HUB-FS Working Paper Series

FS-2022-J-001

ニュースを用いた金融市場分析のための マルチレベル教師ありトピックモデル

姫野 知也

三菱 UFJ 国際投信

横内 大介

一橋大学大学院経営管理研究科

First version: 2022 年 8 月 3 日

ニュースを用いた金融市場分析のための
マルチレベル教師ありトピックモデル

1　はじめに

近年，金融市場の分析において，伝統的に用いられてきた財務データ，市場データ，経済指標などのデータに加えて，ニュースデータ，POS データ，企業間ネットワークデータなどといった，非伝統的なオルタナティブデータの活用が急速に拡大しており，業界のトレンドとなっている．

ことニュースに関しては，ファイナンス領域において，兼ねてより金融市場との関係について盛んに研究が行われている．かつては，ニュースの金融市場へのインパクトは限定的であるとの見方が主流であった．代表的な先行研究として，Roll (1988) は，ニュースの有無によってサンプル期間を分割しても，サンプル間でファクターモデルの株価リターンに対する説明力にほとんど違いがみられないことを示した．Mitchell and Mulherin (1994) と Berry and Howe (1994) は，ニュース件数と，株価の変動および市場の取引量の相関関係は無いかそれほど強くないことを実証した．

しかし，近年になって，理論的背景やニュースの内容を踏まえた詳細な分析が行われ，金融市場分析におけるニュースの重要性を示唆する研究が蓄積している．ニュース件数を用いて，Tetlock (2010) はより詳細な分析を行っており，銘柄のサイズおよび流動性別にニュース件数と株価リターンおよび取引量との関連性を分析し，ニュースの発生が情報の非対称性を解消する効果を持ち，非対称性が大きいと考えられる小型の低流動性銘柄では，ニュースが取引量やニュース発生後の株価変動パターンに影響を与えていることを指摘した．ニュースの内容を考慮した先駆的な研究として，Tetlock (2007) は，新聞のコラムから，語句とセンチメントを紐づけた辞書と主成分分析によって市場センチメントを定量化する手法を提案した．その手法によって算出した悲観度合いに関するスコアの騰落がノイズトレーダーの投資行動を通じて，短期的に取引量と資産価格に影響を与えることを指摘した．これ以後，テキストの内容に注目した研究が活発化し，スコアリング手法が数多く提案されている．これらについては、Kearney and Liu (2014) によるサーベイを参照されたい．

内容に即したニュースの定量化には，ニュースや語句の分類が肝となる．Tetlock (2007) のように辞書をベースにした手法であるルールアプローチは使用する辞書に依存するという弱点があり，本来であれば，用途に応じた辞書を構築することが望ましい．Baker et al. (2016) は，経済不確実性指数の構築にあたって，人間の手で経済政策に関連する語句群を新たに定義している．この方法は正攻法ではあるものの，分析コストが高くなってしまう点は否めない．Li (2010) は，一部のデータにのみポジティブ，ネガティブ等の極性とカテゴリーを人力で付与し，それを訓練データとして学習したナイーブベイズ分類器によって，残りのテキストデー

本論文の内容や意見は著者個人に属し、所属機関の公式見解を示すものではありません．

タをラベリングしている．このように統計モデルや機械学習モデルによって分類する手法は統計的アプローチとよばれる．効率的に独自の基準でラベリングすることができるので，もし適当な訓練データがあれば，すべて機械的に処理することが可能である．

以上の手法では，人間が明示的に付与したラベルに基づいたニュース分類がもちいられるので，ニュースの持つ情報がポジティブとネガティブの 2 つの分類に縮約されてしまう傾向にある．これは，ラベルが一定の客観性や一貫性が担保される基準で分類できるものに限定され，そこにはめ込まれる過程で，文書や語句の持つ潜在的な意味を捨象してしまうためであると我々は推察する．実際に先行研究で算出されている多くのスコアが，その「上昇/低下」と，「ポジティブ/ネガティブ」の他，「リターン上昇/下落」や「リスクオン/オフ」とが対応した作りになっている点はその証左であろう．しかし，実際の金融市場はそこまで単純ではなく，一口に負のインパクトをもたらすネガティブニュースといっても，トピックの違いによって下落幅や変動パターン別に様々な反応を示す．以上の論点を踏まえて，我々は，分析対象や用途に最適化した分類軸で多元的に評価できる新たなデータ分析手法が必要であると考える．

文書を複合的で潜在的なトピックによって分類する手段としては，トピックモデルが有名である．Blei et al. (2003) によって，潜在ディリクレ配分モデル (Latent Dirichlet Allocation; LDA) として提案されたトピックモデルは，文書における語句の出現パターンによって，各文書が持つ潜在的なトピックを確率的に割り振る確率モデルであり，様々な拡張が行われている．Blei and Mcauliffe (2007) は，各文書に教師データを付与する，教師ありトピックモデル (Supervised LDA; sLDA) に拡張した．文書に割り当てられたトピックを入力とした予測関数の出力として教師データを与えることで，教師データの変化をよく説明するトピックを学習する効果がある．[1]数少ないファイナンス領域への応用例として，Yono et al. (2020) は，マクロ経済ニュースを対象に，sLDA を用いて，VIX を教師データとし，不確実性に関連するトピックを抽出し，文書数とトピックの構成比率から不確実性指数の算出を試みている．sLDA は，人間があらかじめ指定したラベル等を用いずに，分析の目的に応じた教師データを軸として，多元的に文書を分類するという意味で，上述の問題点を克服した手法であるといえる．

しかし，現実の金融経済ニュースの分析に用いるには，いくつかの点において不備がある．まず，sLDA は文書と教師データが一対一で対応付けされていなくてはならない．金融経済ニュースは，特定の金融データとの紐づけがなされていないか，逆に，アセットクラスや個別銘柄などが複数紐づけされていることが多い．したがって，複数のデータについて横断的に分析しようとするとき，一つの文書に対して，複数の教師データを対応させる必要がある．また，金融経済ニュースには，カテゴリーやテーマ等に関するラベルが付与されていることが多い．こうしたラベルのみを用いて，ニュースを分類することは，必ずしも分析の目的に沿ったものとはならないため，適切とは言えないが，トピックモデルにおけるトピックの学習の一助になる可能性はあり，一切参照しないというのは得策ではない．

本研究では，こうした問題意識から，トピックモデルをファイナンス分野への応用を念頭に拡張する．本研究で提案するモデルでは，記事に付与されたラベルを文書分類のための補助情報として取り込みつつ，文書に割り当てられたトピックからその文書に対応する複数の教師データを予測する構造を持つ．特に予測部分について，パラメータが銘柄間でランダムな値をとりうるマルチレベルモデルを導入する点が，本研究の最も重要な貢献である．

また，実際のニュースデータを用いた実証分析によって，提案モデルの有用性を確認する．さらに，提案モデルをベースに算出したスコアをもとに株価とニュースの関係について分析し，将来にわたって株価への影響

[1] ここで，分類を教師データとして与える統計アプローチのような手法とは異なることに注意されたい．

が持続する企業のファンダメンタルズに関する情報が，ニュースに含まれる可能性があることを示す．

本論文の構成は次の通りである．第 2 節でトピックモデルに関する先行研究のサーベイを行った後，第 3 節で提案モデルについて説明する．第 4 節にて検証に使用するデータセットおよびその前処理の方法について述べる．第 5 節ではデータ分析の結果を示し，その解釈を行う．最後に第 6 節において本論文の結論を述べる．

2 先行研究

Blei et al. (2003) によって，LDA として最初に考案されたトピックモデルは，文書および語句が，観測されない潜在的なトピックに基いて生成されるように定式化した確率モデルである．語句の頻度を要素として文書をベクトル変換する Bag of Words (BoW) 型の文書について，文書内の語句の頻度や共起パターンから，各文書についてトピックの構成比率 (文書－トピック分布) と，各トピックにおける語彙の出現確率 (トピック－語句分布) によって表現する．LDA では，文書において，文書－トピック分布 (多項分布) からトピックを発生させ，そのトピックをもとに，トピック－語句分布 (多項分布) から，語句を発生させる．これを語句の数だけ，文書の数だけ繰り返すことで，文書集合すなわちコーパスが生成される．ここで，トピックの数については分析者が指定する必要がある．また，文書－トピック分布およびトピック－語句分布について，それぞれディリクレ分布が事前分布として仮定される．観測される文書および語句の生成に，階層構造を持つ潜在変数を導入した，いわゆる階層ベイズモデルである．

LDA は，学習にあたってトピックの数以外の情報を必要としない，教師無し学習モデルである．しかし，実際の文書には，ニュース記事等のようにその内容を理解したり，検索したりするためのラベルやキーワードが付与され，評価レビュー等のようにテキストの内容に応じたスコアが振られていることが少なくない．トピックの学習において，こうした付加的な情報を考慮できる拡張トピックモデルが多く考案された．付加情報の導入の仕方は大きく 2 つに分類できる．

第一に，付加情報にトピックや語句の発生確率を依存させるタイプである．同タイプのモデルは数多く提案されている．文書－トピック分布に付加情報を反映させた例として，著者トピックモデル (Author Topic Model)(Rosen-Zvi et al., 2004) は文書の著者を付加情報として，著者ごとに異なる文書－トピック分布を学習する．ラベル付きトピックモデル (Labeled LDA)(Ramage et al., 2009) は，付加情報である複数のラベルからデザイン行列を作成し，文書に割り当てるトピックを直接制約する．ディリクレ多項回帰 (Dirichlet Multinomial Regression)(Mimno and McCallum, 2008) は文書－トピック分布のパラメータを付加情報によって予測させることで，トピックの出現パターンに付加情報を反映させる．トピック－語句分布に付加情報を反映させた例として，Eisenstein et al. (2011) は語句の発生確率の共変量として付加情報を導入する (Sparse Additive Generative Model; SAGE)．Roberts et al. (2016) の構造トピックモデル (Structural Topic Model) は，文書－トピック分布とトピック－語句分布の両方に付加情報を考慮できる．トピック－語句分布には SAGE を採用し，文書－トピック分布に多変量の正規分布を仮定する (Blei and Lafferty, 2007) ことで相関構造を取り入れつつ，期待値を付加情報の線形結合で表現する．

第二に，トピックを入力とした予測関数の出力として付加情報を付与するタイプである．sLDA (Blei and Mcauliffe, 2007) は，付加情報がトピックの割合の線形結合として生成される．したがって，付加情報の変化にとって重要なトピックを学習することができる．また，トピックの学習時に推定した回帰係数のパラメータの大小をもとに，用途に応じて注目するべきトピックを選別することができる．Zhu et al. (2012) はより強力な予測器として，ソフトマージン法による教師ありトピックモデル (Maximum Margin Supervised Topic Model) を提案している．Perotte et al. (2011) は階層構造をもつ付加情報を教師データとして取り入れられ

るように拡張した (Hierarchically sLDA).

