混合回帰モデルに基づく企業情報を考慮した株価変動の要因分析に関する研究

1X15C091-0　福山武志
指導教員　　後藤正幸

1　研究背景と目的

株式の発行者である企業の多くは，自社の株価の上昇によるブランド力の向上や経営の安定を期待している．株式を発行している企業の責務として，投資家の投資判断の参考になる事業内容や財務状況などの情報を定期的に開示する必要があるが，近年はそのような情報がテキストデータの形で流通し，様々な影響を与えるようになっている．IR 活動の一環として開示している様々なプレスリリースや投資家向け情報がインターネット上で記事データとして流通し，投資家の投資判断に影響を与えている．そのため，どのような記事情報が株価変動に影響を与えのかを明らかにすることで，企業側の情報発信の一助になると考えられる．そこで本研究では，企業の公開情報の作成を支援することを目的として，株価変動の要因分析モデルの構築を目指す．

本研究では，公開情報の文書のトピックは多くの投資家の投資判断に影響を与えると考える．例えば，災害に関するトピックの文書が開示された場合，台風による損失が想定される企業の株は売却され，株価は下落すると考えられる．このような文書のトピックの投資判断への影響は，株価の変動という観点から定量化することが期待される．そこで本研究では単語が文書の内容を構成する要素であると仮定し，pLSA[1] によって得られたトピックを説明変数とし，株価変動に及ぼす影響を分析するための回帰モデルを考える．ここで，文書のトピックが企業の株価変動に及ぼす影響は，企業の特性ごとに異なると考えられ，単一の回帰モデル [2] では，精度の高いモデルの構築が困難である．そのため本研究ではさらに，混合回帰モデル [3] を導入して，文書トピックと共に，企業特性による差異を考慮した要因分析モデルを提案する．提案モデルによって，企業の特性を考慮したうえでトピックが株価に与える影響を分析することが可能となる．また，実データの分析により，提案モデルの適用可能性を検討する．そして，得られた株価変動の要因から，各企業がどのような内容の公開情報を作成すべきかを検討する．

2　準備

2.1　分析対象データ

投資家は，企業の公開情報や新聞記事を投資判断の材料としているため，本研究では，学習データとしてこれらの文書データを用いる．ここで，株価の変動を表す変数として，株価前日比率を以下の式 (1) で定義する．そして，この値が大きいほど記事が投資判断の「買い」に影響を与えたと解釈を行う．

$$\text{株価前日比率} = \frac{\text{新聞記事発行日の終値}}{\text{新聞記事発行前日の終値}} \times 100 \qquad (1)$$

2.2　混合回帰モデル

混合回帰モデルとは目的変数 y と説明変数 $\boldsymbol{x}$ の線形構造の背後に潜在クラスを仮定したモデルである．このモデルはそれぞれの潜在クラスに対して異なる回帰モデルを仮定しており，各データに対する偏回帰係数は，それらの混合により表現される．いま，K 個の混合回帰モデルで用いる潜在クラスを仮定し，k 番目の部分回帰として，$g_k(\cdot)$ を平均 $\boldsymbol{\beta}_k^\top \boldsymbol{x}$，分散 σ_k^2 の正規分布とし，$\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k^\top, \sigma_k^2)^\top$ と表記する．このとき，混合回帰モデルは，混合割合 $\pi_k(\boldsymbol{v}, \boldsymbol{\alpha})$ が補助変数 $\boldsymbol{v}$ に依存するモデルとなっており，式 (2) で表される．

$$h(y|\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{\phi}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{v}, \boldsymbol{\alpha}) g_k(y|\boldsymbol{x}, \boldsymbol{\theta}_k) \qquad (2)$$

$$\pi_k(\boldsymbol{v}, \boldsymbol{\alpha}) \geq 0 \quad \text{and} \quad \sum_{k=1}^{K} \pi_k(\boldsymbol{v}, \boldsymbol{\alpha}) = 1 \qquad (3)$$

