繰り返し測定データに対するトピックモデル： 薬剤処方データへの適用

奥 井 佑

A Topic Model for Repeated-Measures Data:an Application to Prescription Drug Data

Tasuku OKUI

A topic model is an analysis method that extracts latent topics from words of documents. Topic models are used in many fields nowadays, and they are used for marketing or medical data, where data of a subject are often repeatedly measured. Although topic models for longitudinal data have been already proposed, a topic model that can be applied to more diverse repeated measures data needs to be proposed. Then, we proposed a topic model for repeated measures data and evaluated its performance using real-world prescription drug data of a hospital. As a result, the proposed method showed superior performance to existing methods.

Key words: prescription drug data, topic model, repeated measures data

キーワード：処方データ，トピックモデル，繰り返し測定データ

1. 序

トピックモデルは，主に自然言語処理分野において，文書からいくつかの潜在的なトピックを抽出するモデルである (Blei, Ng & Jordan, 2003). トピックを抽出することにより，文書を分類することやその著者を推定することが容易となる．統計学的にはトピックモデルはディリクレ分布を事前分布として潜在変数を仮定した階層ベイズモデルの一種である．各文書が各トピックをどの程度の割合で保有するかを表すトピック分布と，各トピック内での各単語の構成割合を表す単語分布によりモデルは主に構成される．単語分布はトピックごとに異なるとともに，各文書はそれぞれ複数のトピックを有し，そのトピックの構成割合 (トピック分布) に応じて各文書の単語が決定されていると考える．トピックモデルは統計学的には主成分分析などと同様に次元縮小法の一つに分類することが

可能であるが，他の手法と比較した場合のメリットとして，モデルの拡張性が高いのが特徴である．トピックモデルに他の共変量を組み込むことや，時点間でのトピック分布，単語分布の関連をモデル化して時系列データに対応させるなど，多くの拡張法が提案されている (Blei & Lafferty, 2006; Blei & Lafferty, 2007; Blei & McAuliffe, 2007; Mimno & McCallum, 2008; Rosen-Zvi, Griffiths, Steyvers & Smith, 2004).

トピックモデルは文書データにとどまらず，多変数のカウントデータで表すことが可能なデータ一般に適用可能であり，マーケティングデータや薬剤の処方データ，ゲノム・細菌データなど応用事例が近年，急速に増えている (Yan, Chuai, Qi, Shao, Zhou, Zhu, Yang, Yu, Shi, Kang, He & Liu, 2017; Sankaran & Holmes, 2018; La Rosa, Fiannaca, Rizzo & Urso, 2015; 李・照井, 2018; Reisenbichler & Reutterer, 2019; Park, Choi, Lim, Cha, Kim & Moon, 2017; Gaut, Steyvers, Imel, Atkins & Smyth, 2017; Kuhn, 2018). そのうち，マーケティングや医学データでは各対象者で複数回測定を行うことが多くある．特に医学分野において，各対象者で複数回測定を行ったデー

タを経時データと呼ぶことが多くあり、変量効果モデルなど同一の対象者のデータ間の関連を考慮する手法が発展している。一方で、トピックモデルについては繰り返し測定データに対する検討事例が依然として少ない。既存手法としては、マーケティング分野において、顧客の購買内容の変化をトピックモデルによりモデル化したトピック追跡モデル (Iwata, Watanabe, Yamada & Ueda, 2009) や、細菌データ解析において経時データに対するトピックモデルが提案されている (奥井・中路, 2018)。これら手法は同一対象者内での時点間でのトピック分布の関連を考慮可能であるという利点を持つ一方で、各対象者について同一の時点に同一回数データが測定されていることを前提としたモデルとなっている。一方で、現実のマーケティングや医学データにおいては、各対象者について必ずしも同一時点においてデータが測定されているとは限らないと同時に、データの測定回数は各対象者によって異なる場合が多い。

トピックモデルが応用される医学データのうち、処方データに対する応用例が増えてきている (Park et al., 2017; Gaut et al., 2017; Zafari & Ekin, 2019)。処方データとは、病院や薬局において各医薬品がどの程度処方されているかについてのデータであり、患者に投与されている薬剤を分析する、あるいは医師の処方行動を分析するなど様々な用途で分析がなされている。処方データについて、処方を行う医師を基準にデータを集計する場合や、各患者を基準とする場合など場合により異なるが、データ自体は一般的に処方数を医薬品や診療科ごとにカウントした多変量のカウントデータであることが多い。そのため、トピックモデルを用いて医薬品をグループ分けすることが可能となる。処方される医薬品は一般に時間とともに変化するため、トピック分布あるいは処方薬分布 (テキストデータにおける単語分布) を考慮したモデリングが結果の解釈時に有効である可能性がある。そこで本研究では、既存の経時データに対するトピックモデルを拡張した、繰り返し測定データ一般に適用可能なトピックモデルを提案するとともに、近年トピックモデルの応用事例が増えてきている薬剤処方データに適用し、既存手法との性能を比較するとともに、実データへの適用事例を示す。

2. 繰り返し測定データに対するトピックモデル

2.1. モデル

図1に繰り返し測定データに対するトピックモデルのグラフィカルモデルを示す。なお、本研究では処方データを扱うため、自然言語処理において単語に相当するものは処方薬、文書に対応するのは対象者となる。

