統計数理(2018)
第 66 巻 第 2 号 235–247

特集「サービス科学の今」
[研究詳解]

大規模集計POSデータの高次元スパースモデリング

李 銀星[1]・照井 伸彦[2]

(受付 2017 年 11 月 16 日；改訂 2018 年 5 月 17 日；採択 6 月 7 日)

要　　旨

多様な消費者ニーズをきめ細かく捉えて顧客を獲得して維持するための効果的マーケティングのために，主体(消費者)の異質性の統計モデリングが他の分野に先駆けて開発された．他方，実店舗において，個別対応は必ずしも容易ではないのも現実である．本稿では集計 POS データに対して機械学習などの新しい統計分析による高度情報処理を適用することにより，多くの実店舗で活用できる大規模データを活用したマーケティングモデルの可能性を展望する．高次元データについては，2 種類の次元圧縮法，すなわち，トピックモデルによる次元圧縮と購買文脈による部分市場分解，階層因子回帰モデルによる次元圧縮とパラメータの高次元空間への還元が議論される．

全商品データを分析に取り入れることで，目的変数を説明する意外な変数の発見のみならずその量的関係が弾力性の形で測定可能となり，実店舗のきめ細かいマーケティング戦略に有用な情報が提供可能となることを展望する．

キーワード：集計 POS データ，購買状況の異質性，トピックモデル，高次元スパースデータ，階層因子回帰．

1. はじめに

顧客データベースの整備を背景にして，多様な消費者ニーズをきめ細かく捉えて顧客を獲得して維持するための効果的マーケティングのために，主体(消費者)の異質性の統計モデリングが他の分野に先駆けて開発された．例えば，顧客データベースから消費者選択を階層ベイズモデルでモデル化して顧客ごとの市場反応を推定する Rossi et al. (1996)は，個別化モデリングの先駆けである．価格感度の違いに応じて個別化したクーポンの発行による個別対応は Rossi et al. (1996)や Terui and Dahana (2006)などにおいて提案されている．これらの包括的な説明は照井 (2018)でなされている．その後，個別対応のマーケティングは，機械学習の手法も巻き込んで発展し，E ビジネスの世界で日常的に行われている．

実店舗においてはマーケティングの個別対応は必ずしも容易ではないのも現実である．POS システムはほとんどの実店舗で導入されており，集計データは無自覚的に日々蓄積されている．これら集計データのマーケティング分析は古くから回帰や時系列モデルにより行われてきた．Hanssen et al. (2001)では，集計データのマーケティングモデルについて包括的に説明してい

[1] 東北大学大学院 経済学研究科：〒 980–8576 宮城県仙台市青葉区川内；dgod1028@gmail.com
[2] 東北大学大学院 経済学研究科：〒 980–8576 宮城県仙台市青葉区川内；terui@tohoku.ac.jp

るが，これらは分析対象を特定カテゴリーに限定した低次元空間上での少数変数間のモデリングである．他方，アソシエーションルールによるマーケットバスケット分析などカテゴリーの枠を超えた高次元変数の分析も行われているが，マーケティング変数と市場構造の関係などマネジメントに必要なきめ細かい情報は抽出できない．また POS データは実務レベルで十分活用されているとは言えない状況にある．

本稿では，多くの実店舗が保有する集計 POS データに対して，機械学習や高次元データのモデリングなどの新しい統計分析による高度情報処理によって，多くの店舗で活用できるマーケティングの市場反応モデルを視野において，その活用の可能性を展望する．具体的には，(i)自然言語処理分野で提案されたトピックモデルの大規模集計 POS データへの適用による次元圧縮と市場細分化，(ii)大規模高次元スパースデータの市場反応モデルについて詳解して展望する．

2. 大規模集計 POS データのトピックモデルによる次元圧縮と市場細分化

2.1 トピックモデルのマーケティングへの展開

自然言語処理分野で潜在的話題—トピック—を抽出するために開発されたトピックモデル(Blei et al., 2003, 2012; Blei and McAuliffe, 2007)は，モデルの汎用性が高く拡張しやすい特徴をもつため，マーケティングにおいても広く使われるようになってきた．まず，元来の目的であるテキスト情報を直接活用するものとしては，ソーシャルメディアから収集したテキストデータから抽出したトピック(話題)をビジネスやマーケティングの問題に活用した研究がある．Si et al.（2013)は Twitter のテキストデータに対して，トピックモデルの一種の潜在ディリクレ配分(LDA: Latent Dirichlet Allocation)モデルを拡張したディリクレ過程混合 LDA モデルによりトピックを抽出し，それらの動きが株価予測に有効であることを示した．Wang et al.(2016)は，Amazon など商品に対するコメントのテキスト解析により，従来のトピックモデルのような全体的話題ではなく，スクリーン，バッテリーなど商品の特徴をターゲットにした話題を抽出するモデルを提案し，抽出された商品の潜在的特徴が売上に有効な情報をもつことを示した．また Morimoto and Kawasaki（2016)では，時系列テキスト分析であるダイナミックトピックモデル(Blei and Lafferty, 2006)を用いて，ファイナンス市場におけるボラティリティを予測する研究が行われている．

