時系列トピックモデルを用いた株式市場の分析

関西学院大学 理工学研究科 山内 海渡 *
関西学院大学 理工学部　　森本 孝之

1　研究概要

時間情報を考慮したトピックモデル (Online Multiscale Dynamic Topic Model[3]) を用い，時間情報を持ったニュース記事に対してトピックを割り当て，記事集合内のトピックの時間発展を推定する．推定したトピックの時系列変化と東証株価指数 (TOPIX) のボラティリティとの関連を調べる．

2　研究背景

近年，情報科学や経済学において，検索指数やオンラインニュース記事，ブログなどのウェブ上の情報を用いた実世界の動向分析が盛んに研究されている．経済に関する研究としては Google Search Volume Index(SVI) を用いて車や家の売上の予測をおこなった Varian らの研究 [2]，ニュースのヘッドラインや Google 検索指数を株価のボラティリティ予測に用いた Vlastakis らの研究 [6] などがある．これらの研究では，特定のキーワードの検索頻度や，キーワードを含むニュースの数を用いていた．しかしこの手法ではキーワードとして選ぶ単語に大きく結果が左右される場合があるほか，表記ゆれにも弱い．また複数の意味を持つ単語をうまく扱うことができない．また，Google によって提供される検索指数データは週次データであり，日次の分析が困難である．これらを踏まえ，本研究ではキーワードではなく，トピック (話題) を用いた分析を試みる．

*連絡先: 関西学院大学 理工学研究科，〒 669-1337 兵庫県三田市学園 2 丁目 1，Email: bse77252@kwansei.ac.jp

3　トピックモデル

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．われわれは，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．トピックモデルでは，文書を単語の集合とみなしている．これを Bag-of-Words モデルという．例えば，文書 D=” 今日はいい天気だ” を Bag-of-Words 表現で表すと，D={“今日”,“は”,“いい”,“天気”,“だ”} となる．単語以外の情報，例えば単語の順序や係り受けといった情報は捨て，単語の頻度のみに着目している．

3.1　オンライン学習可能な多重スケール時間発展トピックモデル

本研究では特定のトピックの時間発展を追跡するために Online Multiscale Dynamic Topic Model(MDTM)[3] を用いる．MDTM は複数の時間スケールを考慮した，トピックの時間発展を解析するためのモデルである．また MDTM はオンライン学習が可能である．標準的なトピックモデルである Latent Dirichlet Allocation(LDA)[1] を適用した場

1

合，各区間における単語分布およびトピック分布は他の区間の単語分布およびトピック分布と独立なので，特定のトピックの単語分布の時間変化やトピック分布の時間変化を追跡することは難しい．MDTM ではトピックの事前分布のパラメータの分布が一期前の分布に依存し，また単語分布は複数の時間スケールを持つ過去の単語分布に従う．これにより，特定のトピックの単語分布の時間変化やトピック分布の時間変化を追跡することが出来る．ここで，時間スケールを持つ単語分布とは，データに含まれる単語の出現期間を考慮した単語分布を表す．例えばニュース記事中における政治に関する単語として,「憲法」「国会」などは百年以上の長期間に渡って出現する一方，国会議員の名前などであれば数年から数十年，法案名や政策名などの中には数日から数ヶ月程度しか出現しないものもある．MDTM は，このような単語の出現期間を考慮した単語分布の重み付け和を現時点の単語分布の事前分布として用いることで，従来の時間を考慮したトピックモデル (引用する) より，論文誌やブログ記事データ等を用いた実験において低い予測誤差を達成している．