ここで、 z_{ijn} は対象者 i ($i = 1, \dots, N$) の j ($j = 1, \dots, T_i$) 回目の測定時点における n ($n = 1, \dots, N_{ij}$) 番目の処方に関するトピック、 w_{ijn} は対象者 i の j 時点目における n 番目の処方薬、 θ_{ij} は対象者 i の j 時点目におけるトピック分布、 ϕ_k はトピック k の処方薬分布、 β は ϕ_k のハイパーパラメータを表す。 α_{ij} は α_{1ij} , α_{2ij} の2つのパラメータにより構成されるハイパーパラメータベクトルであり、 α_{1ij} は $\theta_{i,j-1}$ と θ_{ij} の関連度合いを示し、 α_{2ij} は $\theta_{i,j+1}$ と θ_{ij} の関連度合いを表すパラメータであるとする。 N は対象者数、 K はトピック数、 N_{ij} は各対象者 i の j 時点目における総処方数を表す記号とする。図1の枠内下部の記号は、枠内の各パラメータが合計いくつ存在するかを表している。 z_{ijn} , w_{ijn} については1処方ごと、 θ_{ij} は各対象者の測定時点ごと、 ϕ_k はトピックごとに定義される。 z_{ijn} は実現値として $1, \dots, K$ の値をとり、 w_{ijn} は実現値として $1, \dots, V$ の値をととする。また、トピック分布 θ_{ij} は $K \times 1$ のベクトル、処方薬分布 ϕ_k と β は $V \times 1$ のベクトルとする。

図2に単回測定者に関するトピックモデルのグラフィカルモデルを示す。測定が1回のみの単回測定者については、後述するように通常のトピックモデル (岩田, 2015) と同様のモデルとなる。 θ_{i1} のハイパーパラメータとして η が設定される。 η は $K \times 1$ のベクトルである。

図1のモデルにおけるデータ生成メカニズムは以下のようになる。

- (1) 各トピック k について、処方薬分布のサンプリング $\phi_k \sim \text{Dirichlet}(\beta)$
- (2) 各対象者 i について、トピック分布をサンプリング
 - (a) 測定回数が1回の対象者について、 $\theta_{i1} \sim \text{Dirichlet}(\eta)$
 - (b) 測定回数が2回以上を対象者について、
 - i. 初回時点 ($j = 1$) のトピック分布のサンプリング、 $\theta_{i1} \sim \text{Dirichlet}(\alpha_{2i1}\theta_{i2})$

図 1. 複数回測定者のグラフィカルモデル

図 2. 単回測定者のグラフィカルモデル

- ii. 中間時点 ($1 < j < T_i$) のトピック分布のサンプリング,
 $\theta_{ij} \sim \text{Dirichlet}(\alpha_{1ij}\theta_{i,j-1} + \alpha_{2ij}\theta_{i,j+1})$
- iii. 最終時点 ($j = T_i$) のトピック分布のサンプリング,
 $\theta_{iT_i} \sim \text{Dirichlet}(\alpha_{1iT_i}\theta_{i,T_i-1})$
- (3) 対象者 i の j 時点目における n 番目の処方薬
 - (a) トピックをサンプリング,
 $z_{ijn} \sim \text{Multinomial}(1, \theta_{ij})$
 - (b) 処方薬をサンプリング,
 $w_{ijn} \sim \text{Multinomial}(1, \phi_{z_{ijn}})$

処方薬分布のサンプリングに関しては通常のトピックモデルにおける単語分布の推定と同様である。トピック分布の推定に関しては単回測定者と複数回測定で異なり、単回測定者については、通常のトピックモデル同様に全対象者で共通であるトピック分布のハイパーパラメータ η をもとにトピック分布が生成される。複数回測定者については測定時点ごとにサンプリング方法が異なり、初回時点については2時点目のトピック分布をもとに、最終時点については1つ前の時点のトピック分布をもとにトピック分布が生成される。初回時点と最終時点以外の中間時点においては、当該時点の一つ前と一つ後のトピック分布をもとにトピックが生成される。

2.2. 尤度関数

トピックモデルの尤度は、処方薬分布とトピック分布、及びそれらの事前分布の積で表現できる。単回測定者のトピックモデルの尤度関数 L は、通常のトピックモデル (佐藤, 2015) と同様に以下の式で表現できる。

$$\begin{aligned}
 L &= \iint \prod_{i=1}^N \prod_{n=1}^{N_{i1}} \\
 &\quad \times \{p(w_{i1n}, z_{i1n}, \theta_{i1}, \phi_k \mid \eta, \beta)\}^{\delta(T_i=1)} d\theta_{i1} d\phi_k \\
 &= \iint \left[\prod_{i=1}^N \prod_{n=1}^{N_{i1}} \{p(w_{i1n} \mid \phi_{z_{i1n}}) p(z_{i1n} \mid \theta_{i1})\}^{\delta(T_i=1)} \right] \\
 &\quad \times \left[\prod_{k=1}^K p(\phi_k \mid \beta) \right] \\
 &\quad \times \left[\prod_{i=1}^N p(\theta_{i1} \mid \eta)^{\delta(T_i=1)} \right] d\theta_{i1} d\phi_k \\
 &= \int \prod_{k=1}^K \left[\prod_{v=1}^V \phi_{kv}^{N_{kv}} \right] \frac{\Gamma(\sum_{v=1}^V \beta_v)}{\prod_{v=1}^V \Gamma(\beta_v)} \prod_{v=1}^V \phi_{kv}^{\beta_v-1} d\phi_k \\
 &\quad \times \left[\int \prod_{i=1}^N \left\{ \prod_{k=1}^K \theta_{i1k}^{N_{i1k}} \right\}^{\delta(T_i=1)} \right. \\
 &\quad \times \left. \frac{\Gamma(\sum_{k=1}^K \eta_k)}{\prod_{k=1}^K \Gamma(\eta_k)} \prod_{k=1}^K \theta_{i1k}^{\eta_k-1} d\theta_{i1} \right] \\
 &= \prod_{k=1}^K \frac{\Gamma(\sum_{v=1}^V \beta_v)}{\prod_{v=1}^V \Gamma(\beta_v)} \frac{\prod_{v=1}^V \Gamma(N_{kv} + \beta_v)}{\Gamma(\sum_{v=1}^V N_{kv} + \beta_v)} \\
 &\quad \times \prod_{i=1}^N \left\{ \frac{\Gamma(\sum_{k=1}^K \eta_k)}{\prod_{k=1}^K \Gamma(\eta_k)} \frac{\prod_{k=1}^K \Gamma(N_{ijk} + \eta_k)}{\Gamma(\sum_{k=1}^K N_{ijk} + \eta_k)} \right\}^{\delta(T_i=1)} \quad (1)
 \end{aligned}$$