これらに対し，トピックモデルを購買や売上など数量データに適用した研究もある．ID-POS データと呼ばれる顧客ごとの非集計データの日々の記録をテキスト解析での一つの文章に対応させ，購買に関するトピックを抽出してマーケティングに活用する研究として，例えば，Ishigakiet al.(2011)がある．Christidis(2010)の研究では，ネット通販での顧客の購買履歴データをトピックモデルで分析し，潜在的マーケットバスケットの発見や個別顧客に対し商品を推薦できるリコメンデーションシステムが可能であることを示した．Iwata et al.（2009)は，提案したトピック・トラッキング・モデルを映画やアニメの購買データに適用し，顧客の趣味・嗜好を潜在トピックにより追跡できることを示した．さらに，Iwata and Sawada（2013)では，ID-POS データに付随する価格情報も考慮した LDA モデル分析を行っている．以上の研究は集計 POS データおよび非集計 ID-POS データを利用するものの，購買行動の背後にある潜在的トピックの情報により顧客を分類することに留まっている．したがって企業が駆使するマーケティング変数による最適化を可能とする制御モデルとは必ずしもなっていない．これに対し，Ishigaki et al.(2017)は，上記を拡張して，顧客の異質性をモデルに取りこみ，潜在購買トピックを割り出すと同時に顧客ごと/商品ごとの価格反応，プロモーション反応を個別に推定するモデルを展開している．

2.2 集計 POS 売上データのトピック分解による市場細分化

集計 POS データは，日々の SKU 単位の商品の売上とその価格およびその商品に対してプロモーションを行ったか否かに関する情報の記録であり，非集計の ID-POS データやレシートデータと異なり顧客の個別の購買状況が見えない．同じ商品の購買であっても購入目的や購買状況の違い，すなわち購買文脈によって，商品の評価やプロモーションへの反応などマーケティング戦略の効果は異なると考えるのが自然である．

Terui and Li (2017)では，集計データに埋没した購買文脈の異質性を潜在変数とし，トピックモデルにより分析に取り入れる．具体的には，集計データをその日に購買された商品の売上情報を用いて複数のショッピングバスケット(ここではトピック)に振り分けて市場細分化を行い，バスケットごとの市場反応を測定して実店舗のきめ細かいマーケティング戦略のための情報提供を可能とするものである．データは高次元でスパースな性質を持っており，これらのための統計モデルを提案している．

自然言語処理の手法として通常使われるトピックモデル(Blei et al., 2003; Blei, 2012)では，単語 v が文書 d に現れる確率は，潜在的トピック k の存在のもとでは，次式の有限混合モデルで表現できると仮定する．

$$p(v|d) = \sum_{k=1}^{K} p(v|k)\, p(k|d) = \sum_{k=1}^{K} \phi_{v|k}\theta_{k|d}. \tag{2.1}$$

Terui and Li (2017)では，店舗の集計売上数量データを同時購買情報のもとで各トピックに分類する．商品 j はテキスト分析における単語 v，日 t は文章 d に対応させる．t 日における商品 j の売上数を Y_{jt} とするとき，$\boldsymbol{Y}_t = (Y_{1t}, Y_{2t}, \ldots, Y_{n_t})'$ は同じ日のすべての商品 j の売上数量ベクトルである．この場合，t 時点で商品 j が購入される確率は $p(j|t) = \sum_{k=1}^{K} p(j|k)p(k|t) = \sum_{k=1}^{K} \phi_{j,k}\theta_{k,t}$ である．さらに，$\phi_{j,k}\theta_{k,t}$ に基づいて商品 j の売上数 Y_{jt} を各トピックごとに分解し，トピック k に入った売上数を $Y_{jt}^{(k)} = Y_{jt} \times E(\phi_{j,k}\theta_{k,t})$ で表す．そのとき t 日における商品 j の売上数は $Y_{jt} = \sum_{k=1}^{K} Y_{jt}^{(k)}$ と表現できる．トピックへの割当て確率に比例する $(\phi_{j,k}\theta_{k,t})$ は均一ではなく，少数のトピックに集中する場合が多いことから，トピックモデルにより次元圧縮が可能である．

2.3 実証分析(1)

実証分析として，Terui and Li (2017)での結果を紹介する．まずデータは，一店舗の 2002 年 5 月 6 日-2003 年 5 月 7 日間の日次集計 POS データを利用する．このデータセットは 363 日(1 年中 3 日休み)で，7912 個の商品種類の総計 3,720,419 回の購買記録が含まれている．POS データは各商品の毎日の売上数量と売上金額，また三種類のプロモーション(実施の有無のバイナリーデータ)の情報が含まれ，購買されていない商品のマーケティング変量の情報は含まれていない．

トピックの数 $k = 10$ と設定し，トピック分布と単語分布はハイパーパラメータ $\boldsymbol{\alpha}$ と $\boldsymbol{\beta}$ によるディレクレー分布を事前分布とし，崩壊型ギッブス・サンプリングで推定を行う．

$$\theta_d \sim Dir(\alpha)\ (d = 1, \ldots, M); \phi_k \sim Dir(\beta)\ (k = 1, \ldots, K)$$

ここでは Griffiths and Steyvers (2004)の研究に従って，$\alpha = 50/k$，$\beta = 0.1$ と設定した(トピックの構成については紙面の都合上 Terui and Li (2017)を参照)．