以下に MDTM の生成モデルを示す．

1. For each topic $k = 1, \ldots, Z$:
   (a) 各トピックの単語分布を生成
   $\boldsymbol{\phi}_{t,z} \sim \mathrm{Dirichlet}(\sum_{s=0}^{S} \lambda_{t,z,s} \hat{\boldsymbol{\omega}}_{t-1,z}^{(s)})$
   (b) トピックの事前分布のハイパーパラメータを生成
   $\alpha_{t,z} \sim \mathrm{Gamma}(\gamma \alpha_{t-1,z}, \gamma)$
2. for each document $d = 1, \ldots, D_t$:
   (a) トピック分布を生成
   $\boldsymbol{\theta}_{t,d} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}_t)$
   (b) for each word $n = 1, \ldots, N_{t,d}$:
      i. トピックを生成
      $z_{t,d,n} \sim \mathrm{Multinomial}(\boldsymbol{\theta}_{t,d})$
      ii. 単語を生成
      $w_{t,d,n} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{t,z_{t,d,n}})$

ここで D_t は時点 t における文書数，$N_{t,d}$ は時点 t の文書 d 中の単語数，Z および S は予め設定したトピック数およびスケール数を表す．$\hat{\boldsymbol{\omega}}_{t-1,z}^{(s)}$ は $t-1$ 時点におけるスケール s のトピック z の単語分布，$\lambda_{t,z,s}$ は各時間スケールの単語分布がどの割合で混合しているかを表す重みである．

また MDTM は確率的 EM アルゴリズムによりオンライン推定が可能であり，大量のデータにも少ないメモリで適用可能である．確率的 EM アルゴリズムでは，現時点の各単語に Collapsed Gibbs Sampling を用いてトピックを割り当て，その後に多重スケール単語分布の重み λ とトピック分布の事前分布の超パラメータ α を不動点反復法 [4] を用いて最尤推定する．この 2 つのステップを Perplexity が収束するまで繰り返すことで，各パラメータを推論することが出来る．Iwata らの論文 [3] では多重スケール単語分布の更新を近似的に行うことで更に必要なメモリの削減を図っている．

4 ボラティリティ時系列モデル

収益率の分散はボラティリティと呼ばれ，株式投資においてリスクの指標として用いられる．ボラティリティを推定するためのモデルとして，ARCH モデル，GARCH モデル，SV モデルなどが提案されてきた．これらは日次の収益率の系列を用いて日次のボラティリティを推定するモデルである．しかしモデルによる推定はモデル自体の制約が推定に加わる．そこで近年，モデルを用いず，高頻度データとよばれる日内のデータの標本分散を用いてボラティリティを推定する，リアライズドボラティリティ(RV) という手法が普及してきている．[5] 本研究においても，ボラティリティの推定量として RV を用いる．以下に高頻度データおよび RV について簡単に述べておく．

2

4.1 高頻度データとリアライズドボラティリティ

高頻度データとは一般に，サンプリング頻度が極めて高い，集約・集計前の個別案件を記録した件数の非常に大きいデータのことである．とりわけ金融証券市場取引においては，一日内の取引を記録したデータを指す．そのサンプリング頻度は，数分程度の等間隔な時点でサンプルしたデータや，サンプリングを行わない，全取引の非等間隔な生データ (ティックデータ) まで様々である．RV の計算に当たっては，非等間隔なティックデータを，適当な補完方法により等間隔に加工したデータを用いる．

ここで，ある時点 t の RV の定義を述べる．

$$RV_t = \sum_{i=1}^{n_t} r_{t,i}^2$$

ここで，$r_{t,i}^2$ はある時点 t の i 番目に観測された収益率の二乗である．n_t は t 時点内のデータ数を表す．RV はボラティリティに対する一致性および普遍性を持つ推定量であることが知られている．

5 株式市場分析

時系列トピックモデルを用いて，ニュース記事集合の各時点のトピックを推定する．ある時点内の特定のトピックの “盛り上がり” を図るためのスコアを定義し，そのスコアを TOPIX の収益率のボラティリティの時系列モデルに外生変数として加え，モデルに効果的に働くかどうかをみる．

5.1 RV データ

本研究では 2008 年 1 月 4 日～2008 年 12 月 30 日のうち東証開場日を分析の対象期間とする．RV の計算に用いたデータはティックデータから 1 分間の終値を取り出したデータである．昼休みや夜間などの閉場時間内のデータは存在しないので，ここでは無視している．以下に 2008 年のリアライズドボラティリティのプロットを示す．

