連続時間フラクショナル・トピックモデル

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

林 晃平
東京大学大学院数理科学研究科
kohei@ms.u-tokyo.ac.jp

藤本 悠吾
野村アセットマネジメント株式会社
yu5fujimoto@gmail.com

概要

LDA の時系列性を考慮するため，DTM や，DTM を連続時間に拡張した cDTM が提案されている．しかしながら，これらの生成パラメータの変化量に各時刻での相関を持たせることで，より現実に即したモデル化が可能になると考えられる．そこで本研究では，cDTM の一般化を行い，生成パラメータの増分の正相関性 (長時間依存性) または負相関性 (ラフさ) を考慮にいれた，連続時間フラクショナル・トピックモデルを提案する．また提案手法のパラメータ推定は簡易的には，トピックモデルと同様であることを示した．そして数値実験によって，提案手法がトピックの正相関性 (長時間依存性) または負相関性 (ラフさ) を捉えることができることを確認する．

1 はじめに

トピックモデルとは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単語を選択していくことで生成される．ここでトピック分布とは文書中の各話題の比率，単語分布とは各話題を構成する単語の分布を意味する．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，文書に含まれるの話題の比率や，話題を構成する単語の分布を知ることができる．トピックモデルの中でも，Blei らによって提案された Latent Dirichlet Allocation (LDA) [1] や，トピック間に相関を加えるモデル [2] も提案されている．

一般に LDA においては時系列性は考慮されないため，トピックの時系列的な推移を確認することはできない．この問題に対応するためにトピックの時系列性を考慮する LDA である Dynamic Topic Model (DTM) が提案された [3]．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．さらに DTM が離散的な時間変化をモデリングしているのに対して，DTM を連続時間に拡張した Continuous Time DTM (cDTM) が提案された [4]．

上述の通り，従来の動的トピックモデルは各時刻における単語生成およびトピック生成のパラメータが時間変化するモデルであった．そこでは，1 ステップ前のパラメータに対し独立に正規分布に従う確率変数を加えることで次の時刻のパラメータを定めており，生成パラメータの増分は各時点で独立である．しかしながら，例えばある時刻で新単語または新トピックに特有の単語が増加傾向にあるとすると，次の観測点の時刻でもその増加傾向に応じて単語とトピックの分布が変化すると考えるのが自然である．即ち，生成パラメータの変化量は，各時刻で正の相関を持つと仮定することで，より現実に即した文書生成を行えると考えられる [5, 6, 7, 8]．また，逆にある時刻で語られていたトピックが次の時刻では語られなくなり，新しいトピックが急に語られるような事象，すなわちトピック分布の負の相関性 (ラフさ) を記したいケースも考えられる．そこで本研究では，生成パラメータの増分の正相関性および長時間依存性またはラフさを考慮にいれた手法である，連続時間フラクショナル・トピックモデルを提案する．提案手法は，標準 Brown 運動の一般化であり，増分が正の相関を持ち，かつ長期記憶性をまたはラフさを持つ非整数階 Brown 運動 [9] を生成パラ

メータの増分過程として利用する．したがって，提案手法は cDTM の一般化になっている．そのため，テキストのタイムステップが均一でない場合あるいは欠損がある場合でもモデル化可能である．そして，実際の経済ニュースデータを用いた数値実験によって，提案手法がトピックの正相関性 (長時間依存性) または負相関性 (ラフさ) を捉えることができることを確認する．

2 提案手法

2.1 非整数階 Brown 運動

ここでは，非整数階 Brown 運動の定義と基本的な性質について概説する [10]．$H \in (0,1)$ を定数 (Hurst 指数) とする．平均 0 の実数値 Gauss 過程 $B^H = \{B_t^H\}_{t\geq 0}$ が Hurst 指数 H の fractional Brown 運動 (fBm) であるとは，ほとんど確実に $B_0^H = \mu$ かつ，任意の $s,t \geq 0$ に対して次を満たすときをいう．

$$\mathrm{Cov}(B_s^H, B_t^H) = \frac{1}{2}(|t|^{2H} + |s|^{2H} - |t-s|^{2H}). \quad (1)$$