後段で売上を価格や各種プロモーションなどマーケティング変数の関数として規定する市場反応モデルを定義することを念頭にして，いまターゲットとする商品として特定ブランドの牛乳(JANcode:4902705065161)を取り上げる．その売上数量 Y は 344 日で総計 21,482 個の売上

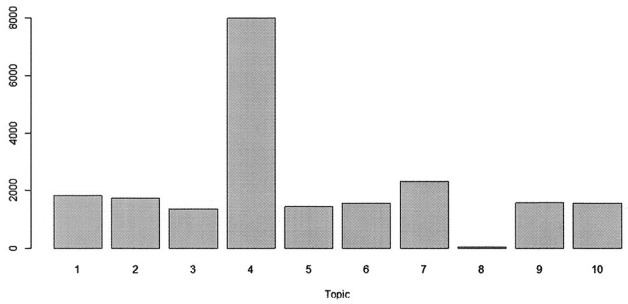

図 1．平均トピック分布．

数である．

図 1 は，7912 個の商品の集計 POS データにトピックモデルを適用したときの平均トピック分布

$$\hat{\phi}_{j,k}\hat{\theta}_{k|\bullet} = \frac{1}{N}\sum_{t=1}^{N}\hat{\phi}_{j,k}\hat{\theta}_{k|t}, k = 1, \ldots, K, \tag{2.2}$$

を表している．ここで，$N = 344$，$K = 10$ であり，一番頻度が高いトピック $k = 4$，と一番頻度が低いトピック $k = 8$ 以外は，ほぼ均等に配分されていることがわかる．図 2 は，各トピックの確率 $\hat{\phi}_{j,k}\hat{\theta}_{k|t}$ でターゲット商品である特定ブランドの牛乳の集計売上数を按分し，トピック分解した $Y_{jt}^{(k)}$ の時系列データである．

他と比べて極端に数量の少ないノイズ的意味をもつトピック 8 を除くと，各トピックは 2 − 4 か月ごとの季節性のトピックを表している．

3. 大規模集計 POS データの市場反応モデル

3.1 高次元スパースデータの統計モデル

高次元スパースデータに関する統計モデルはこれまで多くの研究がある．まず回帰の枠組みでは，Meinshausen and Yu (2009)が Tibshirani (1996)による LASSO 回帰の高次元スパースデータでの変数選択問題の理論的研究を行い，説明変数間に相関が強い場合には問題が生じることを示した．LASSO 回帰は変数間の相関が高い POS データでは有効とは言えない．Chen and Ishwaran (2012)は，バギング (Breiman, 1996)というアルゴリズムを用いる回帰木のランダムフォーレスト (Breiman, 2001)が高次元スパースデータにおいて，チューニングの難しさの問題も指摘しながらも高い予測精度を持つことを示した．ただし，変数間の構造の推定ができない限界も有している．

つぎに次元圧縮手法として，主成分分析，因子分析，正準相関分析を用いる方法がある．まず主成分分析による研究として，Zou et al. (2006)によるスパース主成分モデル，Tipping and Bishop (1999)による確率的主成分分析をスパースデータに拡張する Zeng et al. (2017)によるスパース確率的主成分がある．因子モデルを用いる Lopes and West (2004)は，ベイズ因子分析モデルが計算時間的に通常の因子モデルより優位性を持つため，高次元データに適していると主張した．また，West (2003)は因子分析による説明変数空間の次元圧縮によるスパースベイズ因子回帰モデルを提案し，高い予測精度を持つことを示したが，高次元空間での変数間の構造を推定する議論とはなっていない．

