様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月 １３ 日現在

機関番号：６２６０３
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ０１５９７

研究課題名（和文）テキストデータからの情報抽出を利用した金融時系列予測

研究課題名（英文）Financial time series forecast using information extracted from text data

研究代表者
川崎　能典（Kawasaki, Yoshinori）

統計数理研究所・モデリング研究系・教授

研究者番号：７０２４９９１０

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：テキストデータの系列（例えば日々の新聞記事）から、金融資産の変動（ボラティリティ）に関連しそうな「話題＝トピック」の動向を時系列的に抜き出し、それをボラティリティ予測モデルに組み込んで予測を改善する統計的モデリング法について研究を行った。とりわけ、日次・週次・月次の多重時間スケールを明示的に取り込む提案を行った。その有効性を模擬予測で実証的に研究した結果、実験総ケース全体の35%程度で提案手法が優った。

研究成果の学術的意義や社会的意義
テキストデータ解析の方法自体は潜在ディリクレ分配法を筆頭にさまざまな研究がなされているが、多くは時点を固定した分析であり、テキスト系列からの動的な情報抽出に関する研究は多くない。本研究は経済統計学のテーマ設定で、金融資産の変動性予測の問題とテキスト解析を結びつけて考えたが、時間軸に沿ってテキストデータが流れてくる状況で、そこから抽出された情報を別の予測目的に結びつける問題は他にもあると思われ、今後異分野での展開が期待できる。

研究成果の概要（英文）：We studied a statistical modeling method to improve volatility forecasts by extracting trends in "topics" that may be related to changes in financial assets from a series of text data (e.g., daily newspaper articles) and by incorporating them into a volatility forecasting model. In particular, we proposed a method to explicitly incorporate daily, weekly, and monthly multiple time scales. The effectiveness of the proposed method was empirically studied through simulated forecasts, and the proposed method was superior in about 35% of the total experimental cases.

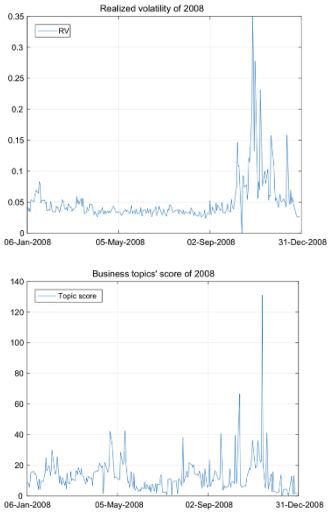
研究分野：経済統計学、統計科学

キーワード：テキストデータ　高頻度データ　動的トピックモデル　時系列モデル　多重スケール　ボラティリティ予測　実現ボラティリティ

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景



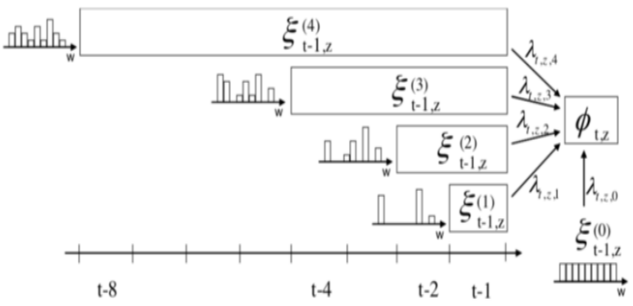
定量的分析のために、大量のテキストデータから情報抽出を行おうという動きは、こと金融データ分析との関連で考えても、過去 10 年以上にわたって研究成果が次々と発表されてきた。しかしその多くは、Google SVI に代表されるような、特定のキーワードとその出現頻度に依存した分析になっていた。テキスト（例えばある日のオンラインニュース記事）は複数のトピック＝話題から構成されていると考え、個々のトピックはその背後にある単語分布の特徴が異なる、というタイプのモデル化をすれば、あるテキストが現在どのようなトピックから構成されているかを有限混合分布の問題として推定できる。研究代表者は Morimoto and Kawasaki (2017, *Asia-Pac. Financ. Mark.*)で、トピックモデルを時系列方向に拡張した手法のひとつである多重スケール動的トピックモデル(Multiscale Dynamic Topic Model, 以下 MDTM)に着目して、オンラインニュース記事からトピック指数時系列を抽出し、どのトピックが金融資産の日次ボラティリティ予測の役に立つかどうかを、実証的に分析した結果を報告した。右の図は、上が 2008 年のとある資産の日次ボラティリティ、下が推定されたトピック時系列のひとつであるが、ボラティリティが激しく上がっている時期に関する情報を有しているように思われる。

２．研究の目的

MDTM では、あるトピックを特徴付ける単語分布が複数の時間スケールに依存する形でモデル化を行う。換言すればどの程度のウィンドウ幅で単語分布を集約するかはラグに依存し、長短取り混ぜて混合する。本研究の目的は、オリジナルの MDTM での実装以外に、金融時系列(特にボラティリティ)の予測に適した定式化を、ひとつのアイディア(後述)に沿って実装し、その有効性を模擬予測等の実証分析によって検証することである。

