情報処理学会論文誌 Vol.60 No.4 1151–1161 (Apr. 2019)

# 販売履歴データに基づく中古ファッションアイテムの販売価格予測モデルに関する一考察

仁ノ平 将人[1,a)] 三川 健太[2,b)] 後藤 正幸[3,c)]

受付日 2018年7月6日, 採録日 2019年1月15日

**概要**：近年の情報技術の発展により，EC（電子商取引）サイトを通じた商品の購買が普及している．本研究で対象とするファッション EC サイト A では，ユーザから中古ファッションアイテムを買取り，値付けを行い再販売を行っている．この EC サイトでは，売れ残りを防ぐため，出品アイテムに対し一定のアルゴリズムで自動的に値下げをする仕組みを採用している．このビジネスモデルにおいて，各アイテムに対し，ある価格で出品された各アイテムが最終的にいくらで販売されるかを予測することは，値付けシステムの構築や経営戦略を考える際に重要である．本研究では，EC サイト A における出品アイテムの販売価格予測モデルの構築のために，潜在クラスを用いた混合回帰モデルを用いた分析を行う．すなわち，アイテムの特徴，季節ごとの値下がり率（オフ率）の傾向をもとに潜在クラスモデルを用いてクラスタリングを行った後に，データの各潜在クラスへの所属確率を用いて潜在クラスごとに回帰式を構築する推定モデルを構築する．さらに，得られた潜在クラスの情報を活用することで，オフ率が定義できない新規出品データに対しても予測が可能となることを示す．本手法が EC サイト A の購買データにおいて販売価格を予測するモデルとして有効なモデルであることを示すとともに，得られたモデルを解釈することで説明変数が持つ販売価格の影響度の定量化を行った．

**キーワード**：EC サイト，中古ファッションアイテム，回帰モデル，潜在クラスモデル，機械学習

# Selling Prices Prediction Model Construction of Second-hand Fashion Items Based on Sales History Data

Masato Ninohira[1,a)] Kenta Mikawa[2,b)] Masayuki Goto[3,c)]

Received: July 6, 2018, Accepted: January 15, 2019

***Abstract:*** Recently, it has become popular for consumers to purchase product items through EC sites. Especially as fashion items, the purchasing actions by consumers for them through EC sites have been rapidly increased. This study focuses on a fashion EC site which operates the resale business of second-hand clothes. They assess the appropriate exhibit prices of second-hand fashion items and resell them on this EC site. A characteristic of this EC site is that if an item is not bought for a certain period, the price force to be discounted automatically. In this EC site, it is important to predict the selling price of each item in condition given information and an exhibit price. When we can predict accurate selling price and clear the effects of factors on selling price, it should help a various marketing strategies. In this paper, we propose a new regression model to predict selling price using linear regression models depending on clusters which are constructed by the relation between the features of items and seasonal off-rate. In order to show the effectiveness of our proposal, simulation experiments with a real data are demonstrated and we discuss the analysis of the results for some insightful marketing policies.

***Keywords:*** EC site, second-hand fashion items, regression model, latent class model, machine learning

[1] 早稲田大学大学院創造理工学研究科
Graduate School of Creative Science and Engineering, Waseda University, Shinjuku, Tokyo 169–8555, Japan
[2] 湘南工科大学工学部
Department of Information Science, Shonan Institute of Technology, Fujisawa, Kanagawa 251–8511, Japan
[3] 早稲田大学創造理工学部
School of Creative Science and Engineering, Waseda University, Shinjuku, Tokyo 169–8555, Japan
[a)] nino0114hira@fuji.waseda.jp
[b)] mikawa@info.shonan-it.ac.jp
[c)] masagoto@waseda.jp

## 1. はじめに

近年の情報技術の発展により，EC（電子商取引）サイトを利用した商品の購買が普及している．これらの EC サイトでは，多種多様なアイテムが取り扱われており，その規模も日々増加している．他方，本研究で対象とするファッションアイテムに関しては，現実店舗に比べ在庫が充実している点や価格の安さといった利点がありながらも，実際の商品を確認することができないため，多くの消費者は期待した商品と実物の品質が異なるというリスクを危惧するとされていた [1], [2]．しかし近年では，ファッションアイテムを取り扱う EC サイトではアイテムに関する詳細情報の提供をはじめ，無料返品サービスやユーザの求めるコーディネートの提案など，売り上げや顧客満足度向上のための様々な施策を行っており，その売上は増大傾向である．

一般に，EC サイトではユーザの閲覧履歴や購買履歴，検索履歴といったログデータが取得できるため，これらの多様なデータを活用し，様々な施策に結び付けようとする取り組みが活発である [3]．たとえば，商品推薦 [4] は多くの EC サイトで一般的であり，様々な方法が提案されている [5], [6]．また，Hou らはユーザの購買行動予測のために木構造モデルによる特徴量変換とシンプルな機械学習を用いた予測モデルを提案している [7]．Dias らは，Web サイトの検索パターンから