DDSTM:Spike and Slab 事前分布を用いた動的スパース・トピックモデル

増田 樹
慶應義塾大学大学院経済学研究科
t.masuda@keio.jp

中川 慧
野村アセットマネジメント株式会社
kei.nak.0315@gmail.com

星野 崇宏
慶應義塾大学経済学部 理化学研究所 AIP センター
bayesian@jasmine.ocn.ne.jp

概要

トピックモデルは文書生成をモデル化する自然言語処理手法の一つで，トピックと単語の分布に基づいて単語を生成する．既存の推定手法には plSA や LDA などがあるが，これらは時系列性を考慮しない．本研究では，時系列変化と分布のスパース性を同時に扱う新しいモデル，動的スパース・トピックモデル (DDSTM) を提案する．DDSTM は動的トピックモデル (DTM) を基にし，Spike and Slab Prior を事前分布として使用し，スパース性を保ちながら時系列変化をモデル化する．この手法により，トピックと単語の分布のスパース性と時系列変化を統合し，解釈性を高めることができる．実証分析では，実データを用いて提案モデルの特徴を確認する．

1 はじめに

トピックモデルとは自然言語処理における文書の確率的生成モデルの一つである．トピックモデルにおいて文書を構成する単語の生成過程はトピック分布にしたがってトピックを選択し，選んだトピックの持つ単語分布にしたがって生成される．ここでトピック分布とは各文書におけるトピックの条件付き分布，単語分布とは各トピックにおける単語の条件付き分布を意味する．文書がトピックモデルから生成されたと仮定した上で，実際に観測された文書から各トピック分布および単語分布を統計的に推定することで，文書に含まれる話題の比率や，話題を構成する単語の分布を知ることができる．トピックモデルの推定手法は多数提案されており，点推定手法の一つとして PLSA [1] があげられる．PLSA では，EM アルゴリズムを用いた推定を行うことで，点推定量を求めることができる．一方で，近年，トピックモデルのベイズ推定も提案されており，例えば Blei らによって提案された Latent Dirichlet Allocation (LDA [2]) があげられる．LDA は，事前分布としてディリクレ分布を仮定し，トピック分布と単語分布に多項分布を仮定することで，MCMC や変分推論を用いたベイズ推定が可能となる．

しかし，一般に LDA において，時系列性は考慮されないため，トピック分布や単語分布の時系列的な推移を追うことはできない．そこでトピックの時系列性を考慮する LDA である DynamicTopic Model(DTM [3]) が提案された．DTM では，データセットは指定された時間ごとに均等に分割され，文書のトピック分布のパラメータおよび各トピックの単語の分布は時間とともに変化する．

一方で，トピックモデルでは文書によっては数百のトピックが分布として仮定されることがあるが，実際の文章としてはその中から特定の重要なトピックによって表現されていると想定するのが妥当である．また単語分布についても同様に，多数の単語が含まれる単語分布のうち，トピックごとに特定の重要な単語で構成されていると想定するのが自然である．すなわち，トピック分布および単語分布にこのようなスパース性を導入することが重要である．この観点から，スパース性を考慮したトピックモデルである Dual Sparse Topic Model [4] や，Sparse Topical Coding [5] が提案されている．しかし，これらのモデルでは前述の時系列性は考慮されていない．

そこで，本研究では，これらの時系列性およびスパース性の問題に対応するために，動的スパース・トピックモデル (DDSTM) を提案する (図 1 およ

び Algorithm 1)．提案手法は，DTM を基礎とし，トピック分布および単語分布に事前分布に Spike and Slab Prior [6] を仮定することで，縮小推定を行うことができる．これによりトピック分布と単語分布のスパース性を先行研究のモデルである Dual Sparse Topic Model と同程度に担保できることを証明する．また，単語分布とトピック分布の時系列変化を確認できることに加えて，スパース性により各トピックや単語分布の解釈が容易になる利点を有する．以上の提案手法の有効性を確認するために人工データおよび実際のテキストデータを用いた実証分析を行う．