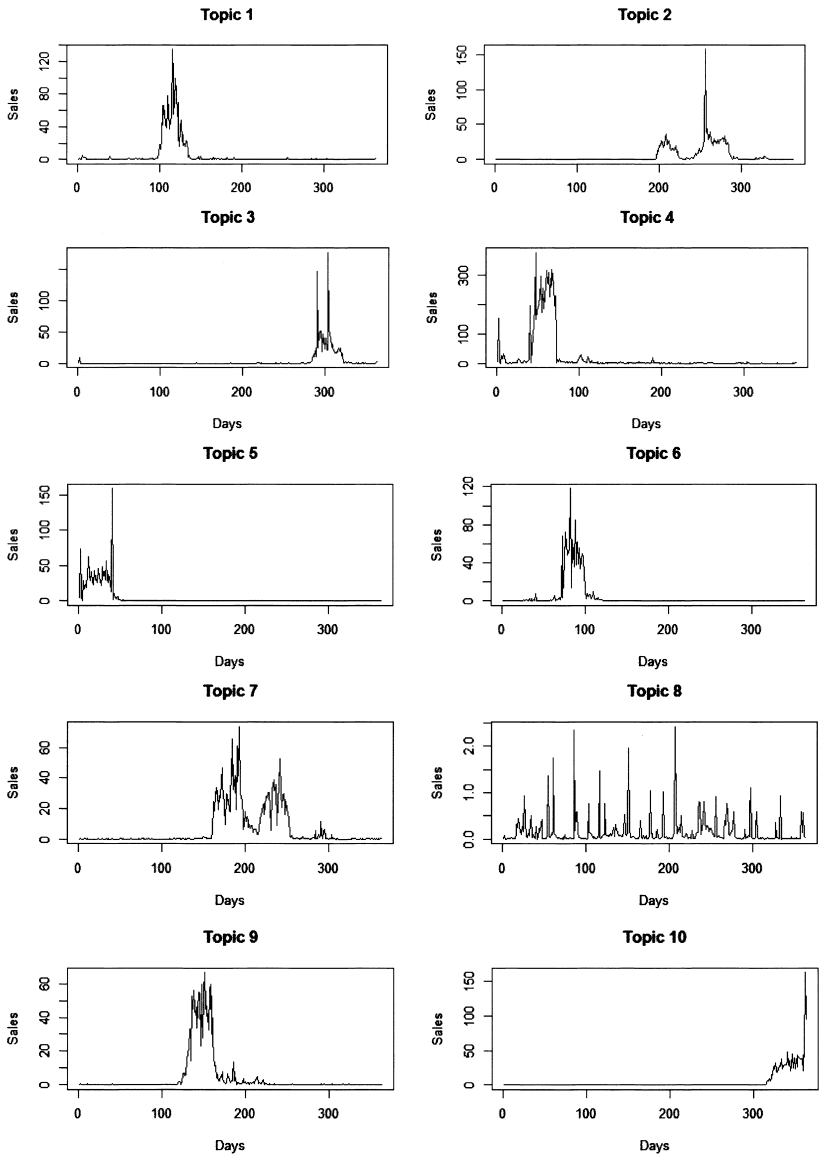

図 2. 売上のトピック分解.

多変数間の関係をモデル化する正準相関分析を用いる最近の研究として，Klami et al.（2013）によるベイズ正準相関分析があり，高次元データ解析での有効性を示した．また Brynjarsdottir and Berliner（2014）は，2 つの高次元時空間大規模データに対し，正準相関分析を基にして変数をそれぞれ圧縮させ，低次元変数間の関係を階層回帰モデルとして定義し，密度の濃い低次元空間で構造を利用して予測モデルを構築している．そのほか，Çiğdem（2009）は正準相関分析をマーケティング問題に応用と評価はしているものの，これらの研究は高次元空間での構造推定は行っていない．マーケティングにおいては，解釈不能な低次元での推定にもとづく予測

だけが目的とはなり得ず，日々のマーケティング意思決定のために，ある特定の商品の売上と価格の弾力性など元の高次元空間での変数間の関係の構造推定が必要となる．

3.2 高次元スパースデータに対する市場反応測定モデル

2 節で行われたショッピングバスケット(トピック)を前提にして，トピックごとの市場反応関数を測定する問題を考える．まずトピック k における商品 j の売り上げ $Y_{jt}^{(k)}$ を説明する市場反応関数を下記で定義する．

$$Y_{jt}^{(k)} = \alpha_{0j}^{(k)} + \boldsymbol{\alpha}'^{(k)}_{j} \boldsymbol{X}_{jt} + \sum_{m \neq j} \beta_m^{(k)} Y_{mt}^{(k)} + \sum_{m \neq j} \boldsymbol{\gamma}'^{(k)}_{m} \boldsymbol{X}_{mt}^{(k)} + \varepsilon_{jt}^{(k)}, \quad t = 1, \ldots, N \tag{3.1}$$

ここで $\boldsymbol{X}_{jt}$ は商品 j のマーケティング変数ベクトル，$Y_{mt}^{(k)}$ はトピック k に入っている目的商品以外($m \neq j$)の商品の売上数，$\boldsymbol{X}_{mt}^{(k)}$ はその商品のマーケティング変数である．ターゲットとする商品の予測は各トピックの予測値の合計 $\hat{Y}_{jt} = \hat{Y}_{jt}^{(1)} + \hat{Y}_{jt}^{(2)} + \cdots + \hat{Y}_{jt}^{(K)}$ で計算できる．

つぎに Terui and Li (2017)で提案した高次元スパースデータの市場反応構造推定のための回帰モデルを紹介する．(3.1)式の回帰モデルの説明変数を改めて $\boldsymbol{X}^{(k)}$ と定義し，同じトピックに入る商品の売上を目的変数 $\boldsymbol{Y}^{(k)}$ とし，各トピックにおける高次元の $\boldsymbol{Y}^{(k)}$ と $\boldsymbol{X}^{(k)}$ の回帰問題について考える．読みやすくするため，ここからはトピックの表記(k)を省略する．

まず P_y 次元の $\boldsymbol{Y}_t$ と P_x 次元の $\boldsymbol{X}_t$ の多変量回帰式を下記で定義する．

$$\boldsymbol{Y}_t = \boldsymbol{F}\boldsymbol{X}_t + \boldsymbol{e}_t, \tag{3.2}$$

ここで，係数行列 $\boldsymbol{F}$ は変数がデータの数を多く超えている高次元であり，いわゆる NP 問題のため直接に推定はできない．そこで，まず $\boldsymbol{Y}_t$ と $\boldsymbol{X}_t$ の周辺分布は，因子構造

$$\boldsymbol{Y}_t = \boldsymbol{U}\boldsymbol{a}_t + \boldsymbol{\eta}_{yt}; \boldsymbol{X}_t = \boldsymbol{V}\boldsymbol{b}_t + \boldsymbol{\eta}_{xt}, \quad t = 1, \ldots, N, \tag{3.3}$$