ただし、 δ は指示関数を表す。複数回測定者については測定時点ごとにトピック分布の事前分布が異なるため以下の式で表現できる。

$$\begin{aligned}
 L &= \iint \prod_{i=1}^N \prod_{j=1}^{T_i} \prod_{n=1}^{N_{ij}} \\
 &\quad p(w_{ijn}, z_{ijn}, \theta_{ij}, \phi_k \mid \theta_{i,j-1}, \theta_{i,j+1}, \alpha_{ij}, \beta)^{\delta(T_i > 1)} \\
 &\quad d\theta_{ij} d\phi_k
 \end{aligned}$$

$$\begin{aligned}
 &= \prod_{k=1}^K \frac{\Gamma(\sum_{v=1}^V \beta_v)}{\prod_{v=1}^V \Gamma(\beta_v)} \frac{\prod_{v=1}^V \Gamma(N_{kv} + \beta_v)}{\Gamma(\sum_{v=1}^V N_{kv} + \beta_v)} \\
 &\quad \times \prod_{i=1}^N \prod_{j=1}^{T_i} \left\{ \frac{\Gamma(\sum_{k=1}^K \alpha_{2ij} \theta_{i,j+1,k})}{\prod_{k=1}^K \Gamma(\alpha_{2ij} \theta_{i,j+1,k})} \right. \\
 &\quad \times \left. \frac{\prod_{k=1}^K \Gamma(N_{ijk} + \alpha_{2ij} \theta_{i,j+1,k})}{\Gamma(\sum_{k=1}^K N_{ijk} + \alpha_{2ij} \theta_{i,j+1,k})} \right\}^{\delta(j=1)} \\
 &\quad \times \left\{ \frac{\Gamma(\sum_{k=1}^K \alpha_{1ij} \theta_{i,j-1,k} + \alpha_{2ij} \theta_{i,j+1,k})}{\prod_{k=1}^K \Gamma(\alpha_{1ij} \theta_{i,j-1,k} + \alpha_{2ij} \theta_{i,j+1,k})} \right. \\
 &\quad \times \left. \frac{\prod_{k=1}^K \Gamma(N_{ijk} + \alpha_{1ij} \theta_{i,j-1,k} + \alpha_{2ij} \theta_{i,j+1,k})}{\Gamma(\sum_{k=1}^K N_{ijk} + \alpha_{1ij} \theta_{i,j-1,k} + \alpha_{2ij} \theta_{i,j+1,k})} \right\}^{\delta(1 < j < T_i)} \\
 &\quad \times \left\{ \frac{\Gamma(\sum_{k=1}^K \alpha_{1ij} \theta_{i,j-1,k})}{\prod_{k=1}^K \Gamma(\alpha_{1ij} \theta_{i,j-1,k})} \right. \\
 &\quad \times \left. \frac{\prod_{k=1}^K \Gamma(N_{ijk} + \alpha_{1ij} \theta_{i,j-1,k})}{\Gamma(\sum_{k=1}^K N_{ijk} + \alpha_{1ij} \theta_{i,j-1,k})} \right\}^{\delta(j=T_i)} \Bigg]^{\delta(T_i > 1)} \quad (2)
 \end{aligned}$$

2.3. ギブスサンプリング式

パラメータの推定方法について経時測定データに対するトピックモデルについて検討した先行研究同様 (Iwata et al., 2009), 周辺化ギブスサンプリングを用いた。周辺化ギブスサンプリングとは、階層ベイズモデルにおいて、階層の中間にあるパラメータを周辺化したサンプリング式を利用することにより、パラメータの推定をより効率的に行う手法のことである。本研究での周辺化ギブスサンプリングで z_{ijn} とハイパーパラメータを直接関連付けた式を用い、 θ_{ij}, ϕ_k については z_{ijn} とハイパーパラメータをもとに推定を行う。

