[S4-P28]

掲示板テキストから得られるセンチメントを利用した日経平均株価ボラティリティ予測

中島秀太，　欅惇志，　渡部敏明，小町守（一橋大）

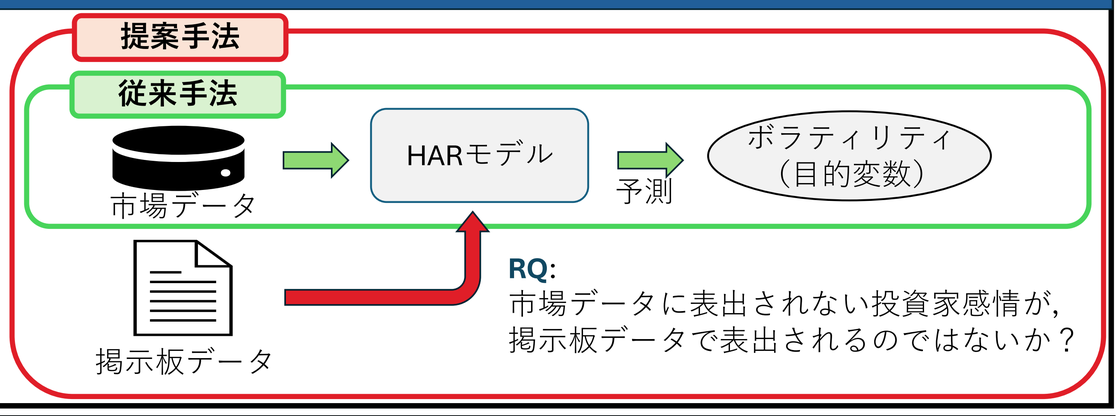
1. 概要

ボラティリティ予測

- ボラティリティとは，「資産価格変化率の分散ないしは標準偏差」.
- 予測タスクは，投資家のリスクを考慮した投資判断等に貢献.
- 近年は，新聞/掲示板テキストデータの活用も検討.

本研究の概要

- 掲示板から得られたポストにセンチメント付与を行い，変数を作成.
- センチメント変数を利用したモデルで，ボラティリティ予測を行う.



2. 手法

従来手法

- 市場データより，RVを算出.
- RVを説明変数とし，HARモデルによる予測を行う.

【RV (Realized Volatility)】

ボラティリティの代理変数であり，目的変数.
t－1からt時点までの日中リターンrを用い，
右式で計算されたRVを用いる.

$$RV_t = \sum_{i=1}^{n} r^2_{t-1+(\frac{i}{n})}$$

（日経平均の場合，昼休み除く9時～15時．5分間隔）

※nは日中リターンの時間間隔を表す.

【HARモデル】（Heterogenous Auto Regressive）

周期の異なる日次/週次/月次のRVを説明変数とし，
最小二乗法で推定されるモデル. [1]

$$\mathbf{logRV}_t = \alpha + \beta_d \mathrm{logRV}_{t-1} + \beta_w \mathrm{logRV}_{t-5:t-1} + \beta_m \mathrm{logRV}_{t-22:t-1} + \delta \bar{R}_{t-1} + \epsilon_t$$

※$\bar{R}_{t-1} = \min\{R_{t-1}, 0\}$，$R_{t-1}$ はt－1期におけるリターンを表す（ボラティリティの非対称性）.

RV_{t-1}，$\mathrm{RV}_{t-5:t-1}$，$\mathrm{RV}_{t-22:t-1}$はそれぞれ日次・週次・月次のRVを表す（ $\mathrm{RV}_{t-n:t-1} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{RV}_{t-i}$ ）.

提案手法

- ①～③：掲示板データより，センチメント変数を作成.
- ④：センチメント変数を説明変数として追加した，HARモデルを拡張したモデルによる予測を行う.

① 掲示板データのノイズ除去

- データ内には，株価動向と無関係のポスト（ノイズ）も散見.
- GPT-3.5 Turboを利用し，ノイズ分類(2値)を実施.

【使用プロンプト】

post: {text}
If the post is related to finance,
please respond with ‘0'; otherwise, respond with ‘1'.
Examples of each type of post are provided below.
（以下略，few-shot事例）

【人手評価との比較】

Precision	0.438
Recall	0.984

※某1日のポスト2,581件に関するアノテーション結果を比較.

② 掲示板データのセンチメント付与

- 非ノイズのポストを対象とし，センチメント付与を行う.
- GPT-3.5 Turboを利用し，「市況感について，強気，弱気，不明，のいずれに該当するポストか」分類(3値)を実施.

【使用プロンプト】

post: {text}
the closing price of the Nikkei Stock Average : {value}
Please return '1' if the post is bullish on the outlook,
'-1' if the post is bearish on the outlook,
and '0' if the outlook is unknown.
Examples of each type of post are provided below.
（以下略，few-shot事例）

【GPT評価の分布】

Bullish	20.60%
Bearish	29.25%
Unknown	50.15%

※全期間のポスト226,655件に関するアノテーション結果.