以上のように，トピックの学習に付加情報を考慮するための様々な手法が提案されているが，金融市場分析を念頭に拡張されたモデルは我々が調べた限りでは見当たらない．金融経済ニュースには，トピックの内容を表すラベルが付与されていることが多いため，トピックの発生確率を依存させる手法の応用が考えられる．さらに，トピックを入力とした予測関数の出力として付加情報を付与し，トピックによる予測モデルを構築することで，金融市場の予測分析を実現できる可能性がある．ただし，しばしば金融経済ニュースには，複数の企業やアセットクラスが関連付けされているので，1 つのニュースに対して複数の銘柄を対応させる工夫が必要になる．ナイーブな対応策として，複数銘柄のデータ集約か，あるいは銘柄別のモデル構築が考えられる．前者については，必ずしも全ての銘柄に対して，同じトピックが同じ影響を与えるとは限らないため問題がある．後者については，モデルごとに異なるトピックが学習されるため，銘柄横断的な比較分析ができないという弱点がある．以上の論点を踏まえ，Roberts et al. (2016) の構造トピックモデルと Blei and Mcauliffe (2007) の sLDA のハイブリッドモデルを土台に，予測部分に銘柄固有のランダム効果を許容することによって，複数銘柄の同時分析を実現するモデルであるマルチレベル教師ありトピックモデル (Multi Labeled and Supervised Topic Model; MLSTM) を提案する．このモデルの詳細は次節で説明する．

3 提案モデル MLSTM

観測できる変数について，文書数を M とし，インデックスを $d \in \{1, 2, ..., M\}$ とする．文書 d の語句数を n_d とし，i 番目の語句を $w_{d,i} (i \in \{1, 2, ..., n_d\})$ とする．語彙数を V とし，インデックスを $v \in \{1, 2, ..., V\}$ とする．また，$w_{d,i} \in \{1, 2, ..., V\}$ である．教師データのグループ数を J とし，インデックスを $j \in \{1, 2, ..., J\}$ とする．文書 d に紐づけられている教師データのグループとデータの集合をそれぞれ J_d，$\{y_{d,j} \mid j \in J_d\}$ とする．ここで，必ずしも $J_d = \{1, 2, ..., J\}$ でなくても良い．文書 d のラベルに基づくベクトルを $\boldsymbol{x}_d$ とする．

語句の生成過程において，潜在変数として K 種類のトピック $(k \in \{1, 2, ..., K\})$ を導入する．$w_{d,i}$ に対応する潜在トピックを $z_{d,i} \in \{1, 2, ..., K\}$ とする．文書 d におけるトピック k の発生確率を $\theta_{d,k}$ とし，ベクトル $\boldsymbol{\theta}_d = (\theta_{d,1}, \theta_{d,2}, ..., \theta_{d,K})^\top$ を文書 − トピック分布のパラメータとする．トピックが k となった語句 $(\{w_{d,i} \mid z_{d,i} = k\})$ における語彙 v の発生確率を $\phi_{k,v}$ とし，確率ベクトル $\boldsymbol{\phi}_k = (\phi_{k,1}, \phi_{k,2}, ..., \phi_{k,V})^\top$ をトピック − 語句分布とする．$\boldsymbol{\theta}_d$ と $\boldsymbol{\phi}_k$ を，それぞれトピックと語句に関する多項分布のパラメータとして，文書 d における語句 $w_{d,i} = v$ の発生確率を $\sum_{k=1}^{K} p(z_{d,i} = k \mid \boldsymbol{\theta}_d) p(w_{d,i} = v \mid \boldsymbol{\phi}_k)$ によって表現する．これが潜在変数 $z_{d,i}$ を介して語句が生成されるトピックモデルのコアである．

トピック − 語句分布 $\boldsymbol{\phi}_k$ には，ベーシックモデルと同様に，パラメータ $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_V)^\top$ のディリクレ分布を事前分布として仮定する．これは，多項分布とディリクレ分布の共役性による．

文書 − トピック分布のパラメータ $\boldsymbol{\theta}_d$ の事前分布として，トピックの発生パターンに相関構造を導入するため，多変量の正規分布を仮定する．ディリクレ分布と異なり正規分布に従う確率変数ベクトルは合計が 1 であることを保証しないため，$\boldsymbol{\theta}_d$ をソフトマックス関数，

$$\mathrm{softmax}(\boldsymbol{\theta}_d) = \frac{1}{\sum_{k=1}^{K} \exp(\theta_{d,k})} (\exp(\theta_{d,1}), \exp(\theta_{d,2}), ..., \exp(\theta_{d,K}))^\top,$$

によって単体に射影し，文書 − トピック分布とする．正規分布の分散を Σ とし，平均を $\boldsymbol{x}_d$ の回帰係数 Γ による線形結合とする．ここで，他の観測変数とは異なり，$\boldsymbol{x}_d$ は確率変数ではなく，外生的な変数である．

教師データ $y_{d,j}$ の生成確率として正規分布を仮定する．分散を ν_j^2 とする．平均は $\bar{\boldsymbol{z}}_d$ の回帰係数 $\boldsymbol{\eta}_j$ による

線形結合とする．ここで，$\bar{\boldsymbol{z}}_d$ は $\vec{\boldsymbol{z}}_{d,i} = (\vec{z}_{d,i,1}, \vec{z}_{d,i,2}, ..., \vec{z}_{d,i,K})^\top$ として，

$$\bar{\boldsymbol{z}}_d = \frac{1}{n_d} \sum_{i=1}^{n_d} \vec{\boldsymbol{z}}_{d,i},$$

$$\vec{z}_{d,i,k} = \begin{cases} 1, & z_{d,i} = k \\ 0, & z_{d,i} \neq k, \end{cases}$$

である．添字からわかる通り，回帰係数 $\boldsymbol{\eta}_j$ はグループによって異なる値をとりうる．$\boldsymbol{\eta}_j$ の事前分布として，平均 $\boldsymbol{\mu}$，分散 Λ の多変量の正規分布を仮定する．さらに，Λ について，自由度 υ，スケール Ω の逆ウィシャート分布を事前分布として仮定する．ここで，逆ウィシャートモデル (Gelman and Hill, 2006) にならい，υ と Ω をそれぞれ $K+1$，$K \times K$ の単位行列 $\boldsymbol{I}_K$ とする．

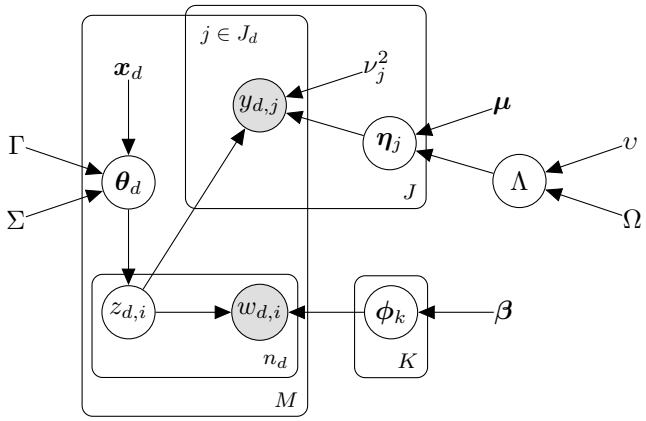
以上より, 外生変数 $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_M\}$ およびハイパーパラメータ υ, Ω, $\boldsymbol{\mu}$, $\boldsymbol{\beta}$, Γ, Σ, $\boldsymbol{\nu}^2 = \{\nu_1^2, \nu_2^2, ..., \nu_J^2\}$ を所与としたとき，$w_{d,i}$ と $y_{d,j}$ の生成過程は以下の通りである．

1. Draw $\Lambda \sim \mathcal{IW}(\upsilon, \Omega)$
2. For each group $j \in \{1, 2, ..., J\}$,
   (a) Draw $\boldsymbol{\eta}_j \sim \mathcal{N}(\boldsymbol{\mu}, \Lambda)$
3. For each topic $k \in \{1, 2, ..., K\}$,
   (a) Draw $\boldsymbol{\phi}_k \sim \mathcal{D}(\boldsymbol{\beta})$
4. For each document $d \in \{1, 2, ..., M\}$,
   (a) Draw $\boldsymbol{\theta}_d \sim \mathcal{N}(\Gamma \boldsymbol{x}_d^\top, \Sigma)$
   (b) For each word $i \in \{1, 2, ..., n_d\}$,
      i. Draw $z_{d,i} \sim \mathcal{M}(softmax(\boldsymbol{\theta}_d))$
      ii. Draw $w_{d,i} \sim \mathcal{M}(\boldsymbol{\phi}_{z_{d,i}})$
   (c) For each group $j \in J_d$,
      i. Draw $y_{d,j} \sim \mathcal{N}(\boldsymbol{\eta}_j^\top \bar{\boldsymbol{z}}_d, \nu_j^2)$

ここで，$\mathcal{IW}(\cdot)$，$\mathcal{N}(\cdot)$，$\mathcal{D}(\cdot)$，$\mathcal{M}(\cdot)$ は，それぞれ逆ウィシャート分布，正規分布，ディリクレ分布，多項分布による生成分布である．

グラフィカルモデルを図 1 に示す．観測される確率変数を色付きの円，潜在変数を色無しの円で囲んでいる．

図 1　グラフィカルモデル



$\boldsymbol{w} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_M\} = \{\{w_{1,1}, ..., w_{1,n_1}\}, \{w_{2,1}, ..., w_{2,n_2}\}, ..., \{w_{M,1}, ..., w_{M,n_M}\}\}, \boldsymbol{z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_M\} = \{\{z_{1,1}, ..., z_{1,n_1}\}, \{z_{2,1}, ..., z_{2,n_2}\}, ..., \{z_{M,1}, ..., z_{M,n_M}\}\}$，$\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_M\}$，$\boldsymbol{\phi} = \{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, ..., \boldsymbol{\phi}_K\}$，$\boldsymbol{y} = \{\{y_{1,j} \mid j \in J_1\}, \{y_{2,j} \mid j \in J_2\}, ..., \{y_{M,j} \mid j \in J_M\}\}$，$\boldsymbol{\eta} = \{\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, ..., \boldsymbol{\eta}_J\}$ として，全ての確率変数の結合分布をベイズの定理により展開すると，

$$
\begin{aligned}
&p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{y}, \boldsymbol{\eta}, \Lambda | \Gamma, \boldsymbol{x}, \Sigma, \boldsymbol{\beta}, \boldsymbol{\nu}^2, \boldsymbol{\mu}, \upsilon, \Omega) \\
&\quad = p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\phi}) p(\boldsymbol{z}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{x}, \Gamma, \Sigma) p(\boldsymbol{\phi}|\boldsymbol{\beta}) p(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\eta}, \boldsymbol{\nu}^2) p(\boldsymbol{\eta}|\boldsymbol{\mu}, \Lambda) p(\Lambda|\upsilon, \Omega) \\
&\quad = \left(\prod_{d=1}^{M} \prod_{i=1}^{n_d} p(w_{d,i}|\boldsymbol{\phi}_{z_{d,i}}) p(z_{d,i}|\boldsymbol{\theta}_d) \right) \left(\prod_{d=1}^{M} p(\boldsymbol{\theta}_d|\boldsymbol{x}_d, \Gamma, \Sigma) \right) \left(\prod_{d=1}^{K} p(\boldsymbol{\phi}_k|\boldsymbol{\beta}) \right) \\
&\qquad \times \left(\prod_{d=1}^{M} \prod_{j \in J_d} p(y_{d,j}|\boldsymbol{z}_d, \boldsymbol{\eta}_j, \nu_j^2) \right) \left(\prod_{j=1}^{J} p(\boldsymbol{\eta}_j|\boldsymbol{\mu}, \Lambda) \right) p(\Lambda|\upsilon, \Omega),
\end{aligned}
$$

が得られる．提案モデルのオリジナリティは，構造トピックモデルによる拡張 $p(\boldsymbol{\theta}|\boldsymbol{x}, \Gamma, \Sigma)$ と，教師ありトピックモデルによる拡張 $p(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\eta}, \boldsymbol{\nu}^2)$ を組み合わせ，教師データの生成に関する部分をマルチラベルモデル $p(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\eta}, \boldsymbol{\nu}^2) p(\boldsymbol{\eta}|\boldsymbol{\mu}, \Lambda) p(\Lambda|\upsilon, \Omega)$ に拡張した点にある．

なお，本研究では提案モデルをフィッティングするための学習アルゴリズムの実装に変分ベイズ法を用いている．実装の詳細については本論文の付録を参照されたい．

4　データセット

4.1　ニュースデータ

本研究で用いるニュースデータは，日本経済新聞社の「日本経済新聞　電子版」の「日経会社情報 DIGITAL」[2] から，スクレイピングによって取得した．「日経会社情報 DIGITAL」では企業ごとのページが提供されており，各企業がタグ付けされたニュースの一覧にアクセスすることができる．スクレイピングのステップは以下の通りである．

1. ホームページの業界一覧より，NEEDS 業種 (中) 分類コードを取得.

[2] 「日経会社情報 DIGITAL」(`https://www.nikkei.com/nkd/`).

2. 業種ページより，業種に属する企業コードを取得.
3. 企業ページより，企業がタグ付けされたニュース記事の ID を取得.
4. ニュース記事ページより，本文，見出し，キーワード，公開/更新日時を取得.