ここで，$\boldsymbol{\phi} = (\boldsymbol{\alpha}^\top, \boldsymbol{\theta}_k^\top)^\top$ は混合回帰モデルのすべてのパラメータを表すベクトルであり，$\boldsymbol{\alpha}$ は補助変数に対するパラメータである．また，このモデルの混合割合 $\pi_k(\boldsymbol{v}, \boldsymbol{\alpha})$ は制約式 (3) を満たすように関数を設定する．

2.3　pLSA(確率的潜在意味解析)

pLSA(確率的潜在意味解析) は，潜在クラスモデルの一つである．文書データに適用した場合，単語と文書の間に潜在クラスを仮定し，それらの共起関係を潜在クラスによる条件付確率分布で表したモデルである．ここで，文書集合を $\mathcal{D} = \{d_m : 1 \leq m \leq M\}$，単語集合を $\mathcal{W} = \{w_g : 1 \leq g \leq G\}$，文書トピックを表わす潜在クラス集合を $\mathcal{U} = \{u_c : 1 \leq c \leq C\}$ と定義する．このとき，pLSA の確率モデルは以下の式 (4) で表される．各パラメータは EM アルゴリズムを用いて推定する．

$$P(d_m, w_g, u_c) = P(u_c)P(d_m|u_c)P(w_g|u_c) \qquad (4)$$

3　提案モデル

3.1　概要

新聞記事のトピックが株価に与える影響は，業種などの企業の特性によって異なる．そこで，説明変数を pLSA を用いて得られたトピックの出現確率，目的変数を株価前日比率，補助変数を企業の基本情報とした混合回帰モデルを考える．本研究では，企業特性を表す基本情報として，企業の業種や従業員規模を用いる．

3.2　提案モデルの定式化

潜在クラス集合を $\mathcal{Z} = \{z_k : 1 \leq k \leq K\}$，$l$ 番目の文書のトピック分布 $P(\boldsymbol{u}|d_l) = (P(u_1|d_l), P(u_2|d_l), \ldots, P(u_C|d_l))^\top (c = 1, 2, \ldots, C)$ を用いて，説明変数 $\boldsymbol{x}_l = (x_{l0}, x_{l1}, x_{l2}, \ldots, x_{lC})^\top$ を $x_{l0} = 1$，$x_{lc} = P(u_c|d_l)$ で定義する．l 番目の文書による株価前日比率を y_l とする．混合回帰モデルは各潜在クラスにおける確率密度関数 $P_k(y_l|\boldsymbol{x}_l)$ の線形結合によりモデル化される．このとき，回帰の誤差が正規分布に従うと仮定したとき，潜在クラス z_k における y_l の確率密度関数は，分散 σ_k^2 を用いて式 (5) で表される．また，z_k における回帰モデルは式 (6) で表される．

$$P_k(y_l|\boldsymbol{x}_l) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{ -\frac{(y_l - f_k(\boldsymbol{x}_l))^2}{2\sigma_k^2} \right\} \qquad (5)$$

$$f_k(\boldsymbol{x}_l) = \sum_{c=0}^{C} \beta_{kc} x_{lc} \qquad (6)$$

次に，補助変数として用いる l 番目の文書に対応する企業の基本情報を表す変数ベクトルを $\boldsymbol{s}_l = (s_{l1}, s_{l2}, \ldots, s_{lJ})^\top$，$s_{lj}$ を l 番目の文書の j 番目の基本情報とする．また，j 番目の基本情報は N_j 種類の要素をもつカテゴリカル変数であり，s_n^j を j 番目の基本情報の n 番目の要素，j 番目の基本

