論文・事例研究

トピックモデルによる顧客データの統合的分析

里村　卓也

1. はじめに

オンライン小売業は幅広い種類の顧客データを収集することが可能である．顧客の購買履歴や，デモグラフィックス，サイトへのアクセス・デバイスのような内容に加えて，自社の顧客へのアンケート調査を行うことで価値観や日常生活での行動などの顧客サイコグラフィックス属性についても知ることができる．そして，これらをもとに購買履歴などからは知りえない顧客に関する知識を得ることができる．しかしながら，購買履歴やアンケートなどのデータはそれぞれ個別に分析されることが多い．たとえば顧客購入履歴データは商品のレコメンデーションに活用され，アンケート調査はサイコグラフィックスに基づくセグメンテーションに利用することができる．それぞれの分析結果は対応するマーケティング戦術を実行するうえで有用であるかもしれないが，異なるデータからは異なる顧客像を描くことになる．一方，もしこれらのデータを統合して分析することで統一した顧客インサイトを得ることができるのであれば，マーケティング戦略を構築するうえでも有用であるといえよう．

そこで本研究では顧客購買履歴データと顧客調査データを結びつける手法の提案を行う．提案手法は購買商品と顧客サイコグラフィックス属性を同時に分析することで統合的な顧客インサイトを獲得することを目指すものである．また，モデルのパラメータの意味について人間が直接解釈することも可能であり，ニューラルネットワークで得られるモデルのように中身を人間が解釈できないものとは異なる．

2. モデル

2.1 提案手法とその特徴

本研究で提案する手法は商品と顧客の同時分析を行うために，ジョイント・トピックモデルを適用するものである．提案手法の特徴としては以下の 3 点が挙げられる．

1.「サイコグラフィックス属性」と「購入商品」を結びつける「潜在的特性」を抽出することができる．潜在的特性を解釈することで顧客インサイトの獲得を行える．
2. 商品とサイコグラフィックス属性の潜在的な共起関係から，「購入可能性の高い商品」「発現可能性の高い顧客サイコグラフィックス属性」を予測することができる．したがって潜在力をもとにした商品やイベントのレコメンデーションを行うことが可能となる．
3.「購入商品」の分布から，「顧客サイコグラフィックス属性」の分布を予測することができる．そのため，購入商品のみの情報しかない顧客についてもサイコグラフィックス属性を個人別に予測することが可能となる．

2.2 ジョイント・トピックモデル

トピックモデルは潜在的意味解析の分野で利用されている手法である．その中でも確率的潜在変数モデルである Latent Dirichlet Allocation (LDA) モデル [1] が近年は中心的に利用されている．LDA モデルでは，各文書内でのトピックの共起関係やトピック内での単語間の潜在的な共起性を抽出することができる [2]．LDA モデルは大規模な文書を解析するために導入されたが，顧客購買データの分析においても応用されている．文書を顧客，単語を商品に置き換えることで，商品の購入確率を予測したり [3]，顧客と商品についての潜在的情報の抽出を行うことが可能となる．また LDA モデルは協調フィルタリングに比べると，顧客数や商品数が増加した場合に必要なメモリサイズの増加が少なくて済むため，ビッグ・データへの適用で有利である [3]．

複数データセットを利用したトピックモデルについても，潜在的意味解析の分野を中心に提案がなされている．このような研究としては Bilingual LSA [4] や Polylingual Topic Model [5]，アノテーションのモデ

さとむら　たくや
慶應義塾大学商学部
satomura@fbc.keio.ac.jp
受付 17.7.25　採択 17.9.30

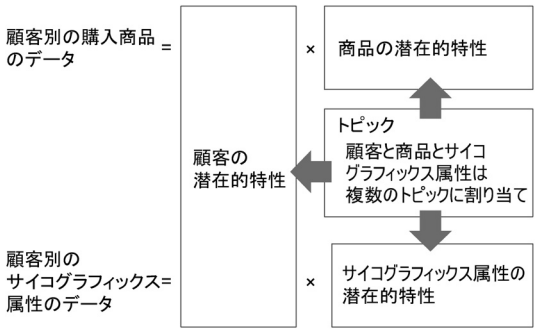

図 1 ジョイント・トピックモデルによる購入商品とサイコグラフィックス属性の同時分析

ル化 [6]，固有表現の抽出 [7]，ファッションのレコメンデーション [8] などを挙げることができる．石垣ら [9] によるカテゴリマイニングでは潜在クラスモデルを利用して顧客アンケートデータと購入商品を同時に分析する手法を提案している．