取得した記事の中には，明らかに金融市場に無関係なニュースも含まれているため，ルールベースで検知できるものについてはあらかじめ除外する．具体的には，人事異動，首相官邸，予定，クイズ，音楽を除外した．例えば，人事異動ニュースであれば，見出しについて「 ^(人事)」や「^(新社長」といった正規表現，首相官邸ニュースについてはキーワード「首相官邸」で検知することができる．また，リンクや動画コンテンツのみで構成されているニュースも除外した.

日本語の文書を BoW 表現に変換するためには，形態素解析によって文章を語句に分割しなくてはならない．本研究ではオープンソースの形態素解析ツール MeCab[3](Kudo et al., 2004) を用いる．形態素解析のための辞書として，新語にも対応した mecab-ipadic-NEologd (NEologd)[4](Sato, 2015)(佐藤他, 2016)(佐藤他, 2017) をベースとする．さらにニュースソースに即した辞書に改良するための工夫として，スクレイピングによって取得したキーワードおよび企業名をユーザー辞書として追加する．これにより，NEologd では分割してしまうが，日本経済新聞の記事のキーワードとして扱われている「日経平均先物」,「脱ガソリン」といった語句 (約 15,000 語) や，NEologd が判別できない企業名 (約 500 社) を分割せずに解析することができる．URL リンクを除外し，記号，アルファベット，数字を半角に統一した後，各ニュース記事について，見出しと本文に形態素解析を施す．解析結果から，「名詞」,「動詞」,「形容詞」,「副詞」をピックアップし，名詞の「数」,「接尾」,「非自立」を除外する．ここに，各ニュース記事のキーワードをそれぞれ加える．さらに,「5%」や「10%」といった語句を同一に扱うため，数値は「#」に変換した.[5]改めて数値 (「#」) を除外し，最後に文書全体における頻度が上位 1% となる頻出語句を除外し，各ニュースの語句集合とする.

分析対象は，タグ付けニュース件数上位 20 社 ($J = 20$) とし，各銘柄をグループとする．20 社のニュースを抽出した後，頻度が 20 回未満の語句を除外し，語句が 5 個未満の文書を除いた．サンプル期間は 2019 年 12 月 27 日から 2021 年 11 月 21 日である．表 1 に分析対象銘柄とそれぞれのニュース件数をまとめた.

[3] MeCab 0.996.
[4] 2021/8/26 時点の辞書を使用.
[5] 「S&P500」や「5G」等の数字を含む固有名詞も，「S&P#」や「#G」と変換されてしまうが，モデル学習上はあくまで記号的に処理を行うため，特に問題はなく，重要な語句については考察の際に数値を補えばよい.

表 1　ニュース件数上位 20 銘柄

(注) 各銘柄のニュース件数は延べ件数.

銘柄コード	企業名	業種	ニュース件数
4755	楽天グループ	インターネットサイト運営	1,111
5401	日本製鉄	製鉄・金属製品	1,127
6501	日立製作所	総合電機	1,495
6502	東芝	総合電機	1,323
6701	ＮＥＣ	総合電機	1,140
6702	富士通	総合電機	1,003
6752	パナソニック	総合電機	1,587
6758	ソニーグループ	総合電機	2,139
7201	日産自動車	自動車	2,116
7203	トヨタ自動車	自動車	5,114
7267	ホンダ	自動車	2,206
8035	東京エレクトロン	製造用機械・電気機械	1,239
9020	ＪＲ東日本	陸運	1,650
9021	ＪＲ西日本	陸運	1,052
9201	日本航空	空運	1,442
9432	ＮＴＴ	通信サービス	1,265
9433	ＫＤＤＩ	通信サービス	1,448
9503	関西電力	電力・ガス	960
9983	ファーストリテイリング	衣料品・服飾品	1,370
9984	ソフトバンクグループ	通信サービス	3,441
		(合計)	24,964

4.2　ラベルデータ

各文書のラベルデータ $\boldsymbol{x}_d$ には，ニュースにタグ付けされている企業の業種データをダミー変数に変換し，定数項と合わせたベクトルを使用する．これにより，業種内で共通するトピックを学習する効果が見込まれる．複数の銘柄がタグ付けされているニュースもあり，複数の業界を組み合わせたようなトピックにも対応できる．

極端な例として，市況や複数社へのインタビューをまとめた記事には，雑多に銘柄がタグ付けされていることが多い．その際，無暗に業種ダミー変数が文書 – トピック分布の期待値に影響してしまう．しかし，提案モデルはラベルによる予測誤差を計測し，学習へのラベルの影響度合いを調整する仕組みを持つ．そのため，このようなニュースには相対的にラベルに依存しないトピックの比率が大きくなることが期待される．すなわち，無関係なラベルが付与されたニュースを，ラベルとは関係のないニュースとして学習することができる．

4.3 教師データ

教師データ $\boldsymbol{y}$ には，ニュースに対する投資家の反応のプロキシーとして取引金額を使用する．[6]ニュースソースの非速報性や市場の反応ラグを考慮して，当日と前後 2 日間を含む 3 日間の平均値をとる．取引金額は市況および銘柄間で水準が異なるため，次の方法で基準化を行う．まず，市場全体の影響を取り除くため，東証 1 部の取引金額で除す．次に，過去 20 日平均を差し引き，取引金額の変化を求める．最後に各銘柄について，サンプル期間で平均 0，分散 1 となるように標準化する．

更新日時が平日の大引け (午後 3 時) までのニュースについては当日，大引け以降のニュースおよび休日のニュースについては翌営業日の教師データに対応させる．なお，同タイミングのニュースは同じ教師データを共有している．

5 データ分析

5.1 モデルのセットアップ

今回の分析ではトピック数 K を 50 に設定する．$\boldsymbol{\beta}$ は全要素について 0.1，$\boldsymbol{\mu}$ は全要素について 0 に固定する．初期値として，$\hat{\boldsymbol{\theta}}_d$ は $[0,1]$ の一様乱数に対数をとり，K 番目の要素を 0 に基準化した値，$\boldsymbol{\phi}_k$ は各要素に対応する語彙の頻度を全語句数で除した値，Σ は $\hat{\boldsymbol{\theta}}_d$ の初期値の分散共分散行列，$\hat{\boldsymbol{\eta}}_j$ は $\boldsymbol{\mu}$，$\hat{\Lambda}$ は単位行列 $\boldsymbol{I}_K$，ν_j^2 は教師データの分散を設定する．なお，本提案モデルのフィッティングには付録でも説明しているように変分ベイズ法を用いており，その反復数は 100 回としている．[7]

5.2 トピックの評価

トピック－語句分布 $\boldsymbol{\phi}$ と文書－トピック分布 $\boldsymbol{\theta}$ を確認することで，学習したトピックの評価を行う．図 2，3 に，トピック－語句分布をもとにワードクラウドを描画した．ワードクラウドは各トピックの確率上位の語句によって構成されており，文字のサイズは確率の大きさを表す．また，表 2，3 に，文書－トピック分布において，各トピックの確率が最大となるニュースの見出しをまとめた．これらの図表より，トピックについて解釈を行うことができる．例えば，トピック 1 は「業績」,「見通し」等の語句が並び，企業業績に関するトピックであることがわかる．トピック 2 は「株主総会」,「取締役会」,「企業統治」等の語句が並び，東芝の株主総会に係る不正に関するニュース記事が当トピックの代表的な文書として示されていることから，株主総会やガバナンスに関するトピックであると考えられる．同様に，トピック 4 は経営トップ人事，トピック 5 は燃料電池自動車，トピック 6 は自動運転技術や電気自動車など次世代の自動車開発，トピック 7 は生産工場の停止・再開，トピック 9 は市況，トピック 10 は M&A やコーポレートアクション，トピック 11 はアンケート，トピック 12 は小売業界，トピック 13 はパンデミック禍の医療体制，トピック 15 はパンデミックによる業績悪化，トピック 16 は金融業界の再編，トピック 17 は企業のパンデミック対応，トピック 18 は自動車販売，トピック 19 は日産自動車元会長による不正問題，トピック 21 はテクノロジー，トピック 22 は宇宙開発，トピック 23 は半導体，トピック 24 は携帯キャリア，トピック 25 はスタートアップ企業，トピック 27 は旅客および空運・陸運業界，トピック 29 は労使交渉，トピック 32 は自治体，都市開発，トピック 33 はオリンピッ

[6] 株式市場に関するデータは全て Datastream から取得した．

[7] 反復終了時点において，各パラメータが十分収束していることを確認している．

ク，演劇，トピック 34 は総合電機業界，トピック 35 は電力業界，トピック 36 は中国における人権問題，トピック 37 は陸運業界，トピック 38 はキャッシュレス決済，トピック 39 は中国におけるビジネス，トピック 43 は ESG，SDGs，トピック 44 はエンターテインメント，トピック 45 は欧米の自動車業界，トピック 49 は環境対策，トピック 50 はロボット技術に関するトピックと解釈できる．中には，人間が解釈困難なトピック (3，8，20，30，31，40，41，46，47，48) や複数のテーマが混合したようなトピック (14: バッテリーとシステム障害，26: 小売と医療，28: 景気・需要と鉄鋼・製鉄，42: トイレとタイヤ) もみられる．

提案モデルは，Σ によってトピック間の相関も許容しているため，相関関係について分析をすることもできる．[8]なお，ここでいう相関係数はトピックの潜在的な意味に関する類似度ではなく，発生パターンに関する類似度を表している．相関係数が高いトピックのペアとしては，トピック 15-28 (0.92)，トピック 1-28 (0.88)，トピック 1-15 (0.88)，トピック 2-19 (0.83)，トピック 1-10 (0.82) がある．例えば，トピック 1，15，28 は業績や景気に関するトピック，トピック 2，19 はガバナンスや経営陣の不正に関するトピックであり，それぞれ共起しやすいトピックとして現実のニュースと合っている．

[8] 識別制約のため，$K-1$ 個についてのみ分析することができる．

図２　各トピックの発生確率上位の語句によるワードクラウド (1)

図 3　各トピックの発生確率上位の語句によるワードクラウド (2)

表 2　各トピックの構成比率が最も高いニュースの見出し (1)

トピック	構成比率	見出し
1	80%	海外投資家の出資を事前審査する重点企業
2	93%	「東芝総会 公正でなかった」 弁護士調査報告
3	25%	オンキヨー再建、車載に活路　祖業売却益どう生かす
4	61%	日立、小島氏が社長兼 COO に昇格　東原氏は会長兼 CEO に
5	38%	SMS 詐欺、筆者が追跡　偽サイトに誘う URL 巧妙に
6	63%	いすゞ、米社からエンジン調達　コスト削減で電動車集中
7	86%	ホンダ、鈴鹿製作所の稼働停止 2 日延長
8	33%	代表幹事　大先輩に思い
9	100%	東証後場寄り　 500 円高、ソフトバンク G など高い
10	84%	ソフトバンク G、通信子会社株の一部売却で 3300 億円調達
11	82%	経営者「規制緩和を」9 割　「景気拡大」38% に増加
12	55%	ユニクロ、ネット注文後 2 時間で店舗に商品
13	29%	新潟の酒蔵改革は異業種から　「稲盛経営」で再建
14	50%	富士通「ご迷惑かけおわび」　東証のシステム障害で
15	60%	JR 西日本、連続最終赤字
16	32%	SBI・新生銀行、深まる対立　まとめ読み
17	78%	新型コロナ: NTT、ワクチン接種 7 月にも開始　 21 年内 4 万人分を計画
18	92%	新車販売、1 月ヤリス首位
19	89%	米に 2 人引き渡し請求
20	47%	日経広告賞の贈賞式、ヤフーが大賞受賞
21	82%	日立、ブロックチェーンを使ったシステム開発支援
22	42%	〈Next ストーリー 宇宙大航海に挑む〉(5) ロケットの「スーパーカブ」
23	68%	半導体製造装置 進む世代交代
24	94%	KDDI、主力「au」で値下げ方針　ドコモに対抗
25	58%	ソフトバンク、マレーシアの広告会社と提携　 65 億円出資

表 3　各トピックの構成比率が最も高いニュースの見出し (2)