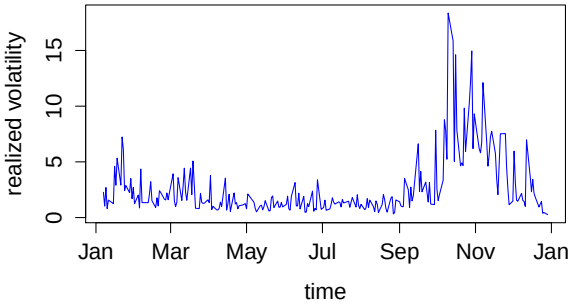

図 1: 2008 年のリアライズドボラティリティ

5.2 トピック分析

ニュース記事のデータは，以下のものを用いた．分析には以下のデータを用いた．

記事	ロイター日本語版の記事
期間	2008/1/7～2008/12/30/のうち東証開場日
単位時間数	243
文書数	14763
語彙数	42498

ただし，以下の様な前処理を施している．

- ロイターの記事には海外情勢に関する記事が多く含まれている．本研究は日本の株式市場が対象なので，ロイター東京支局の記事以外の記事は削除した．
- 形態素解析後，名詞のみを取り出し，更に数詞，接尾語，非自立語，代名詞を削除した．
- 形態素解析語に残った記号類は削除した．

以上のニュース記事データに次の設定で MDTM を適用した．

3

トピック 1	トピック 2	トピック 3	トピック 4	トピック 5
日経 0.109	円 0.208	先物 0.091	予想 0.092	株 0.079
平均 0.107	現在 0.069	上昇 0.079	発表 0.086	証券 0.078
続伸 0.043	週末 0.053	下落 0.077	決算 0.084	投資 0.062
東証 0.038	一時 0.040	米 0.072	利益 0.037	景気 0.054
中心 0.037	上値 0.038	原油 0.067	業績 0.035	市場 0.052
投信 0.035	可能 0.036	安値 0.058	修正 0.023	日本 0.049
国内 0.032	時間 0.034	反落 0.031	連結 0.022	株式 0.045
東京 0.032	反発 0.033	前日 0.025	金融 0.020	声 0.045
寄り付き 0.031	後半 0.030	日本語 0.022	背景 0.018	不動産 0.035
大手 0.027	連休 0.024	石油 0.021	中間 0.018	買い 0.028

表 1: 6 月 2 日の各トピックごとの単語分布 (数字は確率を表す)

トピック数 Z	20
スケール数 S (スケール 0 を含まない)	3
α の事前分布を調節するパラメータ γ	1.0

実行時間は約 4694 秒であった.

表 1 に 6 月 2 日の 20 トピックの単語分布のうち 5 つを示す.

ここではトピック 5 に着目する. 便宜上, トピック 5 を景気トピックと呼ぶ.

5.3 トピックのスコア

以下で定義するトピックに対するスコアを, トピックの盛衰を計る指数として用いる. ある時点 t のトピック i のスコアは,

$$s_t = \sum_{j=d}^{D_t} \theta_{t,j,i}$$

ここで D_t は時点 t における文書数, $\theta_{t,j,i}$ は t 時点の j 番目の文書のトピック分布の i 番目の成分を表す. $\theta_{t,j,i}$ は t 時点の j 番目の文書において, あるトピック i が何割含まれているかを意味している.

図 2 に 2008 年の景気トピックのプロットを示す.

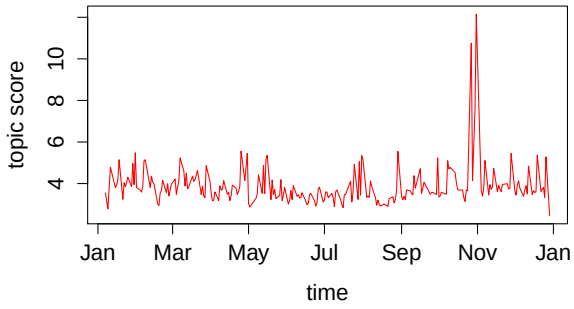

図 2: 2008 年の景気トピックのスコア

5.4 時系列モデル

まず予備実験として, Vlastakis らの先行研究 [6] にならい, 単純な AR(1) モデルについてみる.