本研究では LDA を複数データセットに適用できるように拡張したジョイント・トピックモデルを利用して複数のデータセットの融合を行う．本研究で用いるジョイント・トピックモデルは Mimno et al. [5] や Iwata et al. [8] のモデルと同じ構造であるが，モデルの適用対象が異なる．

本研究では顧客は商品とサイコグラフィックス属性で共通な潜在的特性であるトピックをもっているとする．顧客の潜在的特性，商品の潜在的特性，サイコグラフィックス属性の潜在的特性はトピックにより決まるものと考える（図 1）．また顧客は同時に複数のトピックを確率的にもつものとする．

潜在的意味解析の分野では，トピックモデルにおけるトピックは「潜在的意味のカテゴリー」を表しているものと考えることができる [2]．顧客購買データにトピックモデルを適用した Jacobs et al. [3] では，トピックは商品に対するある種の選好を表していると考えて「モチベーション」と呼んでいる．本研究では，トピックは商品への選好とサイコグラフィックス属性の組み合わせを同時に表している「潜在的ライフスタイルのカテゴリー」と考えることができる．マーケティグにおいて，ライフスタイルは時間とお金をどのように使うかについての個人の選択を反映した消費のパターンを規定するものである [10]．本研究におけるトピックは，顧客がどの商品を購入しどのように生活するのかを規定するものであるため，潜在的ライフスタイルのカテゴリーを表していると考えることができる．

2.3 ライフスタイル研究における本研究の位置づけ

マーケティングにおけるライフスタイル研究には活動 (activities)，関心 (interests)，意見 (opinions) の変数を用いる AIO [11] や，価値をベースにした Rokeach Value Survey (RVS) [12]，SRI International の Values and Lifestyles (VALS) [13]，List of Values (LOV) スケール [14] などがある．VALS を改訂した SRI International の VALS2^{TM} では 35 個の心理属性と 4 個のデモグラフィックス属性を用いて消費者を分類する [10]．ただし，これらの研究では指標は一般的なライフスタイル尺度を作成するためのものであり，個別の商品についての利用状況は消費者の分類には利用されない．また，ライフスタイルは，ある期間中は個人内では一定であると考えられている．

一方，最近のライフスタイル研究においては，一般的なライフスタイル尺度の作成よりも，目的に応じたライフスタイル尺度の作成を行う傾向がある [15]．たとえば，飽戸 [16] では日本におけるファッション・ライフスタイルを五つに分類している．

サイコグラフィックスによるライフスタイル分類は研究によりセグメント数やその内容が異なるため [17]，研究間での比較が難しいといえる．本研究ではサイコグラフィックス属性は提供されたデータを利用するため，分析から得られた結果の解釈に重点を置き，従来の研究との結果の比較については議論をしないことにする．

本研究で提案するアプローチは購入商品とサイコグラフィックス属性を同時に分析するものであり，商品の利用状況も分類に反映される．そのために，分析対象者のサイコグラフィックス属性の値が同じでも，同時に分析対象となる商品カテゴリーが異なればライフスタイルの分類も異なることになる．

従来の研究では，まず一般的なライフスタイルを測定しその後に商品の利用状況について分析を行うか，分析の目的に応じて測定変数を設定する必要があった．一般的なライフスタイル尺度では，必ずしも特定商品の購入について説明力が高いとは限らない．また，目的に応じたライフスタイル尺度の作成のためには，目的に応じてサイコグラフィックス属性を変更する必要がある．一方，本研究で提案する手法では，同じサイコグラフィックス属性を用いても，同時に分析する商品カテゴリーが異なればそれに応じて異なったライフスタイル分類の結果を得ることもありうるため，商品カテゴリー別にサイコグラフィックス属性を変更しなくても商品カテゴリー別など目的に応じた分析を行う

ことが可能となる．

さらに，本研究で用いる LDA モデルでは，トピックは個人内で購入アイテムごと質問項目ごとに異なることを許すモデルである．現代の消費者は一人十色と呼ばれように，一人の消費者が多くのライフスタイルをもっており，場面に応じて使い分けていると考えられる．従来のライフスタイル研究ではこのような消費者の多面性を十分に考慮することはできなかったが，本研究で用いるトピックモデルは消費者内での多面性に対応するモデル構造となっている．