各トピックのサンプリング確率は、処方薬分布とトピック分布の積で表される。トピック z_{ijn} について、各トピック k がサンプリングされる確率は複数回測定者については以下ようになる。ただし、 N_{ijk} は各対象者 i の j 時点目においてトピック k に割り当てられた処方の合計数とする。 N_k はトピック k に割り当てられた処方数であり、 $N_{kw_{ijn}}$ は処方薬 w_{ijn} のうちトピック k に割り当てられた処方数を示す。 $\mathbf{W}$ は全処方のデータを表し、 $\mathbf{Z}$ は全処方に対応する全トピックを表すものとする。なお、 $\backslash ijn$ は ijn を除いてという意味である。

$$\begin{aligned}
& p(z_{ijn} = k \mid \mathbf{W}, \mathbf{Z}_{\setminus ijn}, \boldsymbol{\theta}_{i,j-1}, \boldsymbol{\theta}_{i,j+1}, \boldsymbol{\alpha}_{ij}, \boldsymbol{\phi}_k, \boldsymbol{\beta}) \\
& \propto p(z_{ijn} = k \mid \mathbf{Z}_{\setminus ijn}, \boldsymbol{\theta}_{i,j-1}, \boldsymbol{\theta}_{i,j+1}, \boldsymbol{\alpha}_{ij}) \\
& \quad \times p(w_{ijn} \mid \mathbf{W}_{\setminus ijn}, \mathbf{Z}_{\setminus ijn}, z_{ijn} = k, \boldsymbol{\beta}) \\
& \propto p(z_{ijn} = k \mid \mathbf{Z}_{\setminus ijn}, \boldsymbol{\theta}_{i,j-1}, \boldsymbol{\theta}_{i,j+1}, \boldsymbol{\alpha}_{ij}) \\
& \quad \times \frac{N_{kw_{ijn}} - 1 + \beta_{w_{ijn}}}{N_k - 1 + \sum_{v=1}^V \beta_v}
\end{aligned} \tag{3}$$

$p(z_{ijn} = k \mid \mathbf{Z}_{\setminus ijn}, \boldsymbol{\theta}_{i,j-1}, \boldsymbol{\theta}_{i,j+1}, \boldsymbol{\alpha}_{ij})$ は、測定時点ごとに値が異なり、初回時点の場合には以下の式となる。

$$\frac{N_{i1k} - 1 + \alpha_{2i1}\theta_{i2k}}{N_{i1} - 1 + \alpha_{2i1}} \tag{4}$$

中間時点の場合には以下の式となる。

$$\frac{N_{ijk} - 1 + \alpha_{1ij}\theta_{i,j-1,k} + \alpha_{2ij}\theta_{i,j+1,k}}{N_{ij} - 1 + \alpha_{1ij} + \alpha_{2ij}} \tag{5}$$

最終時点の場合には以下の式となる。

$$\frac{N_{iT_i k} - 1 + \alpha_{1iT_i}\theta_{i,T_i-1,k}}{N_{iT_i} - 1 + \alpha_{1iT_i}} \tag{6}$$

単回測定者については、 $p(z_{ijn} = k \mid \mathbf{Z}_{\setminus ijn}, \boldsymbol{\theta}_{i,j-1}, \boldsymbol{\theta}_{i,j+1}, \boldsymbol{\alpha}_{ij})$ ではなく通常のトピックモデルと同様 (岩田, 2015), 以下の式となる。

$$p(z_{i1n} = k \mid \mathbf{Z}_{\setminus i1n}, \boldsymbol{\eta}) = \frac{N_{i1k} - 1 + \eta_k}{N_{i1} - 1 + \sum_{k=1}^K \eta_k} \tag{7}$$

2.4. 各パラメータの推定式

トピック分布、処方薬分布のハイパーパラメータについて、本研究では経験ベイズ推定した。複数回測定の対象者について、 $\alpha_{1ij}, \alpha_{2ij}$ はそれぞれ以下の式により更新した (奥井他, 2018)。ここで、 ψ はディガンマ関数を表す。

$$\begin{aligned}
& \alpha_{1ij} = \alpha_{1ij} \times \\
& \frac{\sum_{k=1}^K \theta_{i,j-1,k} (\psi(N_{ijk} + \alpha_{1ij}\theta_{i,j-1,k}) - \psi(\alpha_{1ij}\theta_{i,j-1,k}))}{\psi(N_{ij} + \alpha_{1ij}) - \psi(\alpha_{1ij})}
\end{aligned} \tag{8}$$

$$\begin{aligned}
& \alpha_{2ij} = \alpha_{2ij} \times \\
& \frac{\sum_{k=1}^K \theta_{i,j+1,k} (\psi(N_{ijk} + \alpha_{2ij}\theta_{i,j+1,k}) - \psi(\alpha_{2ij}\theta_{i,j+1,k}))}{\psi(N_{ij} + \alpha_{2ij}) - \psi(\alpha_{2ij})}
\end{aligned} \tag{9}$$

単回測定の対象者についてはトピック分布のハイパーパラメータ $\boldsymbol{\eta}$ については通常のトピックモデルと同様の経験ベイズ推定式を用いた (岩田, 2015)。

$$\eta_k = \eta_k \times \frac{\sum_{i=1}^N (\psi(N_{i1k} + \eta_k) - \psi(\eta_k))}{\sum_{i=1}^N (\psi(N_{i1} + \sum_{k=1}^K \eta_k) - \psi(\sum_{k=1}^K \eta_k))} \tag{10}$$

$\boldsymbol{\beta}$ についても同様に、以下の式で値を更新した (岩田, 2015)。

$$\beta_v = \beta_v \times \frac{\sum_{k=1}^K (\psi(N_{kv} + \beta_v) - K\psi(\beta_v))}{\sum_{k=1}^K (\psi(N_k + \sum_{v=1}^V \beta_v) - K\psi(\sum_{v=1}^V \beta_v))} \tag{11}$$