また，fBm から $\{B_{t+1}^H - B_t^H : t = 0,1,\ldots\}$ という新たな離散確率過程を考えることができ，これを fractional Gauss ノイズ (fGn) という．fBm の増分は $H > 1/2$ ($H < 1/2$) のとき正 (負) の相関をもち，$H = 1/2$ のとき Brown 運動に一致することは定義から明らかである．また，$B_t^H - B_s^H \sim \mathcal{N}(0, |t-s|^{2H})$ なので，fBm は定常増分である．次に，確率過程の長期 (短期) 記憶性を定義する．$X = \{X_t\}_{t\geq 0}$ を確率過程とし，その増分を $X_{s,t} = X_t - X_s$ とする．このとき，確率過程 X の増分が長期（resp. 短期）記憶性を持つとは，任意の $h > 0$ に対して，$\sum_{n=1}^{\infty} |\mathrm{Cov}(X_{0,h}, X_{(n-1)h,nh})|$ が発散（resp. 有限）になることをいう．(1) 式から，fBm は $H = 1/2$ のとき，標準 Brown 運動と一致する．また fBm の増分に対しては $H > 1/2$ のときのみ長期依存性を持ち，$H < 1/2$ のときラフさを持つ．

2.2 連続時間フラクショナル・トピックモデル

以下，$\mathbf{K}$ をトピック全体の集合，$\mathbf{W}$ を単語全体の集合とする．連続時間フラクショナル・トピックモデルは，単語分布とトピック分布のパラメータが時間発展するモデルであり，その時間変動を fBm によってモデル化する．$H \in (0,1)$ とし，$\{B_t^{H,(k,w)} : t \geq 0\}_{k\in\mathbf{K}, w\in\mathbf{W}}$ 及び $\{B_t^{H,(k)} : t \geq 0\}_{k\in\mathbf{K}}$ を初期値を 0 とする独立な fBm の列とする．タイムスタンプの列 $\{0 = s_0, s_1, \ldots, s_T = T\}$ における生成パラメータ $\alpha_{s_t}, \beta_{s_t}$ $(t = 0,\ldots,T)$ を，初期分布 $\alpha_0 \in \mathbb{R}^{\mathbf{K}}, \beta_0 \in \mathbb{R}^{\mathbf{K}} \times \mathbb{R}^{\mathbf{W}}$ と次式によって定めることができる．各 $k \in \mathbf{K}$ 及び $w \in \mathbf{W}$ に対し，次の確率微分方程式を解く．

$$\begin{aligned} d\alpha_{s,k} &= f_{\theta_\alpha}(\alpha_{s,k})ds + \sigma_\alpha dB_s^{H,(k)}, \\ d\beta_{s,k,w} &= f_{\theta_\beta}(\beta_{s,k,w})ds + \sigma_\beta dB_s^{H,(k,w)}. \end{aligned} \quad (2)$$

ただし，$\sigma_\alpha, \sigma_\beta \in \mathbb{R}$ である．また，f_{θ_α} および f_{θ_β} は θ_α および θ_β をパラメータに持つ関数であり，方程式 (2) が一意的な解 $\{(\alpha_s, \beta_s) : s \geq 0\}$ が存在し，またそれらの解の密度が持つような十分性質のよいものであると仮定する．これらの関数により，トピック及び単語が生成される傾向を学習することができ，更に fBm により与えたノイズから長期記憶性やラフさを再現することができる．以下では，方程式 (2) を解いた後に，与えられたタイムスタンプ $s = s_0, \ldots, s_T$ における値を利用する．ここでは時間に関して連続な方程式を解くことで生成パラメータを決定しているため，不規則サンプリングに対応可能なことに留意されたい．このパラメータ過程を用いて，各タイムスタンプ s_t における総単語数 N_{s_t} の文書 $d_{s_t} = \{(w_{s_t,k_i}^i)_{1\leq i\leq N_{s_t}} : k_i \in \mathbf{K}, w_{s_t,k_i}^i \in \mathbf{W}\}$ の生成過程は次のように書ける．