トピック	構成比率	見出し
26	40%	がんや認知症、血液 1 滴で早期発見　分析技術が進展
27	96%	新型コロナ: AIRDO ANA　 JAL、2 月は北海道内発着 5 割減便
28	54%	全国商業地 55% で下落、大阪で市況悪化　21 年基準地価
29	81%	トヨタ労組、賃上げ 1 万 100 円要求へ
30	32%	トヨタ販売改革の成算 全車種併売の衝撃（中） いずれは「家電量販店型」も
31	51%	新潟の海と砂が育むワイン　栽培醸造家・本多孝氏
32	69%	トヨタと森ビル、お台場で再開発　アリーナや商業施設
33	29%	新型コロナ: 大阪 4 楽団、生き残りに地域密着　格安公演などファン開拓
34	47%	保有特許に注目　「輸送用機器」関連の技術成長株
35	93%	関電　原発再稼働へ一歩
36	35%	香りの器　高砂コレクション展
37	86%	東北新幹線が通常復旧 所要時間や本数
38	79%	楽天ペイ、新規中小の手数料 1 年無料
39	45%	民主派香港紙、深まる苦境 創業者服役・資産 70 億円凍結
40	36%	新潟―庄内の料亭・イタリアンの味満喫　観光列車「海里」
41	79%	大崎電気工業社長 渡辺光康さん
42	25%	新型コロナ: コロナで脚光 日本発トイレ革新、世界へ　TOTO・LIXIL
43	61%	持続可能な社会へ解決力
44	73%	あつ森、マリオに迫る　任天堂のスイッチソフト累計販売
45	47%	ダイムラー、全ルノー株売却へ
46	60%	大手町や霞が関…　赤門は東大だけじゃなかった
47	57%	秋の叙勲 4100 人　旭日大綬章に仲井真元沖縄知事ら
48	39%	スエズ座礁事故 物流目詰まり懸念
49	76%	東芝系など 6 社、ジェット燃料に CO2 再利用
50	81%	ホンダ、常識覆す車体設計　鋼板強度下げても試験満点

5.3　文書 – トピック分布の期待値と業種ダミー変数の関係

ラベルの文書 – トピック分布の期待値への回帰係数 Γ を確認することで，ラベルと各トピックの関係を調べることができる．分析対象の銘柄が属する「電力・ガス」,「製鉄・金属製品」,「製造用機械・電気機械」,「総合電機」,「自動車」,「衣料品・服飾品」,「陸運」,「空運」,「通信サービス」,「インターネットサイト運営」に絞り，推定結果を観察する (表 4)．ここで，回帰係数がゼロであるという帰無仮説について仮説検定 (t 検定) を行っており，有意水準 95% で帰無仮説を棄却しなかった場合に「-」とした．また，各業種ラベルにおいて推定値が大きかった上位 3 つを太字にしている．ラベルを参考に学習しているため，ほとんどの回帰係数について有意な結果が得られている．電力・ガスではトピック 35，製鉄・金属製品ではトピック 28，製造用機械・電気機械ではトピック 23，総合電機ではトピック 21 や 44，自動車ではトピック 18，衣料品・服飾品は

トピック 12，陸運はトピック 37，空運はトピック 27，通信サービスはトピック 24，インターネットサイト運営はトピック 38 の発生に相対的に大きく寄与している．それぞれの業種に対応した業界や製品・サービスに関連するトピックであることから，いずれもすでに述べたトピックの解釈と整合的である．

図 4 は Σ の対角要素であり，ラベルによる予測精度として解釈できる．値が高ければ，特定の業種とは無関係のトピックであり，値が低ければ，業種固有のトピックである可能性が高い．トピック 9 は著しく高い値を示しており，ラベルの影響をあまり受けていないトピックである．すでに述べた解釈を踏まえると，トピック 9 は市況に関するトピックであるため，業種ラベルをノイズとして，あまり重視しないように学習したと考えられる．反対に，低い値を示したトピック 6，21，44 などは特定の業界の製品や技術に関連した業種固有のトピックとして学習されていることがわかる．

表 4　トピックと業種ラベルの関係

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
電力・ガス	-0.29	1.27	0.86	0.65	0.57	-	0.32	0.29	-0.39	-0.2	0.62	0.15	-	0.66	0.5
製鉄・金属製品	0.85	0.13	-0.22	-	-0.62	-0.37	0.55	0.19	**1.22**	0.24	0.27	-0.67	-0.12	0.56	0.52
製造用機械・電気機械	**1.86**	0.4	0.4	-	-0.27	0.15	0.48	0.87	**3.64**	1.09	0.39	-0.24	0.67	1.1	0.66
総合電機	-0.34	0.29	-0.16	0.15	-0.76	0.05	-0.65	-0.06	-0.68	-0.28	-	-0.87	-0.3	0.18	-0.94
自動車	-	-0.54	-0.14	-	0.32	**1.6**	0.95	0.3	0.13	-0.14	-0.28	-	0.11	0.36	0.15
衣料品・服飾品	0.98	-	0.81	0.26	-0.16	-0.53	0.43	0.4	**2.59**	0.7	0.32	**2.35**	0.44	0.48	0.77
陸運	-0.35	-0.94	-0.37	-0.58	0.48	-0.86	-0.95	-0.09	-0.94	-0.78	0.7	0.08	0.12	-0.73	0.65
空運	0.51	-0.15	0.37	0.27	0.89	-0.27	0.72	0.34	0.73	0.4	0.4	0.45	0.67	0.3	**1.66**
通信サービス	0.58	0.89	0.61	0.38	0.18	0.28	-0.47	0.16	1.27	**2**	0.7	-	0.47	0.39	0.32
インターネットサイト運営	0.58	0.57	0.4	0.53	0.67	0.06	-	0.15	0.88	0.63	0.4	**1.07**	0.7	0.67	0.5

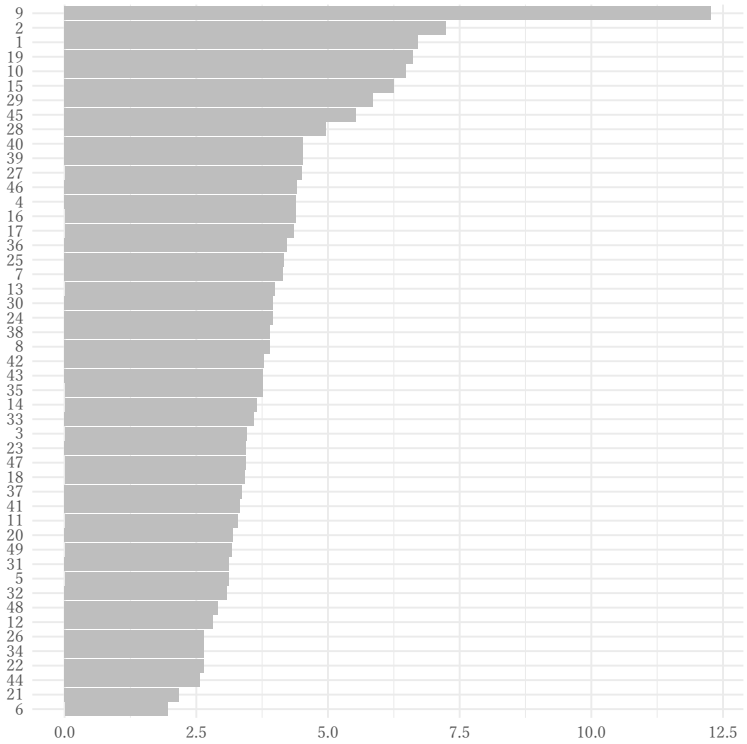
	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
電力・ガス	-	0.58	-0.31	0.31	1.12	0.11	0.21	-0.45	0.37	-	-	0.37	-	0.23	0.32
製鉄・金属製品	0.26	-0.63	-0.4	-0.26	-0.68	-0.93	0.29	-0.45	-1.19	-0.57	-0.61	**1.79**	**2.28**	0.27	0.3
製造用機械・電気機械	0.66	-0.2	0.12	-0.12	-0.09	-0.42	0.45	**2.29**	-	-0.22	0.14	-0.31	1.72	-	0.94
総合電機	-0.15	-0.2	-0.89	-0.82	-0.12	**0.86**	-0.3	0.19	-	-0.18	-0.31	-1.43	-0.51	0.33	-0.62
自動車	-0.51	-0.5	**2.77**	-0.08	-0.32	-0.63	-0.34	-0.58	-0.84	-0.57	-0.08	-0.88	0.17	-0.12	0.51
衣料品・服飾品	0.46	0.15	0.73	-	0.38	-0.57	0.31	-0.56	-0.72	-0.53	0.47	-0.12	0.71	0.15	0.47
陸運	-0.17	0.54	-	-0.6	0.38	-0.27	-0.37	-0.95	-0.55	-0.39	-	**1.54**	0.16	0.34	0.12
空運	0.39	1.28	-	-	0.31	-0.09	0.24	-0.26	0.11	0.53	0.22	**4.01**	0.89	1.28	0.75
通信サービス	1.14	0.56	0.05	1.27	0.44	0.99	0.75	0.85	**2.98**	1.26	0.17	-0.36	0.51	0.27	0.52
インターネットサイト運営	0.76	0.69	0.41	0.88	0.67	0.73	0.31	0.37	**1.53**	1.06	0.6	0.13	0.45	0.21	0.51

	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
電力・ガス	0.25	0.53	0.42	0.1	**4.39**	0.32	0.53	-0.27	-0.17	0.37	0.27	0.82	0.64	-0.26	0.53
製鉄・金属製品	-0.31	-0.75	-0.36	-0.18	0.34	0.55	-0.33	-0.78	0.29	-0.26	-0.36	0.62	-0.13	-0.66	-
製造用機械・電気機械	0.43	-0.47	-0.19	-0.21	0.3	1.07	0.17	-0.19	0.62	0.74	-	0.99	-	-0.1	0.56
総合電機	-0.41	-0.78	-0.21	0.31	-0.49	-0.33	-0.95	-0.68	-0.41	-	-0.23	-0.26	**0.35**	**0.89**	-0.62
自動車	-0.39	-0.61	-0.3	-0.49	-0.33	0.18	-0.42	-0.69	0.54	-0.45	-0.2	-	-0.16	-0.59	**1.3**
衣料品・服飾品	0.71	-0.39	-	-0.16	-0.37	**1.63**	-	0.13	1.49	0.38	0.61	0.12	0.36	-0.12	0.61
陸運	0.94	**1.54**	0.9	-0.49	-0.18	-0.61	**3.45**	-	-0.59	0.6	0.21	-0.28	-0.67	0.06	-0.81
空運	0.63	0.63	0.66	**1.36**	-0.14	0.39	1.16	0.29	0.45	0.84	0.66	0.54	0.22	0.19	0.14
通信サービス	0.61	0.73	0.97	0.67	0.13	0.6	0.25	**1.35**	0.71	0.64	0.69	0.43	0.84	0.69	0.28
インターネットサイト運営	0.39	0.41	0.41	0.52	-	0.39	0.24	**2.53**	0.38	0.44	0.7	0.19	0.76	0.45	0.39

	46	47	48	49	50
電力・ガス	-	**1.46**	-0.27	**1.86**	
製鉄・金属製品	-	0.09	-0.63	0.68	
製造用機械・電気機械	0.66	-	0.24	-0.48	
総合電機	-0.47	-	-0.56	-0.57	
自動車	-0.3	-0.27	-	-	
衣料品・服飾品	0.91	0.62	0.17	-	
陸運	0.49	0.4	-0.16	-0.13	
空運	0.51	0.74	0.5	0.16	
通信サービス	0.45	0.52	0.67	-0.15	
インターネットサイト運営	0.36	-	0.95	-0.19	

図 4　ラベルによる予測精度

(注) 縦軸はトピックの番号であり，予測誤差分散について降順で並べている．



5.4　トピックと取引金額の関係

各トピックに対する各銘柄の取引金額の反応を，回帰係数 $\boldsymbol{\eta}$ によって確認する．図 5，6 に各銘柄について点推定値が大きい順にトピックを 10 個ピックアップし，降順で示した．回帰係数の値が大きければ大きいほど，取引金額へのインパクトが大きいトピックであると解釈できる．銘柄によって，影響のあるトピックの序列が異なっており，また，同じトピックであっても影響度合いが異なっている．これは提案モデル独自の分析結果である．