2.4 モデルの定式化

顧客 $d(=1,\ldots,D)$ がトピック $k(=1,\ldots,K)$ に所属する確率（トピック k の構成比率）を θ_{dk} とする．θ_{dk} の事前分布をパラメータ α のディリクレ分布とする．

$$\theta_d \sim \mathrm{Dirichlet}(\alpha) \tag{1}$$

次に商品購入に関して，商品購入の背後には潜在的特性であるトピックがあると考える．トピックが異なれば，商品の購入のされやすさが異なる．顧客内トピックは商品購入ごとに変化するとする．トピック $k(=1,\ldots,K)$ におけるアイテム $v(=1,\ldots,V)$ の出現確率を ϕ_{kv} とする．ϕ_{kv} の事前分布をパラメータ β のディリクレ分布とする．顧客 d の $n(=1,\ldots,N_d)$ 番目の購買アイテムにおけるトピックを z_{dn} とする．z_{dn} は離散値をとる潜在変数であり，パラメータ θ_d の多項分布に従うとする．また顧客 d の n 番目の購買アイテムを w_{dn} とする．w_{dn} はパラメータ $\phi_{z_{dn}}$ の多項分布に従うとする．

$$\begin{aligned} \phi_k &\sim \mathrm{Dirichlet}(\beta) \\ z_{dn} &\sim \mathrm{Multi}(\theta_d) \\ w_{dn} &\sim \mathrm{Multi}(\phi_{z_{dn}}) \end{aligned} \tag{2}$$

顧客のサイコグラフィックス属性は 0-1 の値をとるものとする．サイコグラフィックス属性の値の背後には潜在的特性であるトピックがあると考える．トピックが異なれば，発現するサイコグラフィックス属性も異なる．顧客内トピックはサイコグラフィックス属性ごとに変化するとする．トピック $k(=1,\ldots,K)$ におけるサイコグラフィックス属性 $s(=1,\ldots,S)$ の出現確率を ψ_{ks} とする．ψ_{ks} の事前分布をパラメータ γ のディリクレ分布とする．顧客 d の $m(=1,\ldots,M_d)$ 番目のサイコグラフィックス属性におけるトピックを y_{dm} とする．y_{dm} は離散値をとる潜在変数であり，パラメータ θ_d の多項分布に従うとする．また顧客 d の m 番目のサイコグラフィックス属性を x_{dm} とする．x_{dm} はパラメータ $\psi_{y_{dm}}$ の多項分布に従うとする．

$$\begin{aligned} \psi_k &\sim \mathrm{Dirichlet}(\gamma) \\ y_{dm} &\sim \mathrm{Multi}(\theta_d) \\ x_{dm} &\sim \mathrm{Multi}(\psi_{y_{dm}}) \end{aligned} \tag{3}$$

式 (1)，式 (2)，式 (3) より顧客 d についての対数尤度 $l(\theta_d,\phi,\psi|w_d,x_d)$ は次のようになる．

$$\begin{aligned} &l(\theta_d,\phi,\psi|w_d,x_d) \\ &=\sum_{n=1}^{N_d}\log\left\{\sum_{k=1}^{K}p(z_{dn}=k|\theta_d)p(w_{dn}|\phi_k)\right\} \\ &\quad+\sum_{n=1}^{M_d}\log\left\{\sum_{k=1}^{K}p(y_{dm}=k|\theta_d)p(y_{dm}|\psi_k)\right\} \end{aligned} \tag{4}$$

また顧客全体の対数尤度は次のようになる．

$$l(\theta,\phi,\psi|w,x)=\sum_{d=1}^{D}l(\theta_d,\phi,\psi|w_d,x_d) \tag{5}$$

3. データとモデルの推定

3.1 データについて

本研究では平成 28 年度データ解析コンペティションで貸与されたファッション EC サイトの購買およびアンケートデータを用いて分析を行う．分析対象者はアンケート回答がある顧客のうち，データ期間中に購買のあった 3,112 名である．分析対象アイテムは 218 カテゴリー（ファッション EC サイトの商品カテゴリー小レベル）である．分析対象者の平均購入アイテム数は 7.8 個であり，購入個数 4 個以下の顧客が分析対象者の 51.7%を占める．

分析対象サイコグラフィックス属性はアンケートの中から 94 項目抽出され，以下の八つに分類される．1) 2015 年参加イベント；2) 意識；3) 人生重視価値；4) 幸福感；6) 購入時期；7) ファッション課題；8) ファッション観．これらのサイコグラフィックス属性は全て 0-1 変数に変換して使用した．