３．研究の方法

MDTM 自体は Iwata et al. (2010, *Proceedings of 16th ACM SIGKDD*)がオリジナルの提案である。上記研究の目的に述べた内容は、MDTM における時間依存性に関する定式化の探索、と換言できる。右の図は Iwata et al. (2010)からの引用であるが、複数の時間スケールを持つ単語分布 $(\xi_{t-1,z}^{(0)}, \ldots, \xi_{t-1,z}^{(4)})$ が重みづけられて、トピック z の単語分布 $\phi_{t,z}$ が生成されることを示している。原典の実装では、ラグを k としたときに時間のスケールを 2^k で広げていく定式化が採用されている。この定式化が一定の合理性を有することは、以下のように直観的に説明できる。もし当該トピックを賑わす事象が発生している場合、直近のウィンドウで集約した単語分布が支配的であるかもしれない。あるいはそうしたことがない場合は、より長いウィンドウで集約した方が、当該トピックに対応させる単語分布としては平均的でバランスが良いかもしれない。



ボラティリティを予測する時系列モデルの文脈にも、類似の発想がある。Corsi (2009, *J. Financial Econ.*) の Heterogeneous Autoregressive (HAR)モデルである。HAR モデルは、分・秒程度の高頻度金融データが利用可能であることを前提に、日次のボラティリティをいわゆる実現ボラティリティ(realized volatility)で推定し、日次ボラティリティの 1 期先を、日次ボラティリティの 1 期ラグ、週次ボラティリティの 1 期ラグ、月次ボラティリティの 1 期ラグに線形回帰するモデルである。ここで週次、月次の集約は営業日ベースで行われる。HAR モデルは極めて単純な方法ながら、既存ボラティリティ予測モデルより優れた予測を生み出すことが経験的に観察されている。

HAR モデルにおける heterogeneous の意味は MDTM における multiscale と殆ど同義である。このことから類推すれば、トピック時系列の抽出にあたって、ラグ構造の定式化を HAR モデルに倣ったタイムホライズン、即ち日次、週次、月次で集約するモデルは、とりわけ金融時系列予測においては有効ではないかと予想される。

４．研究成果

実証分析に用いるデータ

予測対象となる金融資産データは日経 NEEDS の高頻度データ(日経平均、日経平均 300、TOPIX、東証株価指数 33 業種のうち電気機器・輸送用機器・銀行業)で、その標本期間は 2008 年 1 月から 2012 年 12 月とした。一方、テキストデータは同時期のロイタージャパンのニュースデータを Web スクレイピングして抽出したものを用いた。

分析の枠組み

MDTM の枠組みにおける単語分布を、計量ファイナンスを意識して日次、週次、月次に対応するタイムスケールとする方法を実装した。具体的には、オリジナルの MTDM でタイムスケールを 2,6,21 と取ったモデル(それぞれ 1,5 20 営業日に対応)を推定可能なプログラムを作成した。この定式化を、Heterogeneous MDTM (H-MDTM)と呼ぶことにする。比較対象としてのオリジナルの MDTM は、平成 31 年度／令和元年度の分析では 2 に固定していたが、令和 3 年度には時間スケールを 9 まで(つまり 2 の 8 乗まで)カバーして比較分析を行った。なお、いずれの定式化においてもそれぞれ最大 20 までトピックを抽出し、トピックスコア時系列(こちらも 20 系列)を作成した。なお、手法を理解し、自らプログラミングするための要素技術としては、マルコフ連鎖モンテカルロ法を利用したベイズ推定、不動点反復法等の知識が必要である。

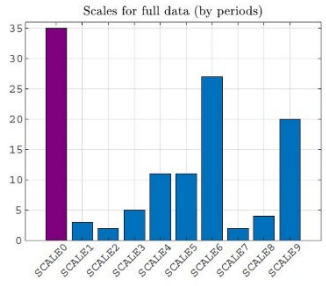
日次ボラティリティ予測のためのモデルは、HAR モデルとその変種を中心に 6 種用意し、全てにトピックスコア時系列を説明変数としてひとつ追加する。ここでいう変種には、Bollerslev et al. (2016, *J. Econom.*) に倣い、高頻度データから計算した実現ボラティリティ (RV) に対し HAR モデルの日次パートの係数を、実現 quarticity (RV の 4 次モーメント版) に依存させた時変係数とするモデルを含めている。

分析結果 1：誤差関数の比較

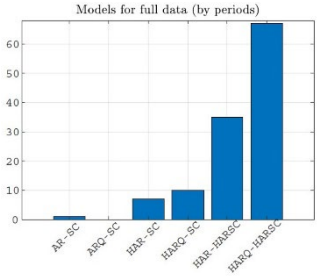
予測手法は固定区間シフト型と区間拡大型の 2 種類、予測評価は Patton (2011, *J. Econom.*) の誤差関数で MSE と QLIKE の 2 通りで分析した。結果については、各ケースでベストなトピック時系列をひとつに決めた上での話であるが、Kawasaki and Morimoto (2021, JAFEE 夏季大会予稿集)に報告されている。総じて、QLIKE を評価関数に取った時には、予測区間の設定が固定区間シフト型であろうと区間拡大型であろうと取り上げた全ての金融資産で大きな差は認められないが、MSE を評価関数に取って固定区間シフト型で予測した場合に、日経平均、TOPIX、東証 33 業種輸送用機器および銀行業では予測誤差の顕著な減少が観察された。