トピック 1 は業績に関するトピックであったが，製造業を中心に上位に現れている．今回の分析対象銘柄が，ニュース件数の多い大型銘柄を中心に構成されているため，トピック 9 も同様に多くの銘柄で確認することができ，市況に関するトピックと取引金額の上昇が同時発生していることを反映していると考えられる．コーポレートアクションに関連するトピック 10 もいくつかの銘柄において際立っている．トピック 14 はバッテリーに関するトピックであるが，自動車セクターと総合電機セクターというようにサプライチェーンにわたって，上位トピックとなっている．自動車セクターでは，中国におけるビジネスに関するトピック 39 が共通して上位に現れている．電力・ガスセクターでは，同業界に関するトピック 35 と環境対策に関するトピック 49 が特徴的である．衣料品・服飾品セクターでは，小売業界に関するトピック 12 が上位となっている．

サンプル期間に発生した事象もトピックおよび回帰係数の学習に影響を与えている．パンデミック禍における人流抑制政策の影響を受けた陸運セクター，空運セクターのＪＲ東日本，ＪＲ西日本，日本航空では，パンデミックに関するトピック 15 やトピック 17 が上位に並ぶ．東芝と関西電力はトピック 2 の回帰係数が最大

となっており，ガバナンスに関する不祥事の発生を反映していると考えられる．

その他，個別銘柄について特徴的な結果について述べる．楽天グループでは，携帯キャリアに関するトピック 24 が上位となっており，同社がモバイルセグメントを持つことによるものと考えられる．ゲーム事業を軸とするソニーグループでは，ゲームやエンターテインメントに関連するトピック 44 がみられる．日産自動車では，同社の元会長による不正問題に関するトピック 19 が確認できる．また，他の自動車メーカーとは異なり，欧米の自動車業界に関するトピック 45 や環境に関連するトピック 49 が上位に入っており，同社の欧州自動車メーカーとのアライアンスと電気自動車生産への注力を反映したもの考えられる．ＮＴＴは携帯キャリアに関連するトピック 24 が上位に現れている一方で，ＫＤＤＩではキャッシュレス決済，電力に関する他セグメントのトピック (35，38) が目立っている．

以上のように，概ねトピックと銘柄の特色やイベントと整合的な結果が得られている．一部解釈が困難であった結果についても，さらに詳しく考察を行うことで解釈できる可能性がある．しかし，本研究ではこれ以上の解釈に立ち入らず，学習したトピックの株価予測性能についての分析に移る．

図 5　銘柄別の取引金額の変化に対する各トピックの回帰係数 (1)

(注) 縦軸はトピック番号，横軸は点推定値 (棒).

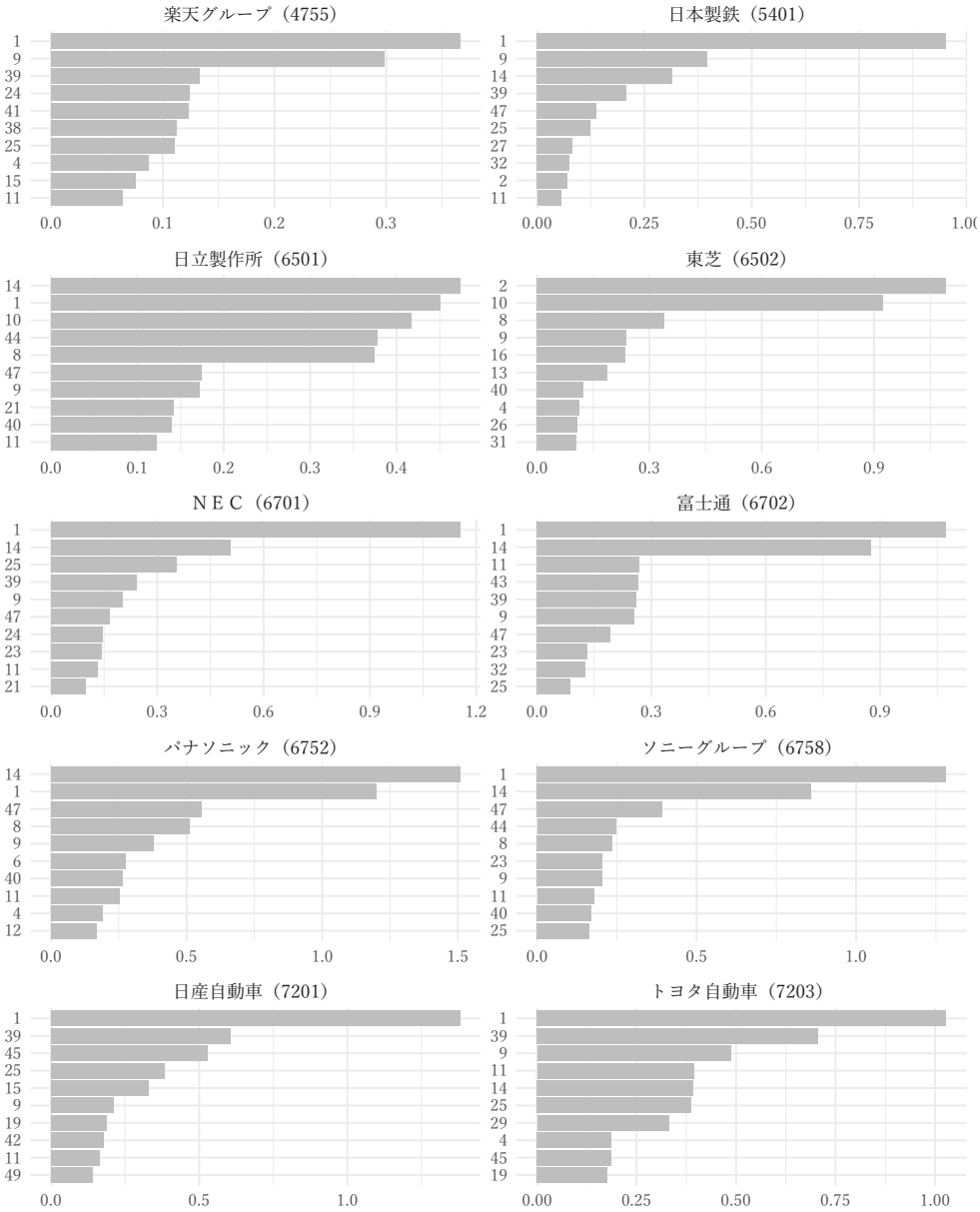
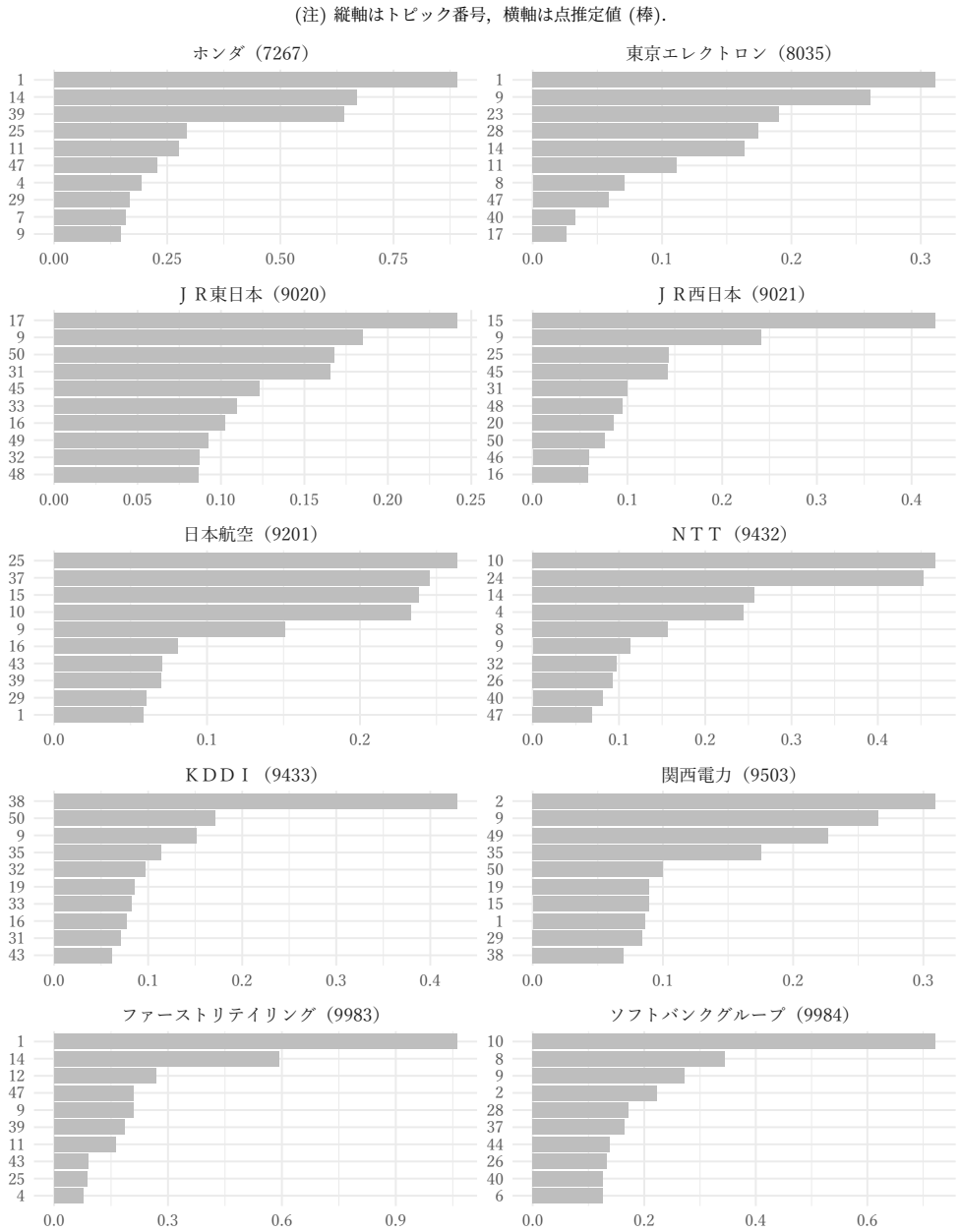


図 6　銘柄別の取引金額の変化に対する各トピックの回帰係数 (2)



5.5　トピックと株価リターンの関係

多くの銘柄において回帰係数上位に共通して現れたトピック 1，トピック 9，トピック 10 と株価リターンとの関係について分析する．分析にあたって，学習したトピックを次の手法によって定量化する．ニュースの文書 – トピック分布を日次で合計し，同日の株価変化の方向性によって符号を調整した値をスコアとする．具体的には，t 期における対東証株価指数 (TOPIX) 相対対数株価と銘柄 j のニュース集合をそれぞれ $P_{j,t}$，$D_{j,t}$

とし，銘柄 j のトピック k による t 期におけるスコアを，

$$\xi_{j,k,t} = \mathrm{sgn}(P_{j,t} - P_{j,t-1}) \sum_{d \in D_{j,t}} \theta_{d,k},$$

によって定義する．ここで，sgn(·) は符号関数である．トピックの発生が同時点の株価に与えるインパクトについては株式市場を参照し，期先のパスへの影響分析を行う．