3.2 モデルの推定とトピック数の決定

モデルの推定は RStan2.16 [18] の自動変分ベイズ法 [19] を用いて変分ベイズ推定を行った．事前分布のパラメータについては $\alpha=1/K,\beta=0.1,\gamma=0.1$ に設定した．トピックモデルでは事前にトピック数を与える必要がある．そこでトピック数を 2 から 10 の間で間隔 1 で変化させて対数周辺尤度を比較した（図 2）．対数周辺尤度が最も高くなるのはトピック数が 6 の場合であったので，トピック数は 6 とした．

トピック数を 6 にした場合，各トピックの比率はトピック 1 が 44.2%，トピック 2 が 17.5%，トピック

表 1　各トピックと総計での性別年齢階層別構成比率（%）

	トピック 1	トピック 2	トピック 3	トピック 4	トピック 5	トピック 6	総計
男性 10 代	2.1	0.6	0.8	0.8	1.0	0.9	1.3
男性 20 代前半	5.4	1.6	2.1	2.1	2.6	2.5	3.5
男性 20 代後半	5.0	1.8	2.4	2.2	2.7	2.6	3.4
男性 30 代前半	8.1	3.2	3.8	4.1	4.9	4.6	5.8
男性 30 代後半	9.0	3.7	4.3	4.4	5.4	5.2	6.4
男性 40 代以上	14.4	6.7	7.7	7.9	9.7	9.4	10.7
女性 10 代	2.6	1.6	1.9	2.1	2.2	2.2	2.2
女性 20 代前半	5.0	5.2	6.8	7.2	6.4	6.3	5.7
女性 20 代後半	9.4	14.9	16.0	15.1	13.4	13.6	12.5
女性 30 代前半	11.8	17.7	16.2	16.1	15.0	15.0	14.4
女性 30 代後半	10.5	17.7	15.2	14.8	14.3	14.5	13.4
女性 40 代以上	16.7	25.4	22.9	23.2	22.4	23.2	20.6

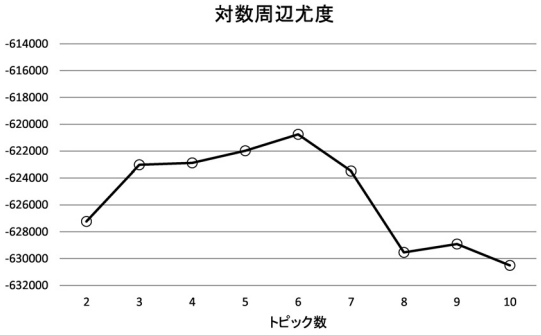

図 2　対数周辺尤度の比較

3 が 12.8%，トピック 4 が 13.8%，トピック 5 が 5.8%，トピック 6 が 5.9%となった．性別年齢についての変数はモデル構造に含まれないため，トピックごとに事後的に集計を行った．各トピックおよびデータ総計での性別年齢階層別の比率は表 1 のとおりである．総計でみると，男性は 31.2%，女性が 68.8%を占める．

4.　推定結果の活用

4.1　複数データの同時利用による顧客インサイトの獲得

「購入商品」と「サイコグラフィックス属性」を結びつけるトピック（潜在的ライフスタイルのカテゴリー）を解釈することで顧客の理解を得ることができる．このとき，顧客の性別年齢階層別の構成比率も解釈で利用する．

各トピックの解釈には商品とサイコグラフィックス属性の出現確率である ϕ と ψ を用いる．ただし，顧客全体との乖離から各トピックの購入商品やサイコグラフィックスの特徴を把握するために，ここではリフト値（各トピックでの出現率 ÷ 全体での出現率）を用いる．このリフト値と，総計と比べた相対的な各トピックでの性別年齢階層別構成比率をもとに解釈を行う[1]．

1：トピック 1 ではインテリアや雑貨・ホビー・スポーツ用品，食器・キッチン用品，帽子などの購入率が高い．好みのファッションがあり，ファッション動向にも敏感である．また，人生では有名・出世・競争・勝利・所属・大望などの価値を重視している．男性の比率が相対的に高く，また年齢が高い層の比率も相対的に高い．

2：トピック 2 ではスーツ・ネクタイやワンピース，マタニティ・ベビー用品の購入率が高い．年間を通して多くのイベントに参加する．ファッションは保守的であり，購入動機も控えめである．20 代後半以上の女性の構成比率が相対的に高い．

3：トピック 3 ではインテリアやマタニティ・ベビー用品，トップス，コスメ・香水などの購入率が高い．服にあまりお金をかけず，またバーゲンやセールスを利用することも多い．自分の現状や将来に対して悲観的である．20 代後半の女性の構成比率が相対的に高い．