をもつと仮定する．ここで $\boldsymbol{a}_t$ は $f_y(< P_y)$ 次元のベクトル，$\boldsymbol{U}$ は $P_y \times f_y$ の行列，$\boldsymbol{b}_t$ は $f_x(< P_x)$ 次元のベクトル，$\boldsymbol{V}$ は $P_x \times f_x$ の行列，$\boldsymbol{\eta}_y \sim (0, \Sigma_y)$，$\boldsymbol{\eta}_x \sim (0, \Sigma_x)$ と仮定する．

さらに Brynjarsdottir and Berliner (2014)に従い，各因子モデルによる低次元の因子空間において関係があり，階層回帰モデル

$$\boldsymbol{a} = \boldsymbol{H}\boldsymbol{b} + \boldsymbol{\varepsilon} \tag{3.4}$$

で規定されると仮定する．ここで $\boldsymbol{a}$ および $\boldsymbol{b}$ は(3.3)の因子スコアベクトルをデータ $t = 1, \ldots, N$ について纏めた $f_y \times N$ および $f_x \times N$ の因子スコア行列，$\boldsymbol{H}$ は $f_y \times f_x$ の回帰係数行列，$\boldsymbol{\varepsilon}$ は $f_y \times N$ の誤差行列で各列は独立に $N(0, \sigma^2 I)$ に従うと仮定する．(3.2)–(3.4)を纏めて**階層因子回帰モデル(Hierarchical Factor Regression)**と呼ぶ．

いま(3.2)でのデータを纏めた $\boldsymbol{Y}$ と $\boldsymbol{X}$ の同時分布は，共通パラメータ $\boldsymbol{H}$ を条件付として独立であると仮定する．このとき，(3.2)，(3.3)のモデルの尤度関数は

$$p(\boldsymbol{Y}, \boldsymbol{X} | \boldsymbol{U}, \boldsymbol{a}, \boldsymbol{V}, \boldsymbol{b}) = p(\boldsymbol{Y} | \boldsymbol{U}, \boldsymbol{a}, \boldsymbol{H})\, p(\boldsymbol{X} | \boldsymbol{V}, \boldsymbol{b}, \boldsymbol{H})\, p(\boldsymbol{H} | \boldsymbol{a}, \boldsymbol{b}). \tag{3.5}$$

と書かれる．

3.3 高次元空間への構造の回復

Brynjarsdottir and Berliner (2014)は，この圧縮次元空間を「結晶化空間」と名付けたが，高次元空間の構造 $\boldsymbol{F}$ を推定する提案は行われていない．$\boldsymbol{X}$ が与えられたとき，構造方程式(3.1)は誤差の条件付期待値 $E_{|x}[\boldsymbol{e}] = 0$ の仮定の下での $\boldsymbol{Y}$ の条件付期待値は

$$E_{|x}[\boldsymbol{Y}] = \boldsymbol{F}\boldsymbol{X} + E_{|x}[\boldsymbol{e}] = \boldsymbol{F}\boldsymbol{X} \tag{3.6}$$

であり，さらに $\boldsymbol{X}$ の確率測度に関して期待値をとって無条件期待値の関係として下記が得られる．

$$E_x\left\{E_{|x}[\boldsymbol{Y}]\right\} = \boldsymbol{F}E_x[\boldsymbol{X}] \quad \text{i.e. } \boldsymbol{\mu}_y = \boldsymbol{F}\boldsymbol{\mu}_x \tag{3.7}$$

他方，因子モデルによる圧縮次元空間上の変量の期待値をとれば，$E_x\{E_{|x}[\boldsymbol{Y}]\} = E[\boldsymbol{Y}] = \boldsymbol{U}\boldsymbol{a}_t$，$E_x[\boldsymbol{X}] = \boldsymbol{V}\boldsymbol{b}_t$ であり，(3.7)式から

$$\boldsymbol{U}\boldsymbol{a}_t = \boldsymbol{F}\boldsymbol{V}\boldsymbol{b}_t \tag{3.8}$$

が得られ，データを纏めた $\boldsymbol{a}$ と $\boldsymbol{b}$ の表記により $\boldsymbol{F}$ は次式により求められる．

$$\boldsymbol{F} = \boldsymbol{U}\boldsymbol{a}\boldsymbol{b}'\boldsymbol{V}'\left(\boldsymbol{V}\boldsymbol{b}\boldsymbol{b}'\boldsymbol{V}'\right)^{-1} \tag{3.9}$$