2 準備: 動的トピックモデル

動的トピックモデル (DTM) [3] は既存のトピックモデルに時系列性を取り入れ，文書のトピック分布のパラメータおよび各トピックの単語分布が時間とともに変化するモデルである．LDA は多項分布の母数の事前分布として共役事前分布であるディリクレ分布からサンプルすることで，各トピックの確率的表現を行うと同時に，各パラメータの反復推定を容易にしている．

一方で，LDA を含むトピックモデルでは，各文書・トピックは独立性が仮定されており，それらの間に存在する時系列的な構造を把握することが難しい．そこで，動的トピックモデルでは，状態空間モデルを参考にトピックモデルの事前分布を改良している．状態空間モデルは，システムモデル H_0 と観測モデル H_1 を仮定し，Kalman-filter や H-M Algorithm 等といった手法を用いることで，パラメータ更新を行う．動的トピックモデルでは，システムモデルを正規分布からなる確率過程，観測モデルを LDA と仮定することで，状態空間モデルのパラメータ更新手法と同様に多項分布の母数の事前分布として正規分布にソフトマックス関数を使用している．これにより，推定コストが増えるものの，時系列的なトピック構造や文書構造を考慮したトピックモデルの構築が可能となった．本研究では，DTM のトピック分布および単語分布にスパース性を導入する．

3 提案手法: 動的スパース・トピックモデル

本稿で提案する動的スパース・トピックモデルは，Spike and Slab 事前分布 [6] のように，一点分布 (ベルヌーイ分布) と連続分布 (平滑化パラメータと弱平滑化パラメータ) の混合を含んでおり，Dual Sparse Topic Model(DSTM) [4] を時系列に拡張した手法である．提案手法における記法およびパラメータは表 1 の通りである．ここで，$\gamma \gg \bar{\gamma}, \pi \gg \bar{\pi}$ である．提案手法のグラフィカルモデルは図 1 に示し，確率生成モデルは Algorithm 1 の通りであり，スパース性を持つ二重の確率生成構造からなる．

表 1: 提案手法における記法

記法	説明
t, T	時点および時点数
d, D	文書および文書集合
k_t, K_t	時点 t におけるトピック k およびトピック数
A_d	文書 d における注目されたトピック集合
r	トピックを定義する，トピック内の単語
V	トピックに含まれる単語の集合（語彙）
B_k	トピックにおいて注目された単語集合
i, N_d	文書 d に含まれる語および語数
γ	単語平滑化パラメータ
$\bar{\gamma}$	単語弱平滑化パラメータ
π	トピック平滑化パラメータ
$\bar{\pi}$	トピック弱平滑化パラメータ
Mult$(\cdot)$	多項分布

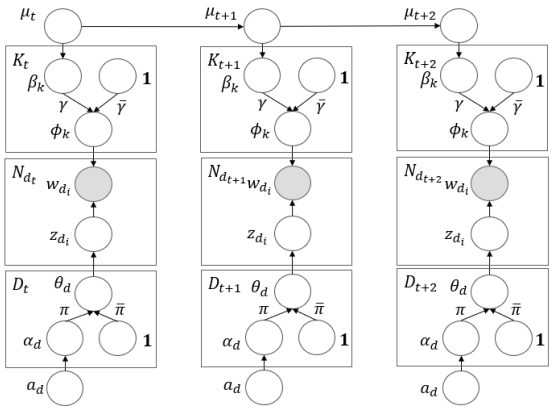


図 1: 提案手法のグラフィカルモデル

提案手法におけるスパース性は，spike and slab と同様のアイデアからもたらされている．ここでの「spike」は，ベルヌーイ分布に従う β_{kr} である．β_{kr} は，事前分布として b_k を持つベルヌーイ分布に従う変数で，b_k の大きさに応じて，ベルヌーイ分布により各トピックが「注目される」か，「注目されない」かを決定している．その後，「slab」として，$(\gamma, \bar{\gamma})$ による平滑化を加えることで，実質的に一点分布と連続分布の混合分布に近い性質を持たせて