4：トピック 4 ではコスメ・香水，インテリアに加えてアクセサリーの購入率が高い．服の原産国や素材を気にするが，ファッションの流行には関心が低い．女性の構成比率が相対的に高い．

5：トピック 5 では生活用品のインテリアやコスメ・香水の購入率が高い．ファッションの流行には関心が低く，人生価値観も安定を重視している．女性の構成比率が相対的に少し高い．

6：トピック 6 では日用品に関連したインテリアやコスメ・香水の購入率が高い．ファッションに関して自信をもっておらず，人生価値観も安定を重視

[1] リフト値の表はサイズが大きいため論文には含めず，論文ではそれらをもとに解釈した結果についてまとめた．

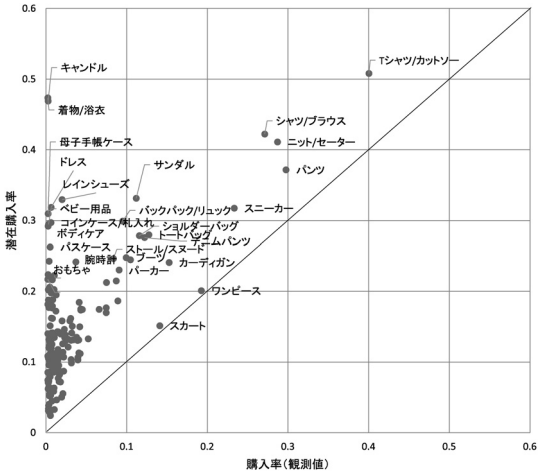

図 3　商品の潜在的購買の可能性

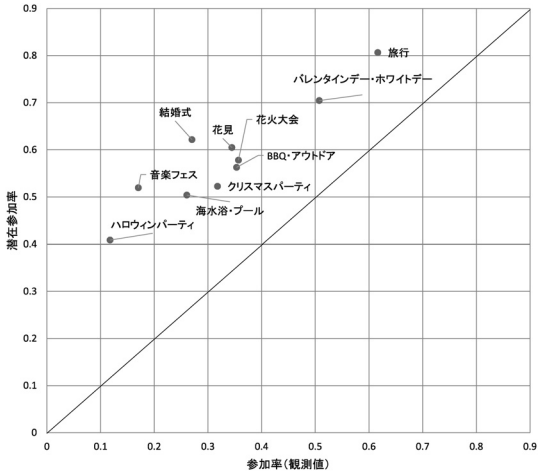

図 4　イベントの潜在的参加の可能性

している．女性 40 代以上の構成比率が相対的に高い．

4.2　購入可能性の高い商品と発現可能性の高い顧客サイコグラフィックス属性の抽出

トピックをもとにした商品やサイコグラフィックス属性の潜在的な共起関係から「購入可能性の高い商品」と「発現可能性の高い顧客サイコグラフィックス属性」を抽出する．

購入可能性の高い商品を抽出するために，まずすべての顧客 d について商品 v の購入確率の予測値を以下の式から求める．

$$\Pr(w_d = v) = \sum_{k=1}^{K} p(v|k)p(k|d) = \sum_{k=1}^{K} \phi_{kv}\theta_{dk} \quad (6)$$

次に，商品 v の購入者の中から $\Pr(w_d = v)$ の 75%点を求め，これを $v^{\#}$ とする．そして，すべての顧客の中で $\Pr(w_d = v) \geq v^{\#}$ となる顧客 d を商品 v の潜在的購買可能性の高い顧客とした．

発現可能性の高い顧客サイコグラフィックス属性を抽出するために，まずすべての顧客 d についてサイコグラフィックス属性 s の出現確率の予測値を以下の式から求める．

$$\Pr(x_d = s) = \sum_{k=1}^{K} p(s|k)p(k|d) = \sum_{k=1}^{K} \psi_{ks}\theta_{dk} \quad (7)$$

次に，サイコグラフィックス属性 s の発現者について $\Pr(x_d = s)$ の 50%点を求め，これを $s^{\#}$ とする．そして，すべての顧客の中で $\Pr(x_d = s) \geq s^{\#}$ となる顧客 d をサイコグラフィックス属性 s の潜在的発現可能性の高い顧客とした．