表 2：各潜在クラスの回帰係数推定値

業種	自動車・その他	自動車・情報	自動車・その他	自動車・電気	銀行・その他	小売・銀行	電気・情報	医薬・化学	情報・食品	情報・化学
従業員規模	A クラス	A クラス	A クラス	B クラス	B クラス	B クラス	B クラス	C クラス	C クラス	C クラス
$\hat{P}(z_k)$	0.047	0.226	0.061	0.249	0.036	0.115	0.046	0.140	0.050	0.030
切片	100.0	100.0	100.0	100.1	100.1	100.1	100.0	100.2	100.1	100.2
人工知能に関するトピック	**4.17**	**3.58**	**4.36**	**−0.12**	**−0.96**	**2.30**	**2.13**	**0.38**	**3.64**	**−0.44**
企業の不正に関するトピック	-1.37	-2.68	-3.30	1.19	2.95	-0.34	1.70	-0.66	1.85	3.37
企業の表彰に関するトピック	4.88	2.11	1.50	4.27	-4.76	1.79	2.14	1.99	-1.04	3.12
上層部の人事異動に関するトピック	0.43	1.05	2.78	-2.23	7.44	-1.56	-2.88	-4.89	-5.90	0.77
テクノロジーに関するトピック	-4.07	-2.09	-4.83	1.90	-2.30	0.89	-0.67	-5.82	-2.51	-1.18
災害に関するトピック	-4.10	-2.10	-4.27	-2.36	-0.91	-0.65	-2.23	-3.14	-1.56	-6.28
自動運転に関するトピック	0.95	0.66	1.05	0.67	-1.01	-0.36	1.72	-2.52	-2.25	-2.26
企業の業績に関するトピック	-1.64	-3.48	-1.03	-1.45	0.95	0.08	-3.27	6.17	-0.28	1.88
株価変動に関するトピック	0.06	1.14	0.61	-2.81	1.05	-2.70	-2.41	0.82	0.86	-0.07
米国の政策に関するトピック	-3.28	-2.12	-3.04	-2.34	0.16	-2.66	-1.16	-3.58	2.97	-0.13
日銀の政策に関するトピック	-0.03	3.58	1.44	-0.57	-2.46	-1.46	-0.12	2.72	5.42	5.58
携帯会社の料金プランに関するトピック	-3.98	-3.67	-3.04	-3.57	-2.52	-0.45	-4.46	-8.47	-6.43	-6.96
他業界に関するトピック	5.05	0.59	7.43	1.79	-4.10	-1.69	3.45	6.38	2.29	2.83
雇用政策に関するトピック	-0.55	-1.87	-0.42	3.94	10.76	4.20	2.40	12.91	11.05	9.00
企業買収に関するトピック	1.65	-0.56	1.31	1.13	3.14	-1.52	-1.41	1.87	5.15	0.95

情報 $(1 \leq j \leq J)$ の要素集合を $\mathcal{S}^j = \{s_n^j : 1 \leq n \leq N_j\}$ とする．このとき，l 番目の文書の生成確率は，式 (7) で表される．$\delta(a, b)$ は a と b が一致していれば 1，さもなければ 0 とする指示関数とする．

$$P(y_l, \boldsymbol{x}_l, \boldsymbol{s}_l) = \sum_{k=1}^{K} P(z_k) P_k(y_l|\boldsymbol{x}_l) \prod_{j=1}^{J} \prod_{n=1}^{N_j} P(s_n^j|z_k)^{\delta(s_n^j, s_{lj})} \quad (7)$$

このモデルのパラメータは EM アルゴリズムを用いて対数尤度関数を局所最大化するように推定される．

4 実験

提案モデルの有用性を示すため，日本経済新聞朝刊の新聞記事データと Yahoo! Finance から取得した株価データに提案手法を適用して分析を行う．また，企業の基本情報は Yahoo! Finance の企業情報から取得した．

4.1 実験条件

分析対象企業として，日経 225 に含まれる 8 業種 45 社の企業データ及び新聞記事を用いた．データ期間は 2017 年 10 月 1 日から 2018 年 9 月 30 日，総新聞記事数は 7,906 件，新聞記事内の単語の総数は 1,574,291 個，単語の種類数は 35,286 個である．新聞記事の検索エンジンは日経テレコンを用い，人事・訃報記事，数表のみの記事，見出しのみの記事，スポーツ面の記事は除外した．企業の基本情報は，業種，従業員規模を用いた．従業員規模は，従業員数が 250,000 人〜400,000 人を A クラス，50,000 人〜250,000 人を B クラス，0 人〜50,000 人を C クラスとした．pLSA，混合回帰モデルの潜在クラス数はそれぞれ $C = 15$，$K = 10$ とした．