1. 各タイムスタンプ s_t における $\alpha_{s_{t+1},k}$ と $\beta_{s_{t+1},k,w}$ を (2) 式で生成する
2. 各単語 $w \in \mathbf{W}$ について，まずトピック $z \in \mathbf{K}$ をパラメータ $\phi(\alpha_{s_t})$ のカテゴリ分布 (トピック分布) から一つ選び，そのトピック z の単語生成パラメータを $\beta_{s_t,z} = (\beta_{s_t,z,w})_{w\in\mathbf{W}} \in \mathbb{R}^{\mathbf{W}}$ として，パラメータ $\phi(\beta_{s_t,k})$ のカテゴリ分布 (単語分布) から w を選ぶ：

$$\begin{aligned} z &\sim \mathrm{Categorical}(\phi(\alpha_{s_t})), \\ w &\sim \mathrm{Categorical}(\phi(\beta_{s_t,z})). \end{aligned}$$

ここで，V-次元単体 $\sigma_V = \{x \in [0,1]^V : x_1 + \cdots + x_V = 1\}^V$ に対してパラメータ $\phi \in \sigma_V$ を持つ，Categorical(ϕ) は確率密度関数が

$$p_{\mathrm{Cat}}(x|\phi) = \prod_{v=1}^{V} \phi_v^{x_v},$$

で表される $\{0,1\}^V$ 上の分布である．また，$\phi : \mathbb{R}^V \to \sigma_V$ は softmax 関数であり次式で定義される．

$$\phi(\beta)_v = \frac{\exp(\beta_v)}{\sum_{1\leq v\leq V} \exp(\beta_v)}.$$

図 1 に提案手法である連続時間フラクショナル・トピックモデルのグラフィカルモデル表現を示す．ここで，ϕ^z_s (resp. ϕ^w_s) は時刻 s におけるトピック (resp. 単語) 分布である．ここでは簡単のため，ドリフト関数のパラメータ $\theta_\alpha, \theta_\beta$ の寄与は省略した．連続時間の微分方程式を解くことによりパラメータ生成を行っているため，サンプリングが不規則であっても対応できることを図では表現している．

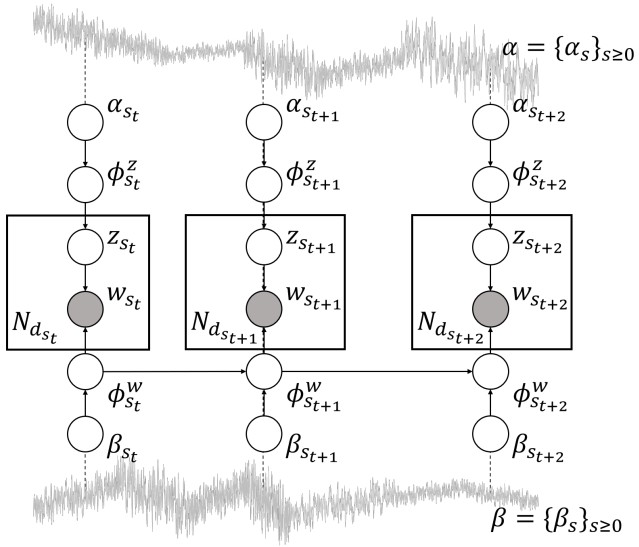


図 1: 連続時間フラクショナル・トピックモデルのグラフィカルモデル表現.