上記の方法にもとづき，図 3 は各商品の潜在的購買可能性の高い顧客の比率（潜在購入率）を計算したものであり，図 4 はサイコグラフィックス属性のうち各イベントへの潜在的参加可能性の高い顧客の比率（潜在参加率）を計算したものである．商品の既存購入者とイベントの既存参加者についても購入や参加の潜在的可能性が高い顧客に含めているため，潜在購入率や潜在参加率は 45 度対角線よりも上に付置される．なお商品購入については購入率（観測値）が 0.002 以下の商品は図 3 からは除いてある．

ここで計算された潜在購入率は，あくまでも既存購入者や既存参加者の ϕ や ψ をもとに任意に設定した閾値から判定したものであり，閾値の設定によって計算結果は変わってくることになるが，ここでは一つの活用例として上に示した閾値をもとに話を進める．

図 3 の商品の潜在的購買可能性を見ると，ワンピースとスカートなどは 45 度対角線上に近く，これ以上の購買の可能性が高くないことがわかる．一方，T シャツ／カットソーはさらに 10%程度の潜在的購入の余地があることがわかる．またシャツ／ブラウス，ニット／セーターとパンツは同程度の購入率であるが，シャツ／ブラウスとニット／セーターのほうがパンツよりも潜在的購入率が高い．また現在は購入率が 10%程度の商品の多くは潜在的には倍以上の購入率となる可能性がある．

次に図 4 のイベントへの潜在的参加の可能性を見ると，多くのイベントは現在の参加率より 20%以上の参加率の増加する可能性がある．その中でも特にハロウィンパーティや音楽フェスは参加の潜在余剰が大きい．結婚式も潜在余剰が大きいが，これはある年には結婚式に参加しなくても別の年には参加する可能性が

あるため，と解釈することができる．顧客が新しいイベントに参加する場合には，それに伴って新しい商品やサービスが必要とされるため，新しい商品を購入することが期待される．参加可能性の高いイベントを顧客に紹介することで，顧客が新しい生活活動を行うことを促進し，その中で新たに商品を購入してもらうという，ライフスタイル・マーケティングの戦略立案がこの分析結果を利用することで可能となる．

4.3 顧客サイコグラフィックス属性の予測

ここまでは，複数データがある顧客について考えてきたが，インターネット小売業ではすべての顧客に対してアンケート調査を行うことは少ないであろう．そこで「購買データ」のみの顧客についても，ほかの顧客のアンケート調査のデータを利用することでサイコグラフィックス属性を予測することを考える．

そのために潜在クラスモデルを用いたデータ融合 [20] をジョイント・トピックモデルに適用する．購買データのみのサンプルは，サイコグラフィックス属性データが欠損していると考える．どの顧客にアンケートをとるかは完全にランダムであるとすると，サイコグラフィックス属性データの欠損は完全にランダム (missing completely at random: MCAR) であると考えることができる．このようなサンプリングの形態はサブ・サンプリング [21] と呼ばれる．MCAR の場合にはサイコグラフィックス属性データを欠損値として尤度を構成しても，推定されたパラメータにバイアスはない．分析ではアンケート回答者を A グループ（D_a=2,000 名）と B グループ（D_b=1,112 名）にランダムに割り振る．推定に利用するデータは A グループは購買データとサイコグラフィックス属性データの完全データ，B グループは購買データのみでサイコグラフィックス属性データは欠損とする．

このときのジョイント・トピックモデルの対数尤度は以下のとおりである．

$$\begin{aligned} & l(\theta^*, \phi, \psi | w, x) \\ = & \sum_{d=1}^{D_a+D_b} \sum_{n=1}^{N_d} \log \left\{ \sum_{k=1}^{K} p(z_{dn}=k|\theta_d^*) p(w_{dn}|\phi_k) \right\} \\ & + \sum_{d=1}^{D_a} \sum_{n=1}^{M_d} \log \left\{ \sum_{k=1}^{K} p(y_{dm}=k|\theta_d^*) p(y_{dm}|\psi_k) \right\} \end{aligned} \tag{8}$$

ただし

$$d = \begin{cases} 1, \ldots, D_a; & A \text{ グループ} \\ D_a+1, \ldots, D_a+D_b; & B \text{ グループ} \end{cases}$$

となる．θ_d^* の推定のために，A グループは購買データとサイコグラフィックス属性データを用い，B グループは購買データのみを用いている．

サイコグラフィックス属性の出現確率は，以下の方法で求める．

$$\begin{aligned} & \Pr(x_d = s | \theta_d^*, \psi) \\ = & \sum_{k=1}^{K} p(x_d = s|\psi_k) p(y_d = k|\theta_d^*) \\ = & \sum_{k=1}^{K} \psi_{ks} \theta_{dk}^* \end{aligned} \tag{9}$$