次にモデルの全パラメータの同時事後確率が下記のように表現される．

$$\begin{aligned} & p\left(\boldsymbol{U},\boldsymbol{a},\boldsymbol{V},\boldsymbol{b},\boldsymbol{H},\sigma^2,\boldsymbol{F},\Sigma_y,\Sigma_x,\Lambda_h|\boldsymbol{Y},\boldsymbol{X}\right) \\ & \quad \propto p\left(\boldsymbol{U},\boldsymbol{a}|\boldsymbol{H},\boldsymbol{Y}\right)p\left(\boldsymbol{V},\boldsymbol{b}|\boldsymbol{H},\boldsymbol{X}\right)p\left(\boldsymbol{H}|\boldsymbol{a},\boldsymbol{b},\sigma^2\right)p\left(\boldsymbol{F}|\boldsymbol{U},\boldsymbol{a},\boldsymbol{V},\boldsymbol{b}\right) \\ & \qquad \times p\left(\sigma^2|\boldsymbol{a},\boldsymbol{b},\boldsymbol{H}\right)p\left(\boldsymbol{U}|\boldsymbol{a},\Sigma_y\right)p\left(\Sigma_y\right)p\left(\boldsymbol{V}|\boldsymbol{b},\Sigma_x\right)p\left(\Sigma_x\right) \\ & \qquad \times p\left(\boldsymbol{a}|\Lambda_a\right)p\left(\Lambda_a\right)p\left(\boldsymbol{b}|\Lambda_b\right)p\left(\Lambda_b\right)p\left(\boldsymbol{H}|\Lambda_h\right)p\left(\Lambda_h\right)p\left(\sigma^2\right) \end{aligned} \tag{3.10}$$

これは因子モデルと多変量回帰モデルの事後分布の組み合わせであり，それぞれ共役な事前分布の利用のもとで，ギッブスサンプリングにより効率的に分布評価が可能である．さらに高次元空間での市場反応係数行列 $\boldsymbol{F}$ の復元は，(3.9)式の関係を利用して MCMC 過程での副産物として，周辺事後分布 $p(\boldsymbol{F}|\boldsymbol{Y},\boldsymbol{X})$ が評価できる．MCMC アルゴリズムについては Terui and Li (2017)に詳細が記載されている．

3.4 実証分析(2)

2 節と同じデータを用いた実証分析の結果を紹介する．

表 1 には，式(3.9)の関係を利用して回復した構造 $\boldsymbol{F}^{(k)}$ の事後分布を評価し，95%HPD 領域の意味で有意な回帰係数推定値(事後平均)と説明変数名の一部が記載されている(紙面の都合上トピック 1–5 までを記載した．全トピックについては Terui and Li (2017)を参照)．回帰係数は弾力性を意味し，商品 ID，カテゴリーの名前が続いて表記されている．これにより，下記のようなマネジメントに有用な知見が得られる．

(i) 各トピックにおける同時購買の説明変数は有意となるケースが多い結果になった．トピックモデルにより，同時購買された商品で回帰モデルを構築したためと考えられる．これらの情報を基に同時購買しやすい商品の組み合わせの発見が可能になり，適切なプロモーションにより特定商品の売上の向上が期待できる．

(ii)「価格」変数はトピック 6 および 10 で多く有意になっている．目的変数は他の商品の価格の影響を多く受け，商品自身の価格の影響は著しくない．

(iii) トピック 5，6，7，9，と 10 のプロモーションの影響は強い．

(iv) トピック 9 では，多くの食料品が有意な説明変数として抽出されている．このトピックにおける目的変数の牛乳は飲み物ではなく，調理品として買われていると推測できる．

(v) 売上数が一番多いトピック 4 では，水やジュースカテゴリーの商品は目的変数の牛乳と同期している反面，クッキーカテゴリーの商品とは逆の関係性を持っている．これは 6

表 1. トピックごとの説明変数の抽出.

Topic	Sales	Coef	Price	Coef	Prom1	Coef	Prom2	Coef	Prom3	Coef
1	[4891]miso	2.318	[3064]rice cracker	0.079	[3064]rice cracker	-0.019	[2092]tomato juice	-0.089	-	-
	[4929]coffee drink	2.184	[3523]cup noodle	0.030	[1639]cup noodle	0.010	[235]yogurt	-0.043	-	-
	[5025]rice cracker	2.141	[7211]barbecue sauce	-0.001	-	-	-	-	-	-
	[5577]suger	2.093	-	-	-	-	-	-	-	-
	[3249]energy drink	1.946	-	-	-	-	-	-	-	-
	[7034]coffee drink	1.942	-	-	-	-	-	-	-	-
	[1927]family ice	-1.885	-	-	-	-	-	-	-	-
	[3287]fruit juice	1.764	-	-	-	-	-	-	-	-
	[494]sauce	1.709	-	-	-	-	-	-	-	-
	[7384]premium ice	1.649	-	-	-	-	-	-	-	-

Topic	Sales	Coef	Price	Coef	Prom1	Coef	Prom2	Coef	Prom3	Coef
2	[5790]biscuit and cookie	3.000	[6012]cheese	-0.014	[2245]snack	0.154	[5550]cup noodle	-0.260	-	-
	[4919]family ice (cream)	1.948	-	-	[2553]chocolat	-0.090	[3128]cup noodle	-0.123	-	-
	[3846]cup noodle	1.764	-	-	-	-	[2112]instant soup	0.108	-	-
	[1345]chocolat	1.764	-	-	-	-	-	-	-	-
	[2727]soy	1.680	-	-	-	-	-	-	-	-
	[1826]dressing	1.674	-	-	-	-	-	-	-	-
	[89]coffee drink	1.532	-	-	-	-	-	-	-	-
	[3587]personal ice	1.462	-	-	-	-	-	-	-	-
	[6860]instant coffee	-1.460	-	-	-	-	-	-	-	-
	[1413]biscuit and cookie	-1.389	-	-	-	-	-	-	-	-