トピック分布については、複数回測定者ではギブスサンプリングの式と同様に以下の式で更新した。初回時点の場合には、

$$\theta_{i1k} = \frac{N_{i1k} + \alpha_{2i1}\theta_{i2k}}{N_{i1} + \alpha_{2i1}} \tag{12}$$

中間時点の場合には以下の式となる。

$$\theta_{ijk} = \frac{N_{ijk} + \alpha_{1ij}\theta_{i,j-1,k} + \alpha_{2ij}\theta_{i,j+1,k}}{N_{ij} + \alpha_{1ij} + \alpha_{2ij}} \tag{13}$$

最終時点の場合には以下の式となる。

$$\theta_{iT_i k} = \frac{N_{iT_i k} + \alpha_{1iT_i}\theta_{i,T_i-1,k}}{N_{iT_i} + \alpha_{1iT_i}} \tag{14}$$

単回測定者については以下の式となる。

$$\theta_{i1k} = \frac{N_{i1k} + \eta_k}{N_{i1} + \sum_{k=1}^K \eta_k} \tag{15}$$

一方、処方薬分布は以下の式で推定できる (岩田, 2015)。

$$\phi_{kv} = \frac{N_{kv} + \beta_v}{N_k + \sum_{v=1}^V \beta_v} \tag{16}$$

2.5. 推定手順

最後にモデルの推定手順を以下に示す。

- (1) $N_{ijk} = 0, N_{kv} = 0, N_k = 0, z_{ijn} = 0$ とする。
- (2) 全対象者の全処方薬について 1 処方ごとに以下の工程を行う。

- (a) $z_{ijn} = 0$ でない場合、 $N_{ijz_{ijn}} = N_{ijz_{ijn}} - 1, N_{z_{ijn}w_{ijn}} = N_{z_{ijn}w_{ijn}} - 1, N_{z_{ijn}} = N_{z_{ijn}} - 1$ とする。
- (b) 各トピック k のサンプリング確率 $p(z_{ijn} = k)$ を計算する。

- (c) $z_{ijn} \sim \text{Multinomial}(1, p(z_{ijn} = k))$ によりトピックをサンプリングする.
- (d) $N_{ijz_{ijn}} = N_{ijz_{ijn}} + 1$, $N_{z_{ijn}w_{ijn}} = N_{z_{ijn}w_{ijn}} + 1$, $N_{z_{ijn}} = N_{z_{ijn}} + 1$ とする.
- (3) ハイパーパラメータを更新する.
- (4) トピック分布, 処方薬分布を推定する.
- (5) (2), (3), (4) を, ベイズシミュレーション回数分繰り返す.

3. 実データによる評価

実データに提案法を適用し, その性能を評価した. 性能評価及び実データ解析において漢方製剤のデータを用い, 各診療科の漢方薬の処方傾向をトピックモデルをもとに分析する. データとして, 九州大学病院において漢方製剤及び漢方生薬を処方された 20 歳以上の患者を対象とした. データは電子カルテデータシステムから抽出した. 2008 年から 2018 年までのデータを使用した. 処方される医薬品の動向は時間とともに変化するため, 診療科ごとに年単位でデータを集約し, 1 年間の間に各診療科において処方された漢方製剤に関するカウントデータを作成した. ただし, 処方件数が 10 に満たない年のデータ, 処方件数の合計が 100 件に満たない薬剤のデータについては情報量が極めて小さいため解析から除外した. また, 診療科を単位に分析するため対象者ではなく, ここでは診療科と表記する. なお, 各漢方製剤についてグループ分けするため, KEGG: Kyoto Encyclopedia of Genes and Genomes の「医療用医薬品の薬効分類 + 商品 (YJ コード), 5 生薬及び漢方処方に基づく医薬品, 5200 漢方製剤」(KEGG) に従って分類を行った.

性能を比較する手法として構造トピックモデル (Roberts, Stewart & Tingley, 2018) と動的トピックモデル (Blei & Lafferty, 2006) を用いた. 提案法は, 同一対象者間のトピック分布の関連性を考慮する手法であるが, 同一対象者間の関連性を考慮する他の方法として, トピック分布のハイパーパラメータを対象者ごとに設定する方法が考えられる. この場合, 対象者の情報はダミー変数化することで共変量として扱うことができるため, 構造トピックモデルによるモデル化が可能である. 構造トピックモデルは補助情報をトピックモデルの推定に組み込むことができるモデルであり, 処方データに対しても診療科の情報を考慮して医

師の処方行動を分析する手法として用いられている (Zafari et al., 2019). また, 今回使用するデータは時系列データであるため, 時系列データに対するトピックモデルとして広く知られている動的トピックモデルを適用した. 動的トピックモデルは時点ごとに処方薬分布 (単語分布) とトピック分布それぞれのハイパーパラメータを設定し, 時点間でのパラメータの関連性をモデル化したモデルである. 処方薬分布やトピック分布が時点により大きく影響される場合には有効な手法となる.