サイコグラフィックス属性についてグループ A は in-sample data でありグループ B は out-of-sample data となっている．そこでサイコグラフィックス属性の予測値についてもデータ内であるグループ A とデータ外であるグループ B について求める．この二つのグループでの予測値の精度を比較する．

まず，サイコグラフィックス属性は 0-1 の変数であるが，式 (9) で求まるのは予測確率であるため，予測確率の閾値を設定する．閾値の設定は 4.2 節と同様の方法で，完全データ（グループ A）のサイコグラフィックス属性の値と予測確率をもとに決定する．グループ A のサイコグラフィックス属性 s の出現者について $\Pr(x_d = s)$ の p%点を求め，これを $s_p^{\#}$ とする．これらをもとに $\Pr(x_d = s) \geq s_p^{\#}$ となる顧客 d についてはサイコグラフィックス属性 s が 1 とし，それ以外は 0 とした．この予測結果をもとに観測値が 1 で予測値も 1 である割合を TPF (True Positive Fraction)，観測値が 0 で予測値が 1 である割合を FPF (False Positive Fraction) とする．TPF は真陽性率（感度），FPF は偽陽性率（1-特異度）である．TPF-FPF が最大になるポイントは Youden's Index であり，最も効率よく予測ができる閾値である．

予測のために完全データであるグループ A の Youden's Index を求めると，そのポイントはグループ A の顧客の $p = 31.7\%$ 点となる．このようにして求めた各サイコグラフィックス属性値の閾値を用いて，グループ B のサイコグラフィックス属性の値を予測する．結果は表 2 のとおりである．

表 2 観測値と予測値の例

観測	予測	観測数	予測数	TPF, FPF
1	1	43,278	30,104	0.696(TPF)
0	1	61,250	31,918	0.521(FPF)

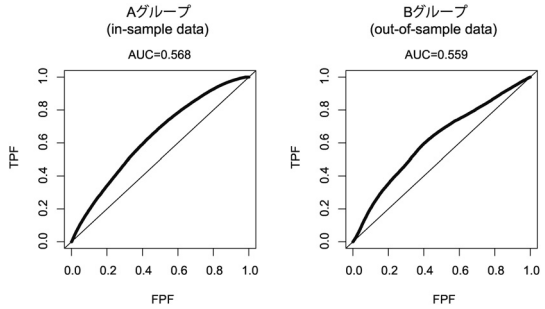

図 5　顧客サイコグラフィックス属性の予測

また，閾値を変化させながら横軸に FPF，縦軸に TPF をプロットすると ROC 曲線が得られる．ROC 曲線下面積 (AUC: area under the curve) は 0.5 から 1 の間のとり，AUC が 1 に近いほど，予測能力が高いと評価できる．図 5 には in-sample data であるグループ A と out-of-sample data であるグループ B の ROC 曲線が描かれている．グループ A の AUC は 0.568 であり，グループ B の AUC は 0.559 となっており，out-of-sample data による予測でも in-sample-data に近い予測力があることがわかる．

5.　おわりに

本研究ではジョイント・トピックモデルを利用して顧客購買履歴データと顧客調査データを結びつける手法の提案を行った．提案手法は購買商品と顧客ライフスタイルを同時に分析することで統合的な顧客インサイトを獲得することを目指すものである．推定結果から得られるモデルの構造は人間が解釈可能なものとなっている．

提案手法の特徴としては以下の 3 点が挙げられる．一つ目は購入商品とサイコグラフィックス属性を結びつける潜在的特性を抽出し顧客インサイトの獲得を行うことができる点である．二つ目に潜在的な共起関係から，購入可能性の高い商品と発現可能性の高い顧客サイコグラフィックス属性を予測し，この潜在力をもとにした商品や生活活動のレコメンデーションを行うことができる点である．三つ目は購入商品の分布から，顧客サイコグラフィックス属性の分布を予測できるため，アンケート調査を実施していない顧客についてもサイコグラフィックス属性の個人別推定を行うことが可能となる点である．

実証分析ではファッション EC サイトの顧客の購買およびアンケートデータを利用し，提案手法の有用性についての検証を行った．提案手法は複数データを統合的に利用することで統一した顧客インサイトを得られるだけでなく，顧客への新しい生活行動の提案を行うことや，購買データから顧客のサイコグラフィックス属性についての予測を行えることが示された．