Topic	Sales	Coef	Price	Coef	Prom1	Coef	Prom2	Coef	Prom3	Coef
3	[1287]cooked curry	-2.088	-	-	[3225]salad oil	0.486	[2031]dressing	-0.178	-	-
	[4825]lactic acid bacteria b	1.913	-	-	-	-	[2030]dressing	0.145	-	-
	[1138]mayonnaise	-1.607	-	-	-	-	-	-	-	-
	[494]sauce	1.516	-	-	-	-	-	-	-	-
	[6402]chocolat	1.379	-	-	-	-	-	-	-	-
	[3116]candy	1.352	-	-	-	-	-	-	-	-
	[469]dessert	-1.316	-	-	-	-	-	-	-	-
	[166]personal ice (cream)	1.294	-	-	-	-	-	-	-	-
	[3559]cheese	1.272	-	-	-	-	-	-	-	-
	[883]carbonated flavor	1.128	-	-	-	-	-	-	-	-

表 1.（つづき）

Topic	Sales	Coef	Price	Coef	Prom1	Coef	Prom2	Coef	Prom3	Coef
4	[258]water	3.963	[3182]cup noodle	0.165	-	-	[869]carbonated flavor	-0.049	-	-
	[5304]biscuit and cookie	-3.253	-	-	-	-	-	-	-	-
	[4309]instant curry	3.042	-	-	-	-	-	-	-	-
	[3731]carbonated flavor	3.008	-	-	-	-	-	-	-	-
	[4792]family ice (cream)	2.941	-	-	-	-	-	-	-	-
	[5184]chocolat	-2.626	-	-	-	-	-	-	-	-
	[2334]snack	-2.54	-	-	-	-	-	-	-	-
	[4094]fruit juice	2.431	-	-	-	-	-	-	-	-
	[6122]chocolat	2.328	-	-	-	-	-	-	-	-
	[6832]cola	2.314	-	-	-	-	-	-	-	-

Topic	Sales	Coef	Price	Coef	Prom1	Coef	Prom2	Coef	Prom3	Coef
5	[7897]instant noodle	2.174	[5046]dessert	0.036	[4710]biscuit and cookie	0.451	[5255]chocolat	-0.328	[2227]rice cracker	-0.108
	[6815]black tea drink	-1.998	[4983]health food	-0.015	[3419]snack	0.276	[5916]family ice	-0.047	-	-
	[3020]dressing	1.781	[2297]snack	0.000	166]Principle of Chinese foo	-0.187	[3015]mayonnaise	-0.020	-	-
	[4297]cooked curry	1.584	-	-	-	-	[6451]100% fruit juice	-0.006	-	-
	[6735]miso	1.538	-	-	-	-	-	-	-	-
	[6862]	1.528	-	-	-	-	-	-	-	-
	[7717]energy drink	1.528	-	-	-	-	-	-	-	-
	[2486]candy	-1.443	-	-	-	-	-	-	-	-
	[505]candy	-1.413	-	-	-	-	-	-	-	-
	[6211]candy	1.410	-	-	-	-	-	-	-	-

月から 8 月までの夏の季節性の影響であると推測できる．

紹介したモデル分析では，実務や Hanssens et al.（2001）などで行われてきたカテゴリーに限定した小変数回帰では得られず，全商品のデータを使うことによってはじめて得られる意外な商品の組み合わせの発見がある．1970 年代の人工知能ブームにおいて注目された「紙おむつとビール」の同時購買のデータマイニングによる発見に類似した知見である．他方，これに加えて変数間の関係性の構造が弾力性の形で評価され，店舗マネジメントに役立つマーケティング戦略への情報を提供できる利点を持っている．

4. おわりに

マーケティングにおいては，主体（消費者）の異質性がいち早く重視され，個別対応のための「個」のモデリングが展開され実用化してきた．本研究詳解では，「買い物状況（トピック）」の異質性をトピックモデルで表現し，買い物状況ごとに異なる購買要因の存在とその関係の構造を仮定し，集計 POS データを用いてきめ細かいマーケティングを実現するモデルを提案する研究を紹介した．全商品データを分析に取り入れることで，目的変数を説明する意外な商品の発見のみならずその量的関係が弾力性の形で測定可能となり，実店舗のきめ細かいマーケティング戦略に有用な情報が提供可能である．

本稿ではトピック数の選択問題は取り上げていないが，無限ディリクレ過程を利用したノンパラメトリックベイズの適用によるトピック数の推定が可能であり，階層因子回帰モデルにおける因子数の推定も同様である．また本研究では対象企業を広範囲とするため，小企業でも保有している集計 POS データの活用を念頭に置いた．対象企業の範囲は狭まるが，レシートデータが扱える状況を想定すれば，1 回の買い物トリップでの同時購買情報を利用して購入者の異質性を反映させたモデル分析が可能である．さらに個人が特定されるメンバーシップ顧客データベースを対象とすれば，さらに消費者異質性と購買状況の異質性の相互作用の分析も可能であろう．これらは今後の研究課題として有望であろう．

謝　辞

JSPS 科研費 Grant Number（A）25245054 の助成を受けた．

参 考 文 献

Blei, D. M. (2012). Introduction to probabilistic topic models, *Communications of the ACM*, **55**, 77–84.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models, *Proceedings of the ICML*, **6**, 113–120.

Blei, D. M. and McAuliffe, J. (2007). Supervised topic models, *Neural Information Processing Systems*, **3**, 993–1022.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation, *Journal of Machine Learning Research*, **3**, 993–1022.