2.3 パラメータ推定

以下, $T+1$ 個のタイムスタンプ $\{0 = s_0, s_1, \ldots, s_T = T\}$ において観測された文書 $\hat{d}_{s_t} = \{\hat{w}^i_{s_t} \in \mathbf{W} : i = 1, \ldots, |\hat{d}_{s_t}|\}$ $(t = 0, \ldots, T)$ が与えられているとする．また，生成パラメータの初期分布を $\alpha_0 \sim \mathcal{N}(\mu_\alpha, \nu_\alpha I_{|\mathbf{K}|})$ および $\beta_0 \sim \mathcal{N}(\mu_\beta, \nu_\beta I_{|\mathbf{K}|\times|\mathbf{W}|})$ とする．ただし，$\mathcal{N}(\mu, \nu)$ は平均 μ，分散共分散行列 ν の正規分布を表す．このとき，連続時間フラクショナル・トピックモデルのパラメータは，$\Phi = (\Phi_\alpha, \Phi_\beta)$，ただし $\Phi_\alpha = (\mu_\alpha, \nu_\alpha, \sigma_\alpha, \theta_\alpha)$ および $\Phi_\alpha = (\mu_\beta, \nu_\beta, \sigma_\beta, \theta_\beta)$，で与えられる．このとき，ドリフト関数 f_{θ_α} (resp. $\mathrm{v}f_{\theta_\beta}$) のパラメータ数を D_{θ_α} (resp. D_{θ_β}) とすると，パラメータ空間は $\Theta = \Theta_\alpha \times \Theta_\beta$，ただし，

$$\Theta_\alpha = \underbrace{\mathbb{R}^{\mathbf{K}}}_{\mu_\alpha} \times \underbrace{\mathbb{R}_+}_{\nu_\alpha} \times \underbrace{\mathbb{R}}_{\sigma_\alpha} \times \underbrace{\mathbb{R}^{D_{\theta_\alpha}}}_{\theta_\alpha}$$

および

$$\Theta_\beta = \underbrace{\mathbb{R}^{\mathbf{K}\times\mathbf{W}}}_{\mu_\beta} \times \underbrace{\mathbb{R}_+}_{\nu_\beta} \times \underbrace{\mathbb{R}}_{\sigma_\beta} \times \underbrace{\mathbb{R}^{D_{\theta_\beta}}}_{\theta_\beta},$$

で表される．以下では，各時刻 $t = 0, \ldots, T$ における次の対数尤度関数をパラメータ $\Phi \in \Theta$ について最大化する．

$$L_{s_t}(\Phi) = \log p(\hat{d}_{s_t}|\Phi) = \sum_{\hat{w}_{s_t} \in \hat{d}_{s_t}} \log p(\hat{w}_{s_t}|\Phi) \tag{3}$$

ここで，各時刻 s について，$\Phi_\alpha = (\mu_\alpha, \nu_\alpha, \sigma_\alpha)$ および $\Phi_\beta = (\mu_\beta, \nu_\beta, \sigma_\beta)$ としてパラメータ $\Phi = (\Phi_\alpha, \Phi_\beta)$ が与えられたときの各単語の確率密度関数は，

$$\begin{aligned} p(w_s|\Phi) &= \sum_{k\in\mathbf{K}} p(z_s = k|\Phi) p(w_{s,k} = w_{s_t}|\Phi) \\ &= \int_{\mathbb{R}^{\mathbf{K}}} \int_{\mathbb{R}^{\mathbf{K}\times\mathbf{W}}} \sum_{k\in\mathbf{K}} p_{\mathrm{Cat}}(z_s = k|\phi(\alpha_s)) \\ &\quad \times p_{\mathrm{Cat}}(w_{s,k} = w_s|\phi(\beta_{s,k})) \\ &\quad \times p(\alpha_s|\Phi_\alpha) p(\beta_s|\Phi_\beta) d\alpha_s d\beta_s. \end{aligned} \tag{4}$$