次に本研究の課題と今後の方向性について述べたい．本研究で用いたデータはファッション EC サイトの購買履歴データおよび顧客調査データであるため，潜在的意味解析における 1 文章中の単語出現数と比べると，1 顧客当たりの購入商品数は少ないといえよう．そのため，購入商品のみから顧客サイコグラフィックス属性を予測しようとした場合には，そのような限られた情報を用いて個人別のトピックの分布を推定することになるため，個人別のトピックの分布が限られた購買データの影響を受けやすい点が問題として挙げられる．そこで顧客特性である性別年齢をトピックの事前分布に関連づけることができれば，顧客の異質性を考慮したトピックの事前分布を用いることができる．たとえば Jacobs et al. [3] ではトピックの事前分布のパラメータに顧客レベルの情報を取り込むことで予測精度を高めている．このような事前分布への顧客情報の利用は今後の課題である．

複数のデータを統合的に分析することで顧客に関して統一した理解をもとにマーケテイング戦略を立案することが可能となる．特にオンライン小売業のように大規模なデータの活用を考えている企業においては，本研究で提案した手法を活用することで，多様な顧客について個々の顧客の行動とその背景を理解して対応することが可能となろう．

謝辞　本研究の分析では経営科学系研究部会連合協議会主催「平成 28 年度データ解析コンペティション」で提供されたデータを使用した．関係者各位に感謝の意を表します．

参考文献

[1] D. M. Blei, A. Y. Ng and M. I. Jordan, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research*, **3**, pp. 993–1022, 2003.
[2] 佐藤一誠，『トピックモデルによる統計的潜在意味解析』，コロナ社，2015.
[3] B. J. D. Jacobs, B. Donkers and D. Fok, “Model-based purchase predictions for large assortments,” *Marketing Science*, **35**, pp. 389–404, 2016.
[4] Y. C. Tam, I. Lane and T. Schultz, “Bilingual LSA-based adaptation for statistical machine translation,” *Machine Translation*, **21**, pp. 187–207, 2007.
[5] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith and A. McCallum, “Polylingual topic models,” In *Proceedings of the 2009 Conference on Empirical*

Methods in Natural Language Processing, **2**, pp. 880–889, 2009.
[6] D. M. Blei and M. I. Jordan, "Modeling annotated data," In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 127–134, 2003.
[7] D. Newman, C. Chemudugunta, P. Smyth and M. Steyvers, "Statistical entity-topic models," In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 680–686, 2006.
[8] T. Iwata, S. Watanabe and H. Sawada, "Fashion coordinates recommender system using photographs from fashion magazines," In *Proceedings of International Joint Conference on Artificial Intelligence*, pp. 2262–2267, 2011.
[9] 石垣司，竹中毅，木村陽一，"日常購買行動に関する大規模データの融合による顧客行動予測システム―実サービス支援のためのカテゴリマイニング技術―，" 人工知能学会論文誌，**26**, pp. 670–681, 2011.
[10] M. R. Solomon, *Consumer Behavior: Buying, Having and Being*, 10th edition, Pearson, 2013.
[11] W. D. Wells and D. J. Tigert, "Activities, interests, and opinions," *Journal of Advertising Research*, **11**(4), pp. 27–35, 1971.
[12] M. Rokeach, *The Nature of Human Values*, The Free Press, 1973.
[13] A. Mitchell, *The Nine American Lifestyles*, Macmillan Publishing, 1983.
[14] L. R. Kahle (ed.), *Social Value and Social Change: Adaption to Life in America*, Praeger, 1983.
[15] 清水聰，『新しい消費者行動』，千倉書房，1999.
[16] 飽戸弘，『売れ筋の法則―ライフスタイル戦略の再構築―』，筑摩書房，1999.
[17] J. P. Peter and J. C. Olson, *Consumer Behavior & Marketing Strategy*, 7th edition, McGraw-Hill, 2007.
[18] Stan Development Team, "RStan: the R interface to Stan," R package version 2.16.2, http://mc-stan.org, 2017（閲覧日 2017 年 7 月 21 日）.
[19] A. Kucukelbir, R. Ranganath, A. Gelman and D. M. Blei, "Automatic variational inference in Stan," arXiv: 1506.03431, 2015.
[20] W. Kamakura and M. Wedel, "Statistical data fusion for cross-tabulation," *Journal of Marketing Research*, **34**, pp. 485–498, 1997.
[21] W. Kamakura and M. Wedel, "Factor analysis and missing data," *Journal of Marketing Research*, **37**, pp. 490–498, 2000.