4.2 実験結果と考察

対象データを pLSA に適用して得られた結果を表 1 に示す．ただし，各トピックで特徴的な単語は太字とした．

表 1：各トピックの Top5 の単語

トピック	解釈	Top1	Top2	Top3	Top4	Top5
u_1	人工知能に関するトピック	**開発**	研究	**データ**	自動	**情報**
u_2	企業の不正に関するトピック	株主	検査	取締役	不正	総会
u_3	企業の表彰に関するトピック	**賞**	位	日経	ゲーム	部門
u_4	上層部の人事異動に関するトピック	会長	**就任**	長	出身	**役員**
u_5	テクノロジーに関するトピック	サービス	**決済**	店舗	**ネット**	スマート
u_6	災害に関するトピック	工場	生産	**停止**	**被害**	影響
u_7	自動運転に関するトピック	電池	**EV**	開発	**技術**	生産
u_8	企業の業績に関するトピック	**利益**	販売	**営業**	毎年	売上
u_9	株価変動に関するトピック	**株価**	**平均**	投資	日経	銘柄
u_{10}	米国の政策に関するトピック	**米国**	中国	**関税**	交渉	輸出
u_{11}	日銀の政策に関するトピック	金融	**金利**	融資	証券	発行
u_{12}	携帯会社の料金プランに関するトピック	通信	**契約**	楽天	**料金**	スマホ
u_{13}	他業界に関するトピック	会長	鈴木	**住友**	**三井**	本社
u_{14}	雇用政策に関するトピック	**女性**	社員	働き	**改革**	取り組み
u_{15}	企業買収に関するトピック	**買収**	投資	**子会社**	出資	武田

表 1 の各トピックの Top5 の単語を見ると，pLSA を用いて単語からトピックを解釈することができる．例えば，u_1 に着目すると，上位に「開発」「データ」「情報」という単語が出現していることから,「人工知能に関するトピック」と解釈できる．

次に，pLSA を用いて得られたトピック分布とトピックの回帰係数を表 2 に示す．また，表 2 における $\hat{P}(z_k)$ は混合割合の推定値とする．ただし，行を新聞記事のトピック，列を企業特性クラスとした．

表 2 の「人工知能」に関するトピックの回帰係数に着目すると，企業特性によってその値が異なることがわかる．例えば，人工知能を用いた技術に関するトピックの記事が公表された場合,「自動車・その他」では株価が上昇しており,「銀行・その他」では株価は下落している．よって，投資家は，自動車業で人工知能を用いた技術を導入すると自動車の性能を高め，企業は発展すると期待している一方で，銀行業で人工知能を用いた技術を導入することは企業に有益でないと考えている可能性を指摘できる．また，表 2 の業種が「銀行・その他」，従業員規模が「B クラス」の列の回帰係数に着目すると，雇用政策に関するトピックの記事が公表された場合，株価は上昇しているが，テクノロジーに関するトピックの記事が公表された場合，株価が下落している．このことから，雇用政策に関するトピックの記事が株価に好影響を与える可能性を示唆している．

5 考察

今回，提案モデルを実データに適用し，企業の特性を考慮した株価変動の要因分析を行った．提案モデルでは pLSA と混合回帰モデルの組み合せにより，結果を分析的に解釈可能な予測モデルが構築できた．しかし，分析対象データの期間によって，本研究の提案モデルを用いて得られる結果は変わってくる．そのため，適切な分析対象期間の検討が必要となる．

6 まとめと今後の課題

本研究では，企業の公開情報が投資家の投資判断に影響を及ぼす要因を抽出するために，新聞記事から企業特性を考慮した株価変動の要因分析を可能とするモデルを提案した．そして，実際の新聞記事データを用いて提案モデルの有用性を示した．

今後の課題として，最適な潜在クラス数の決定や，テストデータを用いたモデルの定量的な評価が挙げられる．

参考文献

[1] T. Hoffman, “Probablistic Latent Semantic Analysis, ” *Proc. of UAI799*,pp.289–296,1999.

[2] Bishop, C. M.,*Pattern Recognition and Machine Learning*, Springer, 2006.

[3] Grun, B., and Leisch, F.,”FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters,” *Journal of Statistical Software*, Vol. 28, Issue 4, pp. 1–35, 2008.