で与えられる．ここで，$p(\alpha_s|\Phi_\alpha)$ および $p(\beta_s|\Phi_\beta)$ は時刻 s における α_s および β_s の確率密度関数である．

本稿では，トピックまたは単語分布の長期記憶性やラフさを再現することを確認することが目的であり，簡単のため単純化された設定の下で文章生成を行う．そのもとで，提案手法の尤度 (4) の最大化について次が成立する．

命題 1. ドリフト関数が $f_\alpha = 0$ かつ $f_\beta = 0$ の場合，尤度 (4) の最適化は古典的な (即ち時間発展のない定常的な) トピックモデルの最適化問題に帰着される．

そのため特に，EM アルゴリズムや変分ベイズ推定などの手法を適用することができる．

3 実証分析

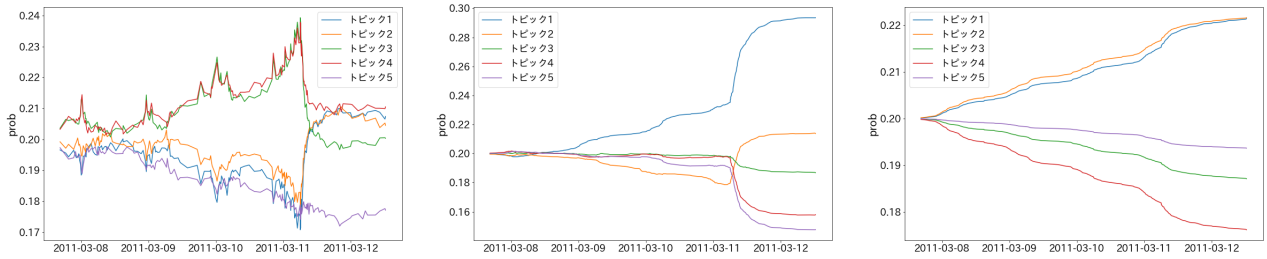
3.1 データセット

提案モデルによる時系列相関を考慮したトピック抽出を評価するため，本実験では重大イベント (東日本大震災) 前後でのトピック推移について定性的な評価を行う．提案モデルの評価のため，ロイターニュース[1)]から東日本大震災を含む前後 5 日間 (2011/3/8〜3/12) のニュース記事 178 件を抽出し，実験を行った．

3.2 実験設定

提案モデルの入力には，ニュース記事から名詞を抽出し計算した Bag of Words 形式の特徴量を利用した．またモデルの学習には，マルコフ連鎖モンテカ

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



(a) $H = 0.1$(ラフな fBm)　(b) $H = 0.5$(cDTM)　(c) $H = 0.9$(長期記憶性を持つ fBm)

図 2: 東日本大震災前後のニュースのトピック分布推移

ルロ法 (MCMC) を利用した．本実験の対象期間は 5 日間と短期であるため，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．よって本実験では，トピック分布の生成パラメータ α のみ時間発展する設定で実験を実施した．提案モデルの評価としては，トピックの時系列相関を制御するパラメータである Hurst 指数を変化させた際に，学習結果として得られるトピック推移，および上位トピックに出現する単語を算出した．提案モデルにおけるハイパーパラメータである Hurst 指数 H をそれぞれ 0.1,0.5,0.9 に設定し，トピック数は 5 に設定し実験を行った．それぞれ，$H = 0.1$ の場合はラフさを，$H = 0.5$ のときは Brown 運動 (cDTM)，$H = 0.9$ の場合は長期記憶性を持つ．

3.3　結果

図 2 は，提案モデルによって得られたトピック分布のパスの推移を示している．Hurst 指数が大きい設定では，イベント前後でトピックの推移が大きく変化することはなく，各トピックの正相関性が保たれてことがわかる．一方で，Hurst 指数を小さくしたケースではイベント後にトピックの発生確率の大幅な変動がみられる．中間の H=0.5，つまり cDTM のケースは両者の中間の位置付けとなっている．このように，Hurst 指数を制御することで，提案モデルが長期依存性を持つトピックや短期依存性を持つトピックの動向を追跡できている．