③ ポストの極性を利用したセンチメント変数の作成

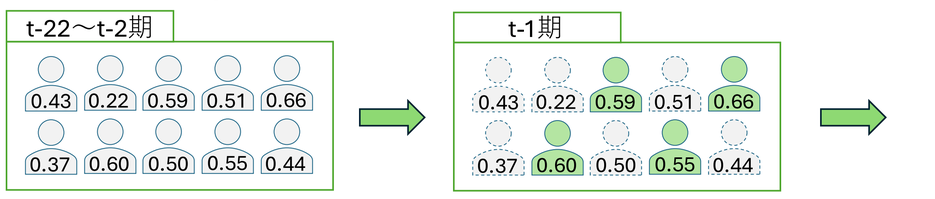
- 各日における強気（Bullish）．弱気（Bearish）のポストの割合を集計し，変数 SF(sentiment factor) を作成

$$\boldsymbol{SF}_t = \frac{Post_{Bull,t} - Post_{Bear,t}}{Post_{All,t}}$$

※ $Post_{All,t}$は，t日におけるノイズ除去後の全ポスト数を表す.
$Post_{Bull,t}$及び$Post_{Bear,t}$は， $Post_{All,t}$中における強気，弱気の投稿数を表す.

③' 優秀な投資家のポストに限定したセンチメント変数の作成

- t-1期において「優秀な投資家」のポストを収集し，SFを作成.
- t-22期からt-2期(1ヶ月) を対象とし，ポストの「市況感予想」のAccuracy = 0.55以上の者を「優秀な投資家」として定義.



④ センチメント変数を利用したHAR_Sモデルによる予測

- 拡張したHAR_Sモデルによる予測を行う.

【HAR_Sモデル】

HARモデルを拡張し，センチメント変数を追加した提案モデル.

$$\mathbf{logRV}_t = \text{HAR 右辺} + \gamma_{SF,d} SF_{t-1} + \gamma_{SF,w} SF_{t-5:t-1} \gamma_{SF,m} SF_{t-22:t-1}$$

$$\boldsymbol{SF}_{t-n:t-1} = \frac{1}{n}\sum_{i=1}^{n} \frac{Post_{SF,i}}{Post_{All,i}}$$

3. 実証分析

データ

- **日経NEEDSティックデータ（2016/01/04－2022/09/30）**
  - 日経225株価指数の5分ごとの価格データ
- **Yahoo!ファイナンス掲示板データ（2015/12/30－2022/09/29）**
  - 特定の株式等の話題について情報交換できる匿名掲示板
  - 「日経平均株価」スレッドにおけるポストデータ
  - 各日300件をサンプリング(全投稿は各日平均4700件)

期間

- 2016/02/03-2022/09/30（1626期，営業日のみ土日祝等除く）
- 予測期間（経済的危機局面の前後で区分）
  - コロナ禍前：2018/2/19-2019/12/30 (456期)
  - コロナ禍後：2020/1/6-2022/9/30 (669期)

比較モデル

- 以下の3つの予測モデルで精度を比較する.
  - HARモデル（ベースライン）
  - HAR_Sモデル（全投資家）
  - HAR_Sモデル（優秀な投資家）

方法

- t期の予測について，t-500期からt-1期内のデータから推定されるモデルを用い，パラメータを都度推定.
- 予測誤差の評価指標として，MSE及びQLIKEを使用. [2]

$$MSE = \frac{1}{N}\sum_{t=1}^{N}(\hat{\sigma}^2_{t+1|t} - \sigma^2_{t+1})^2 \qquad QLIKE = \frac{1}{N}\sum_{t=1}^{N}(\log \hat{\sigma}^2_{t+1|t} + \frac{\sigma^2_{t+1}}{\hat{\sigma}^2_{t+1|t}})$$

※$\hat{\sigma}^2$はボラティリティの予測値，σ^2はRVを表す．Nは評価データの期数を表す.

- 評価指標のモデル間の差が統計的に意味があるかどうか、MCS (Model Confidence Set) によって検証. [3]

結果

- コロナ禍前の期間において，HAR_Sモデル(提案手法)がHARモデル(ベースライン)よりも高精度.
(HAR_S(優)モデルは，QLIKEで10%有意)
- 今後は，SFの改良等が考えられる.
(ボラティリティに即したリスク指標となる変数の作成が目指される)

期間	モデル	MSE	QLIKE
全期間	HAR	2.350	1.045
	HAR_S(全)	3.243	1.054*
	HAR_S(優)	2.351	1.046
コロナ禍前	HAR	0.249	0.762*
	HAR_S(全)	0.245	0.770*
	HAR_S(優)	0.249	0.760
コロナ禍後	HAR	3.783	1.239
	HAR_S(全)	5.286	1.247*
	HAR_S(優)	3.784	1.240*

*は10%有意水準でMCSに含まれないモデル.

[1] Bekaert Geert and Hoerova Marie. The vix, the variance premium and stock market volatility. Journal of econometrics, Vol. 183, No. 2, pp. 181–192, 2014.
[2] Patton Andrew J “Volatility forecast comparison using imperfect volatility proxies”,Journal of Econometrics, Vol.160,No.1,pp. 246-256, 2011
[3] Hansen Peter R, Lunde Asger, and Nason James M. The model confidence set. Econometrica, Vol. 79, No. 2, pp. 453–497, 2011.