分析モデルにはローカル・プロジェクション (Jordà, 2005) を用いて，インパルス反応分析を行う．具体的には，次の線形回帰モデルを推定する．

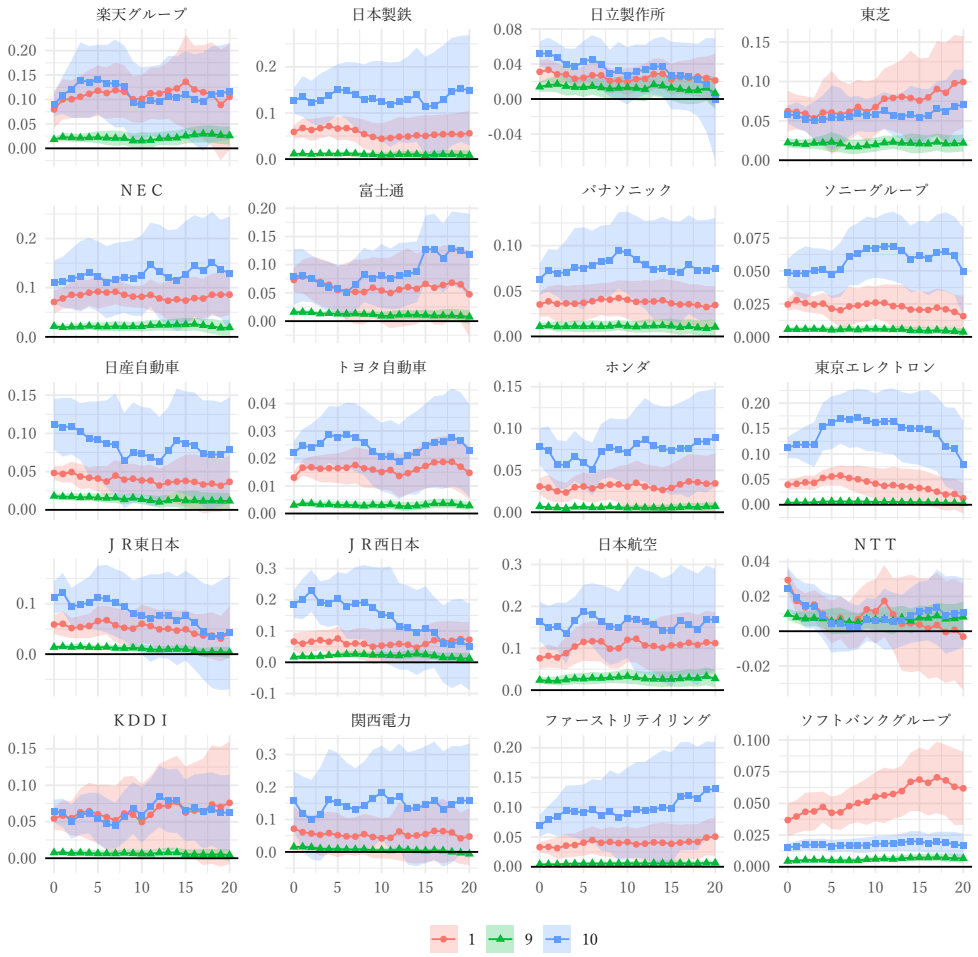
$$\Delta^h P_{j,t} = \alpha^h_{j,k} + \gamma^h_{j,k} \xi_{j,k,t} + \epsilon^h_{j,k,t}.$$

ここで，$\Delta^h P_{j,t} = P_{j,t+h} - P_{j,t-1}$ とし，$\gamma^h_{j,k}$ をスコア $\xi_{j,k,t}$ に対する h 期先までの銘柄 j の累積株価リターンの反応とし，これを $h = 0, 1, ..., 20$ について推定し，インパルス反応関数とする．推定値の信頼区間は誤差項の不均一分散と系列相関に頑健な HAC による標準誤差から計算する．また，株価のモメンタム効果をコントロールするため，$\Delta^0 P_{j,t}$ の 5 次までのラグ項を含めて推定する．以上の設定より，本分析によるインパルス反応関数は各トピックの構成比率が 100% となるニュースが 1 件発生したときの超過収益率への累積効果として解釈できる．

分析結果を図 7 に示した．全体的な傾向として，トピック 1 およびトピック 10 と比較して，トピック 9 への株価の反応は限定的である．すなわち，トピック 9 は取引金額との関連性は高いものの，株価リターンへの影響は小さく，投資家のセンチメントへ作用する部分が大きい可能性がある．一方で，各トピックの株価へのインパクトは一過性ではなく，短期的な株価のリバウンドは見られない．ニュースが将来の株価形成に対するファンダメンタルズ情報を持つという理論を情報理論 (Information Theory) という．概ね全ての銘柄について，明らかに業績や企業活動等のファンダメンタルズに関するトピック 1 とトピック 10 だけでなく，市況に関するトピック 9 も期先の株価に影響を与えており，情報理論をサポートする結果となった．これは，Tetlock (2007) を日本のデータに応用した沖本・平澤 (2014) と同様の結果であるが，個別銘柄を分析対象としている点，既存のカテゴリーではなく，ニュースの潜在的なトピックをスコアのベースとしている点で分析の前提が異なる．最後に銘柄固有の結果として，ＮＴＴや関西電力は 5 日程度で影響が収束し，東芝やソフトバンクグループは影響が逓増するような傾向が確認できる．共通のトピックに対して，銘柄の特性によってニュースへの反応が異なる可能性が示唆された．

図 7　トピックと株価リターンのインパルス反応関数

(注) 縦軸はインパルス反応関数の点推定値 (線) と 95% 信頼区間 (エラーバンド)，横軸は期間.



6　むすび

本研究では，金融市場分析のためのニュースデータのスコアリングの従来手法についての問題点と，既存モデルのファイナンス領域における応用上の不備を指摘し，それらを克服する新しいモデルを開発した．ニュースと金融市場の関連性を分析するためには，単なる件数カウントのような内容を無視した手法ではなく，ニュースの内容と分析の目的を両方考慮した分析が必要である．また，内容の考慮にあたっては，二元的な評価ではなく，金融市場の複雑性やニュースの多義性を踏まえた多元的な評価を行わなくてはならない．左記に部分的に適うモデルとしてトピックモデルがある．ただし，ファイナンス分野におけるニュースは，複数のラベルと複数の銘柄が同時に関連付けられており，前者については補助情報として，後者については予測対象として分析する必要がある．さらに後者については，同じニュースであっても，銘柄によって異なる反応を示すことがあり，そのような構造を全て表現できるトピックモデルは，我々の知りうる限りこれまで提案されていない．そこで，Roberts et al. (2016) の構造トピックモデルおよび Blei and Mcauliffe (2007) の sLDA のハイブリッドモデルを土台に，予測部分にマルチラベルモデルを導入することで，銘柄固有のランダム効果

を許容し，複数ラベルの考慮と複数銘柄の同時分析を実現するモデルを提案した．また，変分ベイズ法による学習アルゴリズムを示した．

「日本経済新聞　電子版」のニュースと日本の株式市場の個別銘柄のデータを使用した実証分析では，ラベルデータをニュースに紐づく銘柄の業種，教師データを各銘柄の取引金額として，提案モデルを応用した．各トピックの特徴的なパラメータを考察することで，分析対象銘柄やラベルに整合的なトピックが学習されたことを確認し，機械的に学習したトピックについて解釈を与えた．提案モデル独自の分析結果として，ニュースが取引金額に与える影響について，トピック・銘柄固有の反応を示すことを定量的に示し，提案モデルの有用性を確認する結果を得た．また，提案モデルによって学習したトピックを用いてスコアを作成し，将来の株価リターンへの影響を分析した．ニュースのトピックの違いによって，影響の大きさが異なり，業績や企業活動に関するトピックの影響が大きく，市況に関するトピックは影響が限定的であることを示した．また，市況に関するトピックも含めて，概ね全てのトピック・銘柄について，ニュースの株価へのインパクトは持続的であり，ニュースが将来の株価のファンダメンタルズに関する情報を持つという情報理論をサポートする結果を得た．

参考文献

Baker, Scott R., Nicholas Bloom, and Steven J. Davis (2016) "Measuring Economic Policy Uncertainty," *The Quarterly Journal of Economics*, Vol. 131, No. 4, pp. 1593–1636.

Berry, Thomas D. and Keith M. Howe (1994) "Public Information Arrival," *The Journal of Finance*, Vol. 49, No. 4, pp. 1331–1346.

Blei, David M. and John D. Lafferty (2007) "A Correlated Topic Model of Science," *The Annals of Applied Statistics*, Vol. 1, No. 1, pp. 17–35.

Blei, David M. and Jon D. Mcauliffe (2007) "Supervised Topic Models," in *Neural Information Processing Systems*, Vol. 20.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003) "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, Vol. 3, pp. 993—1022.

Eisenstein, Jacob, Amr Ahmed, and Eric P. Xing (2011) "Sparse Additive Generative Models of Text," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pp. 1041–1048.

Gelman, Andrew and Jennifer Hill (2006) *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Analytical Methods for Social Research: Cambridge University Press.

Jordà, Òscar (2005) "Estimation and Inference of Impulse Responses by Local Projections," *American Economic Review*, Vol. 95, No. 1, pp. 161–182.

Kearney, Colm and Sha Liu (2014) "Textual Sentiment in Finance: A Survey of Methods and Models," *International Review of Financial Analysis*, Vol. 33, pp. 171–185.

Kudo, Taku, Kaoru Yamamoto, and Yuji Matsumoto (2004) "Applying Conditional Random Fields to Japanese Morphological Analysis," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 230–237.

Li, Feng (2010) "The Information Content of Forward-Looking Statements in Corporate Filings–A Naïve Bayesian Machine Learning Approach," *Journal of Accounting Research*, Vol. 48, No. 5, pp. 1049–

1102.

Mimno, David and Andrew McCallum (2008) “Topic Models Conditioned on Arbitrary Features with Dirichlet-Multinomial Regression,” in *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, Vol. 24, pp. 411–418.

Mitchell, Mark L. and J. Harold Mulherin (1994) “The Impact of Public Information on the Stock Market,” *The Journal of Finance*, Vol. 49, No. 3, pp. 923–950.

Perotte, Adler, Frank Wood, Noemie Elhadad, and Nicholas Bartlett (2011) “Hierarchically Supervised Latent Dirichlet Allocation,” in *Neural Information Processing Systems*, Vol. 24, pp. 2609–2617.

Ramage, Daniel, David Hall, Ramesh Nallapati, and Christopher D. Manning (2009) “Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-Labeled Corpora,” in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 248–256.

Roberts, Margaret E., Brandon M. Stewart, and Edoardo M. Airoldi (2016) “A Model of Text for Experimentation in the Social Sciences,” *Journal of the American Statistical Association*, Vol. 111, No. 515, pp. 988–1003.

Roll, Richard (1988) “R^2,” *The Journal of Finance*, Vol. 43, No. 3, pp. 541–566.

Rosen-Zvi, Michal, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth (2004) “The Author-Topic Model for Authors and Documents,” in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, p. 487–494.

Sato, Toshinori (2015) “Neologism Dictionary Based on the Language Resources on the Web for Mecab,” URL: `https://github.com/neologd/mecab-ipadic-neologd`.

Tetlock, Paul C. (2007) “Giving Content to Investor Sentiment: The Role of Media in the Stock Market,” *The Journal of Finance*, Vol. 62, No. 3, pp. 1139–1168.

——— (2010) “Does Public Financial News Resolve Asymmetric Information?” *The Review of Financial Studies*, Vol. 23, No. 9, pp. 3520–3557.

Yono, Kyoto, Hiroki Sakaji, Hiroyasu Matsushima, Takashi Shimada, and Kiyoshi Izumi (2020) “Construction of Macroeconomic Uncertainty Indices for Financial Market Analysis Using a Supervised Topic Model,” *Journal of Risk and Financial Management*, Vol. 13, No. 4, p. 79.

Zhu, Jun, Amr Ahmed, and Eric P. Xing (2012) “MedLDA: Maximum Margin Supervised Topic Models,” *Journal of Machine Learning Research*, Vol. 13, No. 74, pp. 2237–2278.

沖本竜義・平澤英司 (2014)「ニュース指標による株式市場の予測可能性」,『証券アナリストジャーナル』, 第 52 巻, 第 4 号, 67–75 頁.

佐藤一誠 (2015)『トピックモデルによる統計的潜在意味解析』, コロナ社.

佐藤敏紀・橋本泰一・奥村学 (2016)「単語分かち書き用辞書生成システム NEologd の運用 — 文書分類を例にして —」,『自然言語処理研究会研究報告』, NL–229–15 頁, 情報処理学会.

——— (2017)「単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討」,『言語処理学会第 23 回年次大会 (NLP2017)』, NLP2017–B6–1 頁, 言語処理学会.