分析結果 2：モデル信頼集合による比較

令和2年度までの分析結果はPattonの誤差関数の値を経験比較することにとどまっていたが、令和 3 年度は誤差関数の「差が有意かどうか」を検証するために、Hansen et al. (2011, *Econometrica*)のモデル信頼集合(Model Confidence Set, MCS)の枠組みに基づく比較を行った。この方法では、誤差関数の差の分散をブートストラップで推定することで予測誤差どうしである種の同等性検定を繰り返し、多段階でベストモデルを決める方法である。決勝で複数のモデルが生き残ることもあり、その意味でここでのベストは複数一位(tie)も許している。



全ての分析ケースでの「勝利数」を見ると、H-MDTM(右上図中便宜上 SCALE0 と記載、紫のバー)が 35%で最多だったが、オリジナルの MDTM で時間スケール 6 が 27%、時間スケール 9 が 20%で続いた。時間スケール 6 ということは単語分布のラグとして 1,2,4,8,16,32 を考えることに相当し、MDTM も十分長いラグを確保すれば H-MDTM に遜色ない精度で予測に貢献するトピック時系列を抽出する可能性がある。別の見方をすれば、H-MDTM は他を圧倒するほどの予測性能とまでは言えず、モデル候補は大きめにとっておいた方が良い可能性がある。



また、全ての分析ケースを選択されたベスト時系列モデルごとに整理すると、HARQ-HARSC と HAR-HARSC が殆どを占めていた(右下図)。HARQ-HARSC とは、実現 quarticity で日次 RV のラグ項係数を時変化した上で、トピックスコア時系列を説明変数として追加する際にも HAR モデル同様の日次、週次、月次ラグを取り込むモデルであり、HAR-HARSC は日次 RV のラグ項係数を固定としたものを指す。トピックスコアの多重スケール化が予測改善の主要因となっていることがわかる。

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　0件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Kawasaki, Y. and Morimoto, T.	４．巻 -
２．論文標題 Volatility Forecasting with the Heterogeneous AR-type Multiscale Dynamic Topic Model	５．発行年 2021年
３．雑誌名 2021年度JAFEE夏季大会予稿集	６．最初と最後の頁 12-21
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計9件（うち招待講演　0件／うち国際学会　4件）

１．発表者名 Kawasaki, Y. and Morimoto, T.
２．発表標題 Volatility Forecasting with the Heterogeneous AR-type Multiscale Dynamic Topic Model
３．学会等名 日本金融・証券計量・工学学会
４．発表年 2021年

１．発表者名 Kawasaki, Y. and Morimoto, T.
２．発表標題 On a HAR-type Specification in Dynamic Topic Model and its Application in Volatility Forecasting
３．学会等名 11th CEQURA Conference 2020 on Advances in Financial and Insurance Risk Management（国際学会）
４．発表年 2020年

１．発表者名 Kawasaki, Y.
２．発表標題 Examining the Effects of Expanded Trading Hours Using High Frequency Data in Finance
３．学会等名 Joint Statistical Meeting (JSM) 2020（国際学会）
４．発表年 2020年

１．発表者名
貝淵響，川崎能典，Gilles Stupfler

２．発表標題
A bias-reduced GARCH-EVT approach for financial risk estimation

３．学会等名
2020年度統計関連学会連合大会

４．発表年
2020年

１．発表者名
川崎能典

２．発表標題
RS-Decomp

３．学会等名
2020年度R研究集会(統計数理研究所共同利用研究集会「データ解析環境Rの整備と利用」)

４．発表年
2020年

１．発表者名
Kawasaki, Y. and Morimoto, T.

２．発表標題
Forecasting Financial Market Volatility Using a Dynamic Topic Model

３．学会等名
62nd ISI World Statistics Congress, Kuala Lumpur, Malaysia（国際学会）

４．発表年
2019年

１．発表者名
Kaibuchi, H. and Kawasaki, Y.

２．発表標題
A novel GARCH-EVT approach dealing with bias and heteroscedasticity

３．学会等名
CEQURA Conference 2019 on Advances in Financial and Insurance Risk Management, Munich, Germany（国際学会）

４．発表年
2019年

１．発表者名
貝淵響，川崎能典

２．発表標題
A novel GARCH-EVT approach to VaR estimation dealing with bias and heteroscedasticity

３．学会等名
2019年度中之島ワークショップ「金融工学・数理計量ファイナンスの諸問題2019」

４．発表年
2019年

１．発表者名
川崎能典

２．発表標題
テキスト系列からの動的トピックの抽出によるボラティリティ予測

３．学会等名
リスク解析戦略研究センター第7回金融シンポジウム「金融が直面する新環境への対応と方法論II」

４．発表年
2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関