comments.

Compliance with ethical standards

Conflict of interest

The corresponding author states that there is no conflict of interest.

References

1. Blei D, Ng AY, Jordan MI (2003) Latent dirichlet allocation. Journal of Machine Learning Research 3:993-1022.
2. Dunson D.B & Park J.H.(2008) Kernel stick-breaking process. Biometrika 95:307-323.
3. Eddelbuettel D and Francois R (2011). Rcpp: Seamless R and C++ Integration. Journal of Statistical Software, 40, 1-18. URL http://www.jstatsoft.org /v40/i08/.
4. Hossain M.M, Lawson A.B, Cai B, Choi J, Liu J and Kirby R.S (2013) Space-time stick-breaking processes for small area disease cluster estimation. Environmental Ecology 20:91-107.

5. Iwata T, Watanabe S, Yamada T and Ueda N (2009) Topic tracking model for analyzing consumer purchase behavior. In Proceedings of the twenty-first international joint conference on artificial intelligence (IJCAI-09):1427-1432.
6. Kyoto Encyclopedia of Genes and Genomes. https:// www.genome.jp/kegg/kegg_ja.html. Accessed 31 July 2019
7. Müller P and Fernando A.Q (2004) Nonparametric Bayesian Data Analysis. Statistical Science 19:95-110.
8. Park S, Choi D, Lim M, Cha W, Kim C and Moon I (2017) Identifying prescription patterns with a topic model of diseases and medications. Journal of Biomedical Informatics 75:35-47.
9. R Core Team (2017). R: a language and environment for statistical computing, R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org/. Accessed 31 July 2019
10. Stan Development Team (2018). RStan: the R interface to Stan R package version 2.17.3. http://mc-stan.org. Accessed 31 July 2019
11. Teh Y.W, Jordan M.I, Beal M.J, Blei D.M (2006) Hierarchical Dirichlet Processes. Journal of the American Statistical Association 101:1566-1581.
12. Zafari B and Ekin T (2019) Topic modelling for medical prescription fraud and abuse detection. Journal of the Royal Statistical Society Applied Statistics Series C 68:751-769.