付録 A　モデル学習アルゴリズム

本提案モデルの学習アルゴリズムには，トピックモデルの学習でよく用いられる変分ベイズ法を採用している．変分ベイズ法では，対数周辺尤度と近似事後分布から導出した下限 (変分下限) を，近似事後分布とハイパーパラメータによって最大化する．本章では，対数周辺尤度と変分下限，近似事後分布とハイパーパラメータの更新式を導出する．

A.1　対数周辺尤度と変分下限の導出

対数尤度を潜在変数によって周辺化すると，対数周辺尤度，

$$
\begin{aligned}
L &= \log p(\boldsymbol{w}, \boldsymbol{y} | \Gamma, \boldsymbol{x}, \Sigma, \boldsymbol{\beta}, \boldsymbol{\nu}^2, \boldsymbol{\mu}, \upsilon, \Omega) \\
&= \log \iiiint \sum_{\boldsymbol{z}} p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{y}, \boldsymbol{\eta}, \Lambda | \Gamma, \boldsymbol{x}, \Sigma, \boldsymbol{\beta}, \boldsymbol{\nu}^2, \boldsymbol{\mu}, \upsilon, \Omega) d\boldsymbol{\theta} d\boldsymbol{\phi} d\boldsymbol{\eta} d\Lambda \\
&= \log \iiiint \sum_{\boldsymbol{z}} \Bigg[\left(\prod_{d=1}^{M} \prod_{i=1}^{n_d} p(w_{d,i} | \boldsymbol{\phi}_{z_{d,i}}) p(z_{d,i} | \boldsymbol{\theta}_d) \right) \\
&\quad \times \left(\prod_{d=1}^{M} p(\boldsymbol{\theta}_d | \boldsymbol{x}_d, \Gamma, \Sigma) \right) \left(\prod_{d=1}^{K} p(\boldsymbol{\phi}_k | \boldsymbol{\beta}) \right) \\
&\quad \times \left(\prod_{d=1}^{M} \prod_{j \in J_d} p(y_{d,j} | \boldsymbol{z}_d, \boldsymbol{\eta}_j, \nu_j^2) \right) \left(\prod_{j=1}^{J} p(\boldsymbol{\eta}_j | \boldsymbol{\mu}, \Lambda) \right) p(\Lambda | \upsilon, \Omega) \Bigg] d\boldsymbol{\theta} d\boldsymbol{\phi} d\boldsymbol{\eta} d\Lambda,
\end{aligned}
$$

が得られる．対数関数の中に積分と総和・総乗演算子を含む複雑な式となっており，直接に尤度最大化問題を解くことができない．そこで，潜在変数の近似事後分布 $q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\eta}, \Lambda)$ を用いて，イェンセンの不等式から対数周辺尤度の下限を次の通り求める．

$$
\begin{aligned}
L &= \log \iiiint \sum_{\boldsymbol{z}} q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\eta}, \Lambda) \\
&\quad \times \frac{p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{y}, \boldsymbol{\eta}, \Lambda | \Gamma, \boldsymbol{x}, \Sigma, \boldsymbol{\beta}, \boldsymbol{\nu}^2, \boldsymbol{\mu}, \upsilon, \Omega)}{q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\eta}, \Lambda)} d\boldsymbol{\theta} d\boldsymbol{\phi} d\boldsymbol{\eta} d\Lambda \\
&\geq \iiiint \sum_{\boldsymbol{z}} q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\eta}, \Lambda) \\
&\quad \times \log \frac{p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{y}, \boldsymbol{\eta}, \Lambda | \Gamma, \boldsymbol{x}, \Sigma, \boldsymbol{\beta}, \boldsymbol{\nu}^2, \boldsymbol{\mu}, \upsilon, \Omega)}{q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\eta}, \Lambda)} d\boldsymbol{\theta} d\boldsymbol{\phi} d\boldsymbol{\eta} d\Lambda \\
&\equiv F[q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\eta}, \Lambda)].
\end{aligned}
$$

$F[q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\eta}, \Lambda)]$ を変分下限という．

ここで，近似事後分布 $q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\eta}, \Lambda)$ について，次の因子分解の仮定をおく．

$$
\begin{aligned}
&q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\eta}, \Lambda) \\
&\quad = q(\boldsymbol{z}) q(\boldsymbol{\theta}) q(\boldsymbol{\phi}) q(\boldsymbol{\eta}) q(\Lambda) \\
&\quad = \left(\prod_{d=1}^{M} \prod_{i=1}^{n_d} \prod_{k=1}^{K} q(z_{d,i} = k) \right) \left(\prod_{d=1}^{M} q(\boldsymbol{\theta}_d) \right) \left(\prod_{k=1}^{K} q(\boldsymbol{\phi}_k) \right) \left(\prod_{j=1}^{J} q(\boldsymbol{\eta}_j) \right) q(\Lambda).
\end{aligned}
$$

これにより，変分下限について計算を進めると，

$$
\begin{aligned}
&F[q(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\eta},\Lambda)] \\
&= \iiiint \sum_{\boldsymbol{z}} q(\boldsymbol{z})q(\boldsymbol{\theta})q(\boldsymbol{\phi})q(\boldsymbol{\eta})q(\Lambda) \\
&\quad \times \log \frac{p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\phi})p(\boldsymbol{z}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{x},\Gamma,\Sigma)p(\boldsymbol{\phi}|\boldsymbol{\beta})p(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{\eta},\boldsymbol{\nu}^2)p(\boldsymbol{\eta}|\boldsymbol{\mu},\Lambda)p(\Lambda|\upsilon,\Omega)}{q(\boldsymbol{z})q(\boldsymbol{\theta})q(\boldsymbol{\phi})q(\boldsymbol{\eta})q(\Lambda)} d\boldsymbol{\theta}d\boldsymbol{\phi}d\boldsymbol{\eta}d\Lambda \\
&= \iint \sum_{\boldsymbol{z}} q(\boldsymbol{z})q(\boldsymbol{\theta})q(\boldsymbol{\phi}) \log p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\phi})p(\boldsymbol{z}|\boldsymbol{\theta})d\boldsymbol{\theta}d\boldsymbol{\phi} - \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log q(\boldsymbol{z}) \\
&\quad - \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{x},\Gamma,\Sigma)} d\boldsymbol{\theta} - \int q(\boldsymbol{\phi}) \log \frac{q(\boldsymbol{\phi})}{p(\boldsymbol{\phi}|\boldsymbol{\beta})} d\boldsymbol{\phi} \\
&\quad + \int \sum_{\boldsymbol{z}} q(\boldsymbol{z})q(\boldsymbol{\eta}) \log p(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{\eta},\boldsymbol{\nu}^2) d\boldsymbol{\eta} \\
&\quad - \iint q(\boldsymbol{\eta})q(\Lambda) \log \frac{q(\boldsymbol{\eta})}{p(\boldsymbol{\eta}|\boldsymbol{\mu},\Lambda)} d\boldsymbol{\eta}d\Lambda - \int q(\Lambda) \log \frac{q(\Lambda)}{p(\Lambda|\upsilon,\Omega)} d\Lambda \\
&= \iint \sum_{d=1}^{M} \sum_{i=1}^{n_d} \sum_{z_{d,i}} q(z_{d,i})q(\boldsymbol{\theta}_d)q(\boldsymbol{\phi}) \log p(w_{d,i}|z_{d,i},\boldsymbol{\phi})p(z_{d,i}|\boldsymbol{\theta}_d) d\boldsymbol{\theta}_d d\boldsymbol{\phi} \qquad (1) \\
&\quad - \sum_{d=1}^{M} \sum_{i=1}^{n_d} \sum_{k=1}^{K} q(z_{d,i}=k) \log q(z_{d,i}=k) \qquad (2) \\
&\quad - \sum_{d=1}^{M} \int q(\boldsymbol{\theta}_d) \log \frac{q(\boldsymbol{\theta}_d)}{p(\boldsymbol{\theta}_d|\boldsymbol{x}_d,\Gamma,\Sigma)} d\boldsymbol{\theta}_d \qquad (3) \\
&\quad - \sum_{k=1}^{K} \int q(\boldsymbol{\phi}_k) \log \frac{q(\boldsymbol{\phi}_k)}{p(\boldsymbol{\phi}_k|\boldsymbol{\beta})} d\boldsymbol{\phi}_k \qquad (4) \\
&\quad + \sum_{d=1}^{M} \sum_{j \in J_d} \int \sum_{\boldsymbol{z}_d} q(\boldsymbol{z}_d)q(\boldsymbol{\eta}_j) \log p(y_{d,j}|\boldsymbol{z}_d,\boldsymbol{\eta}_j,\nu_j^2) d\boldsymbol{\eta}_j \qquad (5) \\
&\quad - \sum_{j=1}^{J} \iint q(\boldsymbol{\eta}_j)q(\Lambda) \log \frac{q(\boldsymbol{\eta}_j)}{p(\boldsymbol{\eta}_j|\boldsymbol{\mu},\Lambda)} d\boldsymbol{\eta}_j d\Lambda \qquad (6) \\
&\quad - \int q(\Lambda) \log \frac{q(\Lambda)}{p(\Lambda|\upsilon,\Omega)} d\Lambda, \qquad (7)
\end{aligned}
$$

が得られる．(1)(2)(4) 式はベーシックなトピックモデルと共通する項であり，(3)(5)(6)(7) 式は拡張によって修正または追加された項である．

変分ベイズ法では，近似事後分布の更新とハイパーパラメータの更新を順に繰り返すことによって，変分下限を最大化する．次節以降で，提案モデルにおける拡張によって，先行研究から修正が必要となる固有の近似事後分布とパラメータの更新式の導出を行う[9]．

A.2 近似事後分布の更新

ここで，$KL[q||p]$ は確率分布 q と p のカルバック・ライブラー情報量を表す．ここで，$\mathbf{1}(\cdot)$ は括弧内の条件式が真の時は 1，偽の時は 0 を返す指示関数 $N(\cdot)$ は正規分布の確率密度関数である．ここで，$Dir(\cdot)$ はディ

[9] 本稿で省略した各推定式とその導出については佐藤 (2015) に詳しい．

リクレ分布の確率密度関数である．$n_{k,v}$ は，語彙 v がトピック k となる回数であるここで，$\Psi(\cdot)$ はディガンマ関数である．

A.2.1 周辺分布 $q(\boldsymbol{\eta}_j)$

$q(\boldsymbol{\eta}_j)$ を平均 $\hat{\boldsymbol{\mu}}_j$，分散 $\hat{\Lambda}$ の正規分布 $N(\boldsymbol{\eta}_j|\hat{\boldsymbol{\mu}}_j, \hat{\Lambda})$ と仮定し，変分下限を最大化するパラメータ $\hat{\boldsymbol{\mu}}_j$ と $\hat{\Lambda}$ を求める．まず，変分下限のうち $\hat{\boldsymbol{\mu}}_j$ に関係する項は，

$$\tilde{F}[\hat{\boldsymbol{\mu}}_j] = \frac{1}{2\nu_j^2} \sum_{d \in D_j} \left(2y_{d,j} \hat{\boldsymbol{\mu}}_j^\top \mathbb{E}_{q(\boldsymbol{z}_d)}[\bar{\boldsymbol{z}}_d] - \hat{\boldsymbol{\mu}}_j^\top \mathbb{E}_{q(\boldsymbol{z}_d)}[\bar{\boldsymbol{z}}_d \bar{\boldsymbol{z}}_d^\top] \hat{\boldsymbol{\mu}}_j \right) - \frac{1}{2} (\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu})^\top \mathbb{E}_{q(\Lambda)}[\Lambda]^{-1} (\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}).$$