トピックモデルの評価方法として, 実際のデータにモデルを適用し, モデルの予測性能を評価する perplexity が一般的であり (Blei, Ng & Jordan, 2003; 佐藤, 2015), 評価方法として用いた. データの総処方数の半分の処方をランダムにサンプリングし, サンプリングしたデータを訓練データ, 残りを検証データとした. 訓練データにおいてモデルを学習し, 検証データにおいて以下の指標を算出した.

$$\text{Perplexity} = \exp\left\{-\frac{\sum_{i=1}^N \sum_{j=1}^{T_i} \sum_{n=1}^{N_{ij}} \log p(w_{ijn})}{\sum_{i=1}^N \sum_{j=1}^{T_i} N_{ij}}\right\},$$

$$p(w_{ijn}) = \hat{\theta}_{ijz_{ijn}} \hat{\phi}_{z_{ijn}w_{ijn}} \quad (17)$$

ここで, $\hat{\theta}_{ijz_{ijn}}$ は診療科 i の j 時点目におけるトピック分布においてトピック z_{ijn} が占める割合に関する訓練データでの推定値であり, $\hat{\phi}_{z_{ijn}w_{ijn}}$ はトピック z_{ijn} の処方薬分布内での処方薬 w_{ijn} の割合に関する推定値を表す. Perplexity は訓練データにおけるトピック分布の推定値と処方薬分布の推定値を検証データに適用して検証データにおける尤度を計算しており, 値が小さいほど検証データの予測性能がよいことを示している. Perplexity の値を算出する際, 訓練データで推定した処方薬分布をもとに検証データにおけるトピック分布を推定した. なお, 動的トピックモデルでは時点ごとに処方薬分布が設定されるため, 診療科 i の j 回目の測定時点における時点 $t_{ij}(1, \dots, T)$ としたとき $p(w_{ijn})$ は以下の式となる.

$$p(w_{ijn}) = \hat{\theta}_{ijz_{ijn}} \hat{\phi}_{t_{ij}z_{ijn}w_{ijn}} \quad (18)$$

本研究ではベイズ推定におけるシミュレーション回数を 2000 回として最初の 1000 回は burn-in 期間として, 残りの 1000 回のシミュレーションでの平均を取り θ_{ij} , ϕ_k を推定した. なお, 本研究のすべてのデータ分析は R3.5.1 を用いて行った (R core team., 2017).

表 1. Perplexity の結果

モデル	トピック数				
	10	20	30	40	50
構造トピックモデル	18.92	18.44	19.05	18.49	17.02
動的トピックモデル	16.75	15.04	14.53	14.13	13.99
提案法	14.53	13.74	13.52	13.46	13.49

構造トピックモデルは R パッケージ `stm` (Roberts et al., 2018) で実行可能であるが、処方薬分布の数値情報が得られないため、R パッケージ `rstan` を用いてコードを作成した。動的トピックモデルについても `rstan` を用いてコードを作成した。提案法と Perplexity の算出には R パッケージ `Rcpp` (Eddelbuettel & Francois, 2011) を用いてコードを作成し実行した。

表 1 が提案法とトピックモデルを評価した結果である。トピック数は 10 から 50 まで変化させて計算している。提案法は構造トピックモデルと比較し Perplexity が低くなることが確認された。

4. 実データ解析

本研究では性能評価同様、漢方製剤の処方データに対して提案法を適用した。使用データは前節と同様である。漢方製剤は複数の生薬を混合した製剤であることから、複数の漢方製剤を重複して服用した場合、特定の生薬を過剰に摂取する恐れがある。一方で、その処方実態や診療科間での処方傾向が大規模データを用いて分析されているわけではない。そこで、実データ解析の事例として各診療科で処方されている漢方製剤のデータをトピックモデルをもとに要約した。

診療科、処方薬の基本属性が表 2 である。ただし、ここでの観測値とは各診療科の各年を指す。図 3 が、全体での処方件数が最も高い上位 10 について、全処方に占める処方割合を診療科ごとに折れ線グラフで示した結果である。各診療科において、各医薬品の処方割合は年によって変化することがわかる。そのため、繰り返し測定を考慮したトピックモデルを適用する意義があると考えられる。

表 3 が抽出されたトピックの処方薬分布であり、表 4 が抽出されたトピックのトピック分布の値の要約統計量である。表 4 では各トピックについて、処方薬分布値が大きい順に処方薬分布の値が 5% 以上である

表 2. 各診療科の基本属性

項目	値
観測値数*	345
診療科数*	35
処方薬数*	90
各診療科の時点数†	9.86(2.48)
各観測値の処方薬の種類数†	17.76(16.04)
各観測値の処方回数†	768.59(1192.29)

* 件数

† 平均 (標準偏差)

処方薬までを示している。それぞれ内容が異なる 10 個のトピックが得られた。

図 4 が各診療科におけるトピック分布の経時推移をトピックごとに示した結果である。図 3 の医薬品の処方件数と同様、各トピックのトピック割合は年を通して一定であるわけではなく、診療科ごとに年とともに変化していることがわかる。

表 5 が各診療科のトピック分布である。トピック割合が高い上位 3 つのトピック番号とそのトピック分布のみ示している。各診療科のトピック分布は年ごとに存在するが、ここではその平均値をもとにした結果を示している。処方件数が最も多い大建中湯、六君子湯で構成される 9 番目のトピックが各診療科で上位を占めることが多かったが、診療科ごとに処方傾向が異なっていることが確認できる。また、1 つのトピックが処方の大半を占める診療科と様々なトピックが存在する診療科があり、診療科により用いる漢方製剤の種類や使用目的も異なることが示唆された。

5. 考 察

本研究では繰り返し測定データ一般に拡張可能なトピックモデルを定式化し、その性能を確認した。性能評価の結果、提案法は構造トピックモデル、動的トピックモデルを上回る予測性能を示した。構造トピックモデルでは対象者に関する情報をトピック分布の事前分布と関連づけることが可能であるが、提案法では同一対象者内のトピック分布間の関連性をモデル化しているため、より同一対象者内でのトピック分布の相関を考慮可能である可能性がある。