ベースラインとなる RV の AR(1) モデルを以下に示す.

$$\log(RV_t) = \delta \log(RV_{t-1}) + \epsilon_t \qquad (1)$$
$$\epsilon_t \sim N(0, \sigma^2)$$

ただし $0 \leq \delta < 1$ とする. ここで δ は回帰係数である.

4

次に，RV の AR(1) モデルにトピックのスコアを外生変数として加えたモデルを示す．

$$\log(RV_t) = \alpha s_t + \delta \log(RV_{t-1}) + \epsilon_t \quad (2)$$
$$\epsilon_t \sim N(0, \sigma^2)$$

ただし，$0 \leq \alpha < 1, 0 \leq \delta < 1$ とする．ここで α, δ は回帰係数，s_t は注目したいトピックの t 時点でのスコアである．

5.5 時系列モデルの推定結果

ベースラインとなる RV の AR(1) モデルの推定結果は次のようになった．

	推定値	t 統計量の p 値
δ	0.75532	2×10^{-16}
自由度調整済決定係数	0.5661	
平均二乗誤差	0.440064	

表 2: AR(1) の推定結果

RV の AR(1) に，景気トピックのスコアを外生変数として加えたモデルは次のような結果になった．

	推定値	t 統計量の p 値
δ	0.75532	2×10^{-16}
α	0.025741	1.38×10^{-6}
自由度調整済決定係数	0.6047	
平均二乗誤差	0.3992485	

表 3: AR(1) にスコアを加えたモデルの推定結果

トピックのスコアを加えたモデルにおいて，トピックのスコアの回帰係数である α の t 統計量の p 値は 1.38×10^{-6} であり，よって十分に低い有意水準でこの係数が有意 (係数 $\alpha \neq 0$) であるといえる．またベースラインに対して自由度調整済み決定係数が上昇しているので，現在のデータに対するモデルの当てはまりは向上していると言える．よって景気トピックのスコアはこの時系列モデルに対して有効に働いているといえる．

6 まとめ

本研究では，時系列トピックモデルによってオンラインニュース記事集合から推定したトピックのスコア系列を用いて，TOPIX のリアライズドボラティリティの時系列モデルの改善を試みた．結果としては，トピックのスコアは AR(1) モデルに有効に働いた．しかしモデルの改善の程度は大きなものではなかった．その原因の 1 つとして，時系列トピックモデルではトピックの推定に以前の期間の単語分布およびトピック分布を用いるので，唐突な経済の動きやニュース記事の増加にトピックモデルが反応できていないことが考えられる．これを克服するには，MDTM のトピックの時間に対する一貫性を調整するパラメータ γ を大きな値にし，トピック分布の事前分布の分散を大きく設定することが考えられる．これによってある時点のトピックが以前のトピックに引きずられにくくなるので，唐突な変化にも対応できると考えられる．

今回の実験に続き，単純な AR モデルだけではなく，より複雑なモデル，例えば多変量の時系列モデルである VAR(Vector Autoregressive) モデル等について，トピックのスコアが有効に働くかどうかを検証していく．また，今回は株式市場を代表する指標として TOPIX を用いたが，日経平均等他の指標についても分析する．

参考文献

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[2] Hyunyoung Choi and Hal Varian. Predicting the present with google trends. Technical report, 2009.

[3] Tomoharu Iwata, Takeshi Yamada, Yasushi Sakurai, and Naonori Ueda. Online multiscale dynamic topic models. In *Proceedings of the*

5

16th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '10, pages 663–672, New York, NY, USA, 2010. ACM.

[4] Thomas P. Minka. Estimating a Dirichlet distribution. Technical report, 2009.

[5] Kimio Morimune. Volatility models. *The Japanese Economic Review*, 58(1):1–23, 2007.

[6] Nikolaos Vlastakis and Raphael N. Markellos. Information demand and stock market volatility. *Journal of Banking & Finance*, 36(6):1808 – 1821, 2012.