ここで，$\mathbb{E}[\cdot]$ は期待値を表し，$D_j = \{d \mid j \in J_d\}$ である．$\hat{\boldsymbol{\mu}}_j$ について，1 階条件より，

$$\begin{aligned} &\frac{\partial \tilde{F}[\hat{\boldsymbol{\mu}}_j]}{\partial \hat{\boldsymbol{\mu}}_j} = 0 \\ &\Leftrightarrow \hat{\boldsymbol{\mu}}_j = \left(\frac{1}{\nu_j^2} \sum_{d \in D_j} \mathbb{E}_{q(\boldsymbol{z}_d)}[\bar{\boldsymbol{z}}_d \bar{\boldsymbol{z}}_d^\top] + \mathbb{E}_{q(\Lambda)}[\Lambda]^{-1} \right)^{-1} \\ &\qquad \times \left(\frac{1}{\nu_j^2} \sum_{d \in D_j} y_{d,j} \mathbb{E}_{q(\boldsymbol{z}_d)}[\bar{\boldsymbol{z}}_d] + \mathbb{E}_{q(\Lambda)}[\Lambda]^{-1} \boldsymbol{\mu} \right). \end{aligned}$$

次に，変分下限のうち $\hat{\Lambda}$ に関係する項は，

$$\tilde{F}[\hat{\Lambda}] = -\frac{J}{2} \left(-\log|\hat{\Lambda}| + \mathrm{Tr}(\mathbb{E}_{q(\Lambda)}[\Lambda]^{-1} \hat{\Lambda}) \right) - \sum_{d=1}^{M} \sum_{j \in J_d} \frac{1}{2\nu_j^2} \mathrm{Tr}(\mathbb{E}_{q(\boldsymbol{z}_d)}[\bar{\boldsymbol{z}}_d \bar{\boldsymbol{z}}_d^\top] \hat{\Lambda}).$$

$\hat{\Lambda}$ について，1 階条件より，

$$\frac{\partial \tilde{F}[\hat{\Lambda}]}{\partial \hat{\Lambda}} = 0 \Leftrightarrow \hat{\Lambda} = \left(\mathbb{E}_{q(\Lambda)}[\Lambda]^{-1} + \frac{1}{J} \sum_{d=1}^{M} \sum_{j \in J_d} \frac{1}{\nu_j^2} \mathbb{E}_{q(\boldsymbol{z}_d)}[\bar{\boldsymbol{z}}_d \bar{\boldsymbol{z}}_d^\top] \right)^{-1}.$$

A.2.2 周辺分布 $q(\Lambda)$

変分下限のうち $q(\Lambda)$ に関係する項は，

$$\tilde{F}[q(\Lambda)] = \int q(\Lambda) \sum_{j=1}^{J} \int q(\boldsymbol{\eta}_j) \log p(\boldsymbol{\eta}_j|\boldsymbol{\mu}, \Lambda) d\boldsymbol{\eta}_j d\Lambda - \int q(\Lambda) \log \frac{q(\Lambda)}{p(\Lambda|\upsilon, \Omega)} d\Lambda.$$

変分法により，

$$\frac{\delta \tilde{F}[q(\Lambda)]}{\delta q(\Lambda)} = 0 \Leftrightarrow \sum_{j=1}^{J} \int q(\boldsymbol{\eta}_j) \log p(\boldsymbol{\eta}_j|\boldsymbol{\mu}, \Lambda) d\boldsymbol{\eta}_j - \log \frac{q(\Lambda)}{p(\Lambda|\upsilon, \Omega)} - 1 = 0.$$

これを解くと，

$$q(\Lambda) = IW(\Lambda|\hat{\upsilon}, \hat{\Omega}).$$

ここで，$IW(\cdot)$ は逆ウィシャート分布の確率密度関数であり，スケールと自由度パラメータはそれぞれ，

$$\begin{aligned}\hat{\upsilon} &= \upsilon + J,\\ \hat{\Omega} &= \Omega + \sum_{j=1}^{J}(\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu})(\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu})^\top.\end{aligned}$$

期待値は，

$$\mathbb{E}_{q(\Lambda)}[\Lambda] = \frac{1}{\hat{\upsilon} - K - 1}\hat{\Omega}.$$

A.2.3 周辺分布 $q(z_{d,i} = k)$

変分下限のうち $q(z_{d,i} = k)$ に関係する項は，

$$\begin{aligned}&\tilde{F}[q(z_{d,i} = k)]\\ &= \sum_{k=1}^{K} q(z_{d,i} = k) \iint q(\boldsymbol{\theta}_d)q(\boldsymbol{\phi}_k)\log p(w_{d,i}|\boldsymbol{\phi}_k)p(z_{d,i} = k|\boldsymbol{\theta}_d)d\boldsymbol{\phi}_k d\boldsymbol{\theta}_d\\ &\quad - \sum_{k=1}^{K} q(z_{d,i} = k)\log q(z_{d,i} = k)\\ &\quad + \sum_{j\in J_d}\frac{1}{2\nu_j^2}\left(2y_{d,j}\hat{\boldsymbol{\mu}}_j^\top \mathbb{E}_{q(\boldsymbol{z}_d)}[\bar{\boldsymbol{z}}_d] - \hat{\boldsymbol{\mu}}_j^\top \mathbb{E}_{q(\boldsymbol{z}_d)}[\bar{\boldsymbol{z}}_d\bar{\boldsymbol{z}}_d^\top]\hat{\boldsymbol{\mu}}_j - \mathrm{Tr}(\mathbb{E}_{q(\boldsymbol{z}_d)}[\bar{\boldsymbol{z}}_d\bar{\boldsymbol{z}}_d^\top]\hat{\Lambda})\right).\end{aligned}$$

ここで，$(\hat{\Lambda})_{k,k}$ と $(\hat{\Lambda})_k$ をそれぞれ $\hat{\Lambda}$ の k 番目の対角要素，k 列目を抜き出したベクトルとすると，

$$\begin{aligned}\hat{\boldsymbol{\mu}}_j^\top \mathbb{E}_{q(\boldsymbol{z}_d)}[\bar{\boldsymbol{z}}_d] &= \frac{1}{n_d}\sum_{i=1}^{n_d}\sum_{k=1}^{K}\hat{\mu}_{j,k}q(z_{d,i} = k),\\ \hat{\boldsymbol{\mu}}_j^\top \mathbb{E}_{q(\boldsymbol{z}_d)}[\bar{\boldsymbol{z}}_d\bar{\boldsymbol{z}}_d^\top]\hat{\boldsymbol{\mu}}_j &= \frac{1}{n_d^2}\sum_{i=1}^{n_d}\left(\sum_{k=1}^{K}\hat{\mu}_{j,k}^2 q(z_{d,i} = k) + \sum_{k=1}^{K}\hat{\mu}_{j,k}q(z_{d,i} = k)\sum_{i'\neq i}^{n_d}\boldsymbol{\zeta}_{d,i'}^\top\hat{\boldsymbol{\mu}}_j\right),\\ \mathrm{Tr}(\mathbb{E}_{q(\boldsymbol{z}_d)}[\bar{\boldsymbol{z}}_d\bar{\boldsymbol{z}}_d^\top]\hat{\Lambda}) &= \frac{1}{n_d^2}\sum_{i=1}^{n_d}\left(\sum_{k=1}^{K}(\hat{\Lambda})_{k,k}q(z_{d,i} = k) + \sum_{k=1}^{K}q(z_{d,i} = k)\sum_{i'\neq i}^{n_d}\boldsymbol{\zeta}_{d,i'}^\top(\hat{\Lambda})_k\right).\end{aligned}$$

これらを用いて，1 階条件より，

$$\begin{aligned}&\frac{\partial\tilde{F}[q(z_{d,i})]}{\partial q(z_{d,i} = k)} = 0\\ &\quad\Leftrightarrow \int q(\boldsymbol{\phi}_k)\log\phi_{k,w_{d,i}}d\boldsymbol{\phi}_k + \int q(\boldsymbol{\theta}_d)\log\frac{\exp(\theta_{d,k})}{\sum_{k'=1}^{K}\exp(\theta_{d,k'})}d\boldsymbol{\theta}_d\\ &\qquad - \log q(z_{d,i} = k) - 1\\ &\qquad + \sum_{j\in J_d}\left[\frac{y_{d,j}\hat{\mu}_{j,k}}{\nu_j^2 n_d} - \frac{1}{2\nu_j^2 n_d^2}\left(\hat{\mu}_{j,k}^2 + 2\hat{\mu}_{j,k}\sum_{i\neq i'}^{n_d}\boldsymbol{\zeta}_{d,i'}^\top\hat{\boldsymbol{\mu}}_j\right.\right.\\ &\qquad\qquad\left.\left. + (\hat{\Lambda})_{k,k} + 2\sum_{i'\neq i}^{n_d}\boldsymbol{\zeta}_{d,i'}^\top(\hat{\Lambda})_k\right)\right] = 0.\end{aligned}$$

これを解くと，

$$
\begin{aligned}
&q(z_{d,i}=k)\\
&\quad\propto \exp\left(\mathbb{E}_{q(\boldsymbol{\phi}_k)}[\log\phi_{k,w_{d,i}}]\right)\exp\left(\mathbb{E}_{q(\boldsymbol{\theta}_d)}[\theta_{d,k}]\right)\\
&\qquad\times\prod_{j\in J_d}\exp\Bigg(\frac{y_{d,j}\hat{\mu}_{j,k}}{\nu_j^2 n_d}-\frac{1}{2\nu_j^2 n_d^2}\Bigg(\hat{\mu}_{j,k}^2+2\hat{\mu}_{j,k}\sum_{i\neq i'}^{n_d}\boldsymbol{\zeta}_{d,i'}^\top\hat{\boldsymbol{\mu}}_j\\
&\qquad\qquad+(\hat{\Lambda})_{k,k}+2\sum_{i'\neq i}^{n_d}\boldsymbol{\zeta}_{d,i'}^\top(\hat{\Lambda})_k\Bigg)\Bigg).
\end{aligned}
$$

A.3 ハイパーパラメータの更新

A.3.1 パラメータ $\boldsymbol{\mu}$

変分下限のうち，$\boldsymbol{\mu}$ に関係する項は，

$$
\tilde{F}[\boldsymbol{\mu}]=-\frac{1}{2}\sum_{j=1}^{J}\Bigg((\hat{\boldsymbol{\mu}}_j-\boldsymbol{\mu})^\top\mathbb{E}_{q(\Lambda)}[\Lambda]^{-1}(\hat{\boldsymbol{\mu}}_j-\boldsymbol{\mu})\Bigg).
$$

$\boldsymbol{\mu}$ について，1 階条件より，

$$
\frac{\partial\tilde{F}[\boldsymbol{\mu}]}{\partial\boldsymbol{\mu}}=0\Leftrightarrow\boldsymbol{\mu}=\frac{1}{J}\sum_{j=1}^{J}\hat{\boldsymbol{\mu}}_j.
$$

A.3.2 パラメータ $\boldsymbol{\nu}^2$

変分下限のうち ν_j^2 に関係する項は，

$$
\tilde{F}[\nu_j^2]=-\frac{1}{2}\sum_{d\in D_j}\Bigg(\log(\nu_j^2)+\frac{1}{\nu_j^2}\mathbb{E}_{q(\boldsymbol{z}_d)q(\boldsymbol{\eta}_j)}\left[(y_{d,j}-\boldsymbol{\eta}_j^\top\bar{\boldsymbol{z}}_d)^2\right]\Bigg).
$$

ν_j^{-2} について，1 階条件より，

$$
\begin{aligned}
&\frac{\partial\tilde{F}[\nu_j^2]}{\partial\nu_j^{-2}}=0\\
&\quad\Leftrightarrow\nu_j^2=\frac{1}{\sum_{d\in D_j}1}\sum_{d\in D_j}\Bigg(y_{d,j}^2-2y_{d,j}\hat{\boldsymbol{\mu}}_j^\top\mathbb{E}_{q(\boldsymbol{z}_d)}[\bar{\boldsymbol{z}}_d]\\
&\qquad\qquad+\hat{\boldsymbol{\mu}}_j^\top\mathbb{E}_{q(\boldsymbol{z}_d)}[\bar{\boldsymbol{z}}_d\bar{\boldsymbol{z}}_d^\top]\hat{\boldsymbol{\mu}}_j+\mathrm{Tr}(\mathbb{E}_{q(\boldsymbol{z}_d)}[\bar{\boldsymbol{z}}_d\bar{\boldsymbol{z}}_d^\top]\hat{\Lambda})\Bigg).
\end{aligned}
$$