動的トピックモデルは、処方薬分布 (単語分布) とトピック分布のハイパーパラメータの時間変化をモデル

図 3. 処方数が多い医薬品における処方割合の経時推移

表 3. 各トピックの処方薬分布

トピック番号	処方薬分布
1	桂枝茯苓丸, 抑肝散, 補中益気湯, 温清飲, 大建中湯
2	六君子湯, 補中益気湯, 半夏厚朴湯, 当帰芍薬散, 加味逍遙散
3	十全大補湯, 半夏瀉心湯, 大建中湯, 当帰芍薬散
4	芍薬甘草湯, 抑肝散, 五苓散
5	抑肝散, 補中益気湯, 大建中湯, 半夏厚朴湯, 加味逍遙散, 六君子湯, 防風通聖散, 十全大補湯
6	葛根湯, 大建中湯, 小青竜湯, 牛車腎気丸
7	加味逍遙散, 抑肝散加陳皮半夏, 人參養栄湯, 五苓散, 苓桂朮甘湯, 大建中湯
8	桂枝茯苓丸, 人參湯, 真武湯
9	大建中湯, 六君子湯
10	牛車腎気丸, 加味逍遙散, 桂枝茯苓丸, 当帰芍薬散, 葛根湯

表 4. 各トピックのトピック分布の要約統計量

トピック番号	最小値	25% 点	中央値	平均値	75% 点	最大値
1	0.00	0.00	0.00	0.03	0.00	0.80
2	0.00	0.00	0.04	0.10	0.11	0.93
3	0.00	0.00	0.01	0.05	0.05	0.74
4	0.00	0.00	0.02	0.10	0.12	0.52
5	0.00	0.00	0.01	0.08	0.09	0.88
6	0.00	0.00	0.10	0.18	0.25	0.94
7	0.00	0.00	0.00	0.08	0.01	1.00
8	0.00	0.00	0.00	0.04	0.00	0.80
9	0.00	0.00	0.19	0.27	0.45	0.96
10	0.00	0.00	0.03	0.09	0.13	0.62

化したモデルであるが、同一対象者間のトピック分布の関連をモデル化しているわけではない。そのため、各対象者のカウント値が測定時点に大きく影響を受ける場合には有効であると考えられるが、その他の場合には同一対象者間のトピック分布の関連をモデル化したほうが各対象者のトピック分布を正確に推定できると考えられる。今回適用した処方薬のデータでは、各診療科のトピック分布は時間とともに変化するものの、年の変化に伴う処方トレンドの変化よりも診療科ごとの処方傾向の違いが強いため予測性能も提案法が上回ったと考えられる。ただ、データの種類や規模によって他のモデルとの評価結果は異なる可能性もあるため、今後マーケティングデータなど他種類のデータを用いたさらなる検証が必要であると考えられる。

なお、時系列データに対する代表的なトピックモデルである動的トピックモデルは単語分布の時間変化を

考慮しており、単語分布の変化に着目したい場合に有効な手法である。各対象者の測定時点がそれぞれ異なるデータに適用する場合、データに存在する時点の数だけ単語分布を作成して、それら複数の単語分布の時系列変化をモデル化することになるためモデルの解釈も困難になると考えられる。繰り返し測定データ及び、時系列データにトピックモデルを適用する際、提案法と動的トピックモデルは使用目的に応じて使い分ける必要がある。

今後の課題として、トピックモデルの性質上、設定するトピック数により得られるトピックの内容は変わりうる。トピック数の設定の問題に対処する方法として、階層ノンパラメトリックベイズモデルを用いたモデリングも複数提案されているが、提案法についてもノンパラメトリックベイズ推定が可能であるか検討する意義があると考えられる。

図 4. 各診療科におけるトピック分布の経時推移

表 5. 各診療科における上位 3 トピックのトピック番号とトピック割合

診療科	トピック番号（トピック割合）		
	1 位	2 位	3 位
第一外科	9(0.73)	10(0.07)	6(0.05)
顎口腔外科	7(1.00)	9(0.00)	1(0.00)
肝臓・膵臓・胆道内科	9(0.41)	6(0.26)	4(0.15)
眼科	6(0.47)	5(0.20)	10(0.13)
顔面口腔外科	7(0.78)	2(0.22)	5(0.00)
救命救急センター	9(0.86)	4(0.09)	5(0.05)
形成外科	4(0.52)	2(0.17)	7(0.16)
血液・腫瘍・心血管内科	6(0.31)	9(0.30)	10(0.20)
呼吸器科	6(0.34)	2(0.22)	9(0.15)
口腔総合診療科	7(1.00)	3(0.00)	5(0.00)
歯科麻酔科	7(0.90)	10(0.03)	9(0.02)
耳鼻咽喉・頭頸部外科	3(0.43)	6(0.19)	2(0.19)
循環器内科	9(0.39)	6(0.35)	4(0.19)
小児科	6(0.37)	2(0.31)	3(0.13)
小児外科, 成育外科, 小腸移植外科	9(0.57)	8(0.23)	2(0.11)
消化管内科	9(0.47)	2(0.26)	6(0.15)
心臓血管外科	9(0.59)	6(0.14)	10(0.11)
心療内科	2(0.68)	5(0.16)	7(0.10)
脳神経内科	5(0.45)	4(0.17)	9(0.13)
腎・高血圧・脳血管内科	6(0.48)	9(0.22)	4(0.21)
整形外科	6(0.26)	4(0.21)	10(0.18)
精神科神経科	5(0.65)	6(0.11)	9(0.10)
先端医工学診療部	9(0.52)	5(0.38)	7(0.10)
先端分子・細胞治療科	9(0.28)	4(0.19)	3(0.19)
総合診療科	8(0.77)	2(0.09)	10(0.04)
内分泌代謝・糖尿病内科	6(0.41)	4(0.24)	9(0.17)
第二外科	9(0.77)	10(0.10)	4(0.05)
脳神経外科	9(0.25)	4(0.23)	6(0.22)
泌尿器・前立腺・腎臓・副腎外科	9(0.56)	10(0.26)	5(0.07)
皮膚科	1(0.66)	5(0.10)	6(0.08)
産科婦人科（周産母子）	9(0.47)	2(0.29)	6(0.23)
産科婦人科	10(0.45)	9(0.40)	3(0.09)
放射線科	9(0.24)	2(0.19)	10(0.17)
麻酔科蘇生科	6(0.61)	10(0.17)	4(0.15)
免疫・膠原病・感染症内科	6(0.31)	10(0.25)	9(0.16)