Breiman, L. (1996). Bagging predictors, *Machine Learning*, **24**(2), 123–140.

Breiman, L. (2001). Random forests, *Machine Learning*, **45**(1), 5–32.

Brynjarsdottir, J. and Berliner, L. K. (2014). Dimension-reduced modeling of spatio-temporal process, *Journal of American Statistical Association*, **109**, 1647–1659.

Chen, X. and Ishwaran, H. (2012). Random forests for genomic data analysis, *Genomics*, **99**(6), 323–329.

Christidis, K., Apostolou, D. and Mentzas, G. (2010). Exploring customer preferences with probabilistic topics models, *In European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2010*.

Çiğdem Şahin, B. (2009). An evaluation and an application of using canonical correlation analysis in marketing research, *International Journal of Economic and Administrative Studies*, **1**(3), 41–68.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics, *Proceedings of the National Academy of Sciences*, **101**, 5228–5235.

Hanssens, D. M., Parsons, L. J. and Shultz, R. L. (2001). Market response models, *Econometric and Time Series Analysis*, 2nd ed., Kluwer Academic Press Inc., Boston, MA.

Ishigaki, T., Takenaka, T. and Motomura, Y. (2011). Customer behavior prediction system by large scale data fusion in a retail service, 人工知能学会論文誌, **26**(6), 670–681.

Ishigaki, T., Terui, N., Sato, T. and Allenby, G. (2017). Personalized market response analysis for a wide variety of products from sparse transaction data, *International Journal of Data Science and Analytics*, **5**(4), 233–248.

Iwata, T. and Sawada, H. (2013). Topic model for analyzing purchase data with price information, *Data Mining and Knowledge Discovery*, **26**, 559–573.

Iwata, T., Watanabe, S., Yamada, T. and Ueda, N. (2009). Topic tracking model for analyzing consumer purchase behavior, *International Joint Conference on Artificial Intelligence '09*, 1427–1432.

Klami, A., Virtanen, S. and Kaski, S. (2013). Bayesian canonical correlation analysis, *Journal of Machine Learning Research*, **14**, 965–1003.

Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis, *Statistica Sinica*, **14**(1), 41–67.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data, *The Annalsof Statistics*, **37**(1), 246–270.

Morimoto, T. and Kawasaki, Y. (2016). Forecasting financial market volatility using a dynamic topic model, *Asia-Pacific Financial Markets*, **24**(3), 149–167.

Rossi, P. E., Allenby, G. and McCulloch, R. (1996). The value of purchase history data in target marketing, *Marketing Science*, **15**(4), 321–340.

Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H. and Deng, X. (2013). Exploiting topic based twitter sentiment for stock prediction, *Publisher Association for Computational Linguistics*, **2**, 24–29.

照井伸彦 (2008).『ベイズモデリングによるマーケティング分析』, 東京電機大学出版局, 東京.

Terui, N. and Dahana, W. D. (2006). Price customization using price thresholds estimated from scanner panel data, *Journal of Interactive Marketing*, **20**(3), 58–70.

Terui, N. and Li, Y. (2017). Measuring large scale market responses from aggregated sales regression model for high dimensional sparse data, Discussion Paper of DSSR No.66, Graduate School of Economics and Management, Tohoku University, Sendai.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society*, **58**(1), 267–288.

Tipping, M. and Bishop, C. (1999). Probabilistic principal component analysis, *Journal of the Royal Statistical Society, Series B*, **21**(3), 611–622.

Wang, H., Chen, Z., Fei, G., Liu, B. and Emery, S. (2016). Targeted topic modeling for focused analysis, *Association for Computing Machinery*, **13**(17), 1235–1244.

West, M. (2003). Bayesian factor regression models in the large p, small n paradigm, *Bayesian Statistics*, **7**, 723–732.

Zeng, J., Liu, K., Huang, W. and Liang, J. (2017). Sparse probabilistic principal component analysis model for plant-wide process monitoring, *Korean Journal of Chemical Engineering*, **34**(8),

2135–2146.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis, *Journal of Computational and Graphical Statistics*, **15**(2), 265–286.

High-dimensional Sparse Modeling of Large-scale Aggregated POS Data

Yinxing Li and Nobuhiko Terui

Graduate School of Economics and Management, Tohoku University

Micro-marketing based on consumer heterogeneity using disaggregated ID-POS data has been well studied in the literature, but the implementation and operation of this approach remain limited, particularly in real store management. On the other hand, while aggregated POS data are collected by most retailers, it is widely recognized that these data have never been well utilized.

We discuss the possibility of using aggregated POS data by applying new techniques in high-dimensional sparse data modeling and machine learning. In particular, we propose a procedure comprising two sub-models: the topic model first decomposes the aggregate number of sales to several different market baskets, and then hierarchical factor regression is used to reduce dimensionality and ultimately recover from the reduced dimension to the original space in order to detect the marginal effect among all products in each market basket.

The proposed model, which uses a large amount of product data, not only makes it possible to discover unexpected predictors, but also measures the quantitative relation in the form of elasticity for managerial implications.

Key words: Aggregated POS data, heterogeneity in shopping context, topic model, high-dimensional sparse-data, hierarchical factor regression.