$H = 0.1$ について，関連するニュースの増加に伴うトピック分布の変動を捉えられていると考えられる．具体的に，5 つのトピックはそれぞれ，トピック 1(青線) は国内ニュース，トピック 2(オレンジ線) は国内市況，トピック 3(緑線) は海外ニュース，トピック 4(赤線) は海外市況，トピック 5(紫線) は中国関連と解釈できる．イベント前はリビア情勢不安や中国の政府・経済動向といった海外ニュースおよび海外市況のトピックが多かった．一方で，イベント後は震災に関する国内ニュースおよび国内市況のトピックが大幅に上昇した．

4　まとめ

本研究の貢献は次の通りである．

- cDTM の一般化を行い，生成パラメータの増分の正相関性 (長時間依存性) および負相関性 (ラフさ) を考慮にいれた手法である，連続時間フラクショナル・トピックモデルを提案した．
- 提案手法のパラメータ推定は簡易的には，トピックモデルと同様であることを示した．
- 実際の経済ニュースデータを用いた数値実験によって，提案手法がトピックの長時間依存性やラフさを捉えることができることを確認した．

本研究の限界として，トピックまたは単語分布の長期記憶性やラフさの再現が主目的のため，ドリフト項を考慮しなかった．加えて，fBm は一般には独立増分性を持たないため [3, 4] と同様に Kalman Filter をもとにトピック分布あるいは単語分布の事後分布を効率的に計算することができない．

今後の発展として，トピック分布あるいは単語分布の事後分布の効率的な計算方法を考察することが挙げられる．また，本研究では考慮していないドリフト項を非線形なニューラルネット関数を用いて学習させる ODE-Net (SDE-Net) ベースの拡張 [11] または本研究と同様に fBm による ODE-Net ベースの拡張 [12] が挙げられる．

参考文献

[1]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.

[2]David M Blei and John D Lafferty. A correlated topic model of science. *The annals of applied statistics*, Vol. 1, No. 1, pp. 17–35, 2007.

[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]Chong Wang, David M Blei, and David Heckerman. Continuous time dynamic topic models. In *UAI'08 Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, 2008.

[5]Liangjie Hong and Brian D Davison. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pp. 80–88, 2010.

[6]Xuanhui Wang, ChengXiang Zhai, Xiao Hu, and Richard Sproat. Mining correlated bursty topic patterns from coordinated text streams. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 784–793, 2007.

[7]Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 424–433, 2006.

[8]Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 497–506, 2009.

[9]Benoit B Mandelbrot and John W Van Ness. Fractional brownian motions, fractional noises and applications. *SIAM review*, Vol. 10, No. 4, pp. 422–437, 1968.

[10]Francesca Biagini, Yaozhong Hu, Bernt Øksendal, and Tusheng Zhang. *Stochastic calculus for fractional Brownian motion and applications*. Springer Science & Business Media, 2008.

[11]Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, Vol. 31, , 2018.

[12]Kohei Hayashi and Kei Nakagawa. Fractional sde-net: Generation of time series data with long-term memory. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2022.

A Appendix: 命題の証明

ドリフト関数が $f_\alpha = 0$ かつ $f_\beta = 0$ の場合，式 (2) は次のように解くことができる．

$$\alpha_s = \alpha_0 + \delta B_s^H, \quad \beta_s = \beta_0 + \sigma B_s^H$$

よって，時刻 s において α_s および β_s の分布は Gauss 分布を用いて明示的に表すことができる．これは古典的な (即ち時間発展のない定常的な) トピックモデルの最適化問題と等しく，そのため尤度 (4) の最適化はトピックモデルの最適化問題に帰着される．