参 考 文 献

- Blei, D.M., Ng, A.Y. & Jordan, M.I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, **3**, 993–1022.
- Blei, D.M. & Lafferty, J.D. (2006). Dynamic topic models. *Proceedings of the 23rd international conference on machine learning*, 113–120.

謝 辞

的確なご指摘をいただいた査読者の先生方にお礼申し上げます。

- Blei, D.M. & Lafferty, J.D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, **1**, 17–35.
- Blei, D.M. & McAuliffe, J.D. (2007). Supervised topic models. In Platt, J.C., Koller, D., Singer, Y., & Roweis, S., editors, *Advances in Neural Information Processing Systems (NIPS)*, 121–128, Cambridge, MA. MIT Press.
- Eddelbuettel, D. & Francois, R. (2011). Rcpp: Seamless R and C++ Integration. *Journal of Statistical Software*, **40**, 1–18. URL <http://www.jstatsoft.org/v40/i08/>.
- Gaut, G., Steyvers, M., Imel, Z.E., Atkins, D.C. & Smyth, P. (2017). Content coding of psychotherapy transcripts using labeled topic models, *IEEE journal of Psychotherapy*, **21**, 476–487.
- 岩田具治 (2015). トピックモデル (機械学習プロフェッショナルシリーズ). 講談社.
- Iwata, T., Watanabe, S., Yamada, T. & 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.
- Kyoto Encyclopedia of Genes and Genomes, https://www.genome.jp/kegg/kegg_ja.html.
- Kuhn, K.D. (2018). Using structural topic modelling to identify latent topics and trends in aviation incident reports. *Transportation research, part C*, **87**, 105–122.
- Mimno, D. & McCallum, A. (2008). Topic models conditioned on arbitrary features with Dirichlet multinomial regression. In *Uncertainty in Artificial Intelligence*, Helsinki, 411–418.
- Rosen-Zvi, M., Griffiths, T., Steyvers, M. & Smyth, P. (2004). The Author-Topic Model for Authors and Documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*.
- 奥井佑・中路重之 (2018). 経時的に測定された腸内細菌組成データに対する潜在ディリクレ配分モデルによる関連解析法. 計量生物学会誌, **39**, 37–54.
- Park, S., Choi, D., Lim, M., Cha, W., Kim, C. & Moon, I. (2017). Identifying prescription patterns with a topic model of diseases and medications. *Journal of Biomedical Informatics*, **75**, 35–47.
- R Core Team (2017). R: a language and environment for statistical computing, R Foundation for Statistical Computing. Vienna, Austria. <http://www.R-project.org/>.
- Reisenbichler, M. & Reutterer, T. (2019). Topic modeling in marketing: recent advances and research opportunities. *Journal of Business Economics*, **89**, 327–356.
- 李銀星・照井伸彦 (2018). 大規模集計 POS データの高精度スパースモデリング, 統計数理, **66**, 235–247.
- Roberts, M.E., Stewart, B.M. & Tingley, D. (2018). stm: R Package for Structural Topic Models, URL <http://www.structuraltopicmodel.com>.
- La Rosa, M., Fiannaca, A., Rizzo, R. & Urso, A. (2015). Probabilistic topic modeling for the analysis and classification of genome sequences. *BMC Bioinformatics*, **16**, S2.
- Sankaran, K. & Holmes, S.P. (2018). Latent variable modeling for the microbiome. *Biostatistics*, <https://arxiv.org/pdf/1706.04969.pdf>.
- Stan Development Team. (2018). RStan: the R interface to Stan. R package version 2.17.3. <http://mc-stan.org>.
- 佐藤一誠 (2015). 自然言語処理シリーズ 8 トピックモデルによる統計的潜在意味解析. 講談社.
- Yan, J., Chuai, G., Qi, T., Shao, F., Zhou, C., Zhu, C., Yang, J., Yu, Y., Shi, C., Kang, N., He, Y. & Liu Q. (2017). MetaTopics: an integration tool to analyze microbial community profile by topic model. *BMC Genomics*, **18**, 962.
- Zafari, B. & 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.

(2019年6月30日受理, 2019年9月19日採択)

(この間審査3回・審査期間合計66日)