Algorithm 1 確率生成モデル

Input: D: 文書集合，K_t: トピック数，$\gamma, \bar{\gamma}, \pi, \bar{\pi}$: パラメータ

for $t \in \{1, 2, ..., T\}$ **do**
  for $k \in \{1, 2, ..., K_t\}$ **do**
    $\mu_t \leftarrow N(\mu_{t-1}, \sigma^2))$
    $b_k \leftarrow \text{logistic}(\mu_t)$
    for $r \in \{1, 2, 3, ..., |V|\}$ **do**
      $\beta_{kr} \sim \text{Bernoulli}(b_k)$
    end for
    $\vec{\beta}_k \leftarrow \{\beta_{kr}\}_{r=1}^{|V|}$
    $\vec{\phi}_k \sim \text{Dirichlet}(\gamma\vec{\beta}_k + \bar{\gamma}\vec{1})$
  end for
  for $d \in \{1, 2, ..., |D|\}$ **do**
    $a_d \leftarrow \text{Beta}(s, v)$
    for $k \in \{1, 2, ..., K_t\}$ **do**
      $\alpha_{dk} \sim \text{Bernoulli}(a_d)$
    end for
    $\vec{\alpha}_d \leftarrow \{\alpha_{dk}\}_{k=1}^{K_t}$
    $\vec{\theta}_d \sim \text{Dirichlet}(\pi\vec{\alpha}_d + \bar{\pi}\vec{1})$
  end for
  for $i \in \{1, 2, ..., N_d\}$ **do**
    $z_{d_i} \sim \text{Mult}(\vec{\theta}_d)$
    $w_{d_i} \sim \text{Mult}(\vec{\phi}_{z_{d_i}})$
  end for
end for

いる．このアプローチは noise OR model [7] や aspect Bernoulli model [8] でも使用されている．なお，もし $\bar{\gamma} = 0$ すなわち，$\bar{\gamma}$ による弱平滑化がない場合，$\vec{\beta}_k = \mathbf{0}$ の時に，$\text{Dirichlet}(\gamma\vec{\beta}_k)$ を定められない [9, 4]．そのため，本稿における DDSTM でも，平滑化パラメータと弱平滑化パラメータという二種類のパラメータで平滑化を行うことで，上記の問題に対処している．

3.1 理論的性質

先行研究 [9, 4] 同様にスパース性を以下のように定義する．

定義 1 (スパース性). B_k をトピック k において注目されている単語の集合としたとき，スパース性 $\textbf{sparsity}(k)_t$ は

$$\textbf{sparsity}(k)_t := 1 - \sum_{r=1}^{|V|} \frac{\beta_{kr}}{|V|} = 1 - \frac{|B_k|}{|V|} \tag{1}$$

と定義される．

定理 1 (informal). 無情報事前分布 (ベータ分布において $a = b = 1$) を仮定した場合，提案手法と Dual Sparse Topic Model のスパース性の期待値は変わらない．

証明 Dual Sparse Topic Model [4] において，トピック生成過程における sparsity の期待値を考える．まず，ベルヌーイ分布のパラメータ b_k を使用して，条件付期待値を考える．また，途中で，無情報なベータ分布として $a = b = 1$ であるとすると，

$$\begin{aligned}
\mathbb{E}(\textbf{sparsity}(k)_t) =& \mathbb{E}_{b_k}(\mathbb{E}(\textbf{sparsity}(k)_t|b_k)) && (2)\\
=& 1 - \mathbb{E}_{b_k}(b_k) && (3)\\
=& 1 - \frac{a}{a+b} \because b_k \sim \text{Beta}(a, b) && (4)\\
=& 0.5 && (5)
\end{aligned}$$

本稿で提案する DDSTM において，トピック生成過程における sparsity の条件付き期待値を考えると，同様に，

$$\begin{aligned}
\mathbb{E}(\textbf{sparsity}(k)_t) =& \mathbb{E}_{b_k}(\mathbb{E}(\textbf{sparsity}(k)_t|b_k)) && (6)\\
=& 1 - \mathbb{E}_{b_k}(b_k) && (7)\\
=& 1 - \mathbb{E}(\text{logistic}(N(\mu_{t-1}))) && (8)
\end{aligned}$$

ここで，logistic 関数は全ての実数で微分可能な関数なため，デルタ法を使用し，漸近平均を考えると，

$$\begin{aligned}
(8) \text{ 式} \xrightarrow{d}& 1 - \text{logistic}(\mu_{t-1}) && (9)\\
=& 1 - \frac{\exp(\mu_{t-1})}{1 + \exp(\mu_{t-1})} && (10)
\end{aligned}$$

ここで，$\mu_{t-1} = 0$ であるならば，(10) 式 $= 0.5$ となり，(5) と同じ値になることが確認できた． □

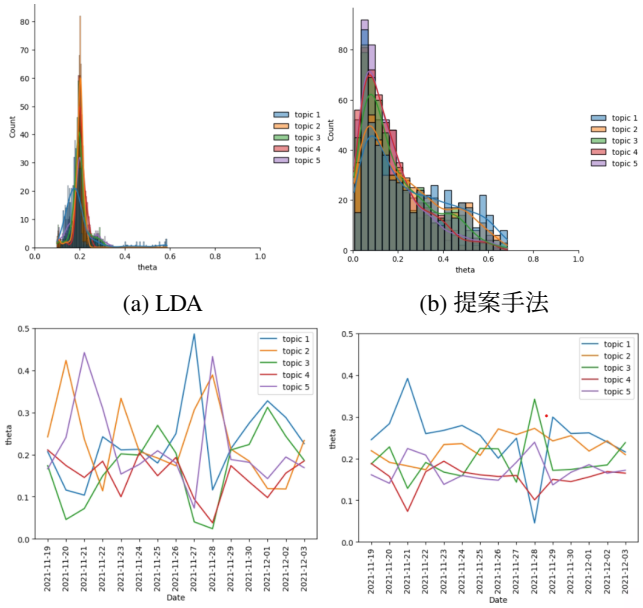
上記の定理から，無情報事前分布や，無情報な母数を与えた場合には，動的にモデルを組んだとしても，その全体としてのスパース性能には影響を与えないことを示すことができた．

4 実証分析

4.1 実験設定

本実験では，ロイターニュース[1)]から 2021 年のビジネスニュース記事のヘッドラインを利用する．特にオミクロン株の発生した 2021 年 11 月 26 日の前後 1 週間，すなわち 2021 年 11 月 19 日から 12 月 3 日までの 15 日分のニュース，計 542 記事を抽出した．本実験の対象期間は 5 日間と短期であるた

1) https://jp.reuters.com/



(a) LDA　(b) 提案手法

(c) DSTM によるトピック分布　(d) 提案手法の推移

図 2: 各手法によるトピック分布の比較

め，トピック分布は期間中変化する一方で，トピック内の単語分布の変化は相対的に小さいことが想定される．そのため，実験ではスパース性をトピック分布 θ のみに持たせる．なお，提案モデルの入力は，ヘッドラインに含まれる語を正規化，数詞を取り除いた後に，辞書として Mecab Neologd[10] を使用し名詞[2] の表層形を抽出し，一日を一つの時間間隔 t として集計した BoW 形式の特徴量とする．提案モデルの評価としては，学習結果として得られるトピック推移を算出し，LDA および DSTM との比較を定性的に行う．提案手法の各パラメータについては次のように設定した．$K_t = 5, \sigma = 1, \gamma = 0.1, \bar{\gamma} = 1e-7, \pi = 0.1, \bar{\pi} = 1e-7, s = 1, v = 1$ なお，モデルの学習には，マルコフ連鎖モンテカルロ法 (MCMC) を使用した．

4.2 結果

図 2 は，各手法によって得られたトピック分布を示している．まず図 2a の通常の LDA では，ほとんどの値が 0.2 周辺に値が集まっていることがわかる．一方で，図 2b に示す提案手法の DDSTM では二種類の平滑化により，多くの値が 0 周辺に分布しており，スパース性が確認できる．また，slab の影響により，theta の中央値が 0.05 周辺であることも確認できた．提案手法のトピックの単語分布 ϕ について確認すると，topic 1 が「インフレ」，topic 2 が「労働市場」，topic 3 が「コロナ」，topic 4 が「中央銀行」，topic 5 が「海外」に関連する単語の重要性が高くなっている．オミクロン株の発生後に topic 3 が大きく上昇していることから，動的にすることで，コロナの影響を正しく推定することがすることができたと考える．

次に，図 2c は全時点のデータを使用し推定した結果を各時点 t で期待値をとったもの，図 2d は DSTM[3] における，時点 t，トピック k ごとの θ の期待値の推移を示している．DSTM では時点毎に独立を仮定しているため各パラメータが大きく上下に振れているものの，提案手法では時系列性を考慮しているため上下の変動を比較的抑えることができたといえる．その一方で，オミクロン株の発生した 2021 年 11 月 26 日の前後でトピックを表すパラメータの変動を捉えることができている．

5 まとめ

本研究の貢献は以下の通り．

- DSTM の一般化を行い，新しいモデルである動的スパーストピックモデル (DDSTM) を提案した．
- DDSTM の理論的性質を確認し，スパース性能は不変であることを確認した．
- 実際の経済ニュースデータを用いた数値実験によって，トピック分布のスパース性を確認した．

なお，本稿では，トピックの時間推移をとらえることが目的であるため，β のみを動的にしたが，α についても同様に動的にすることは容易に可能である．また，今回は離散的な時間として動的トピックモデルを発展したが，モデルを連続時間に発展させることで，より精緻にトピックの趨勢を把握すること可能であると考えられる．

2) なお，本稿では計算時間の削減のため，品詞のうち以下の属性を持つものに絞った．[”名詞, 一般,*,*”, ”名詞, 引用文字列,*,*”, ”名詞, 形容動詞語幹,*,*”, ”名詞, 固有名詞, 一般,*”, ”名詞, 固有名詞, 人名, 一般”, ”名詞, 固有名詞, 人名, 姓”, ”名詞, 固有名詞, 人名, 名”, ”名詞, 固有名詞, 組織,*”, ”名詞, 固有名詞, 地域, 一般”, ”名詞, 固有名詞, 地域, 国”]

3) なお，本稿では提案手法との比較の観点から ϕ についての sparse 性は実装していない．

参考文献

[1] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pp. 289–296, 1999.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, Vol. 3, No. Jan, pp. 993–1022, 2003.

[3] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120, 2006.

[4] Tianyi Lin, Wentao Tian, Qiaozhu Mei, and Hong Cheng. The dual-sparse topic model: mining focused topics and focused terms in short text. In *Proceedings of the 23rd international conference on World wide web*, pp. 539–550, 2014.

[5] Jun Zhu and Eric P Xing. Sparse topical coding. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 831–838, 2011.

[6] Hemant Ishwaran and J Sunil Rao. Spike and slab variable selection: frequentist and bayesian strategies. 2005.

[7] Eric Saund. A multiple cause mixture model for unsupervised learning. *Neural Computation*, Vol. 7, No. 1, pp. 51–71, 1995.

[8] Ella Bingham, Ata Kabán, and Mikael Fortelius. The aspect bernoulli model: multiple causes of presences and absences. *Pattern Analysis and Applications*, Vol. 12, pp. 55–78, 2009.

[9] Chong Wang and David Blei. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. *Advances in neural information processing systems*, Vol. 22, , 2009.

[10] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pp. 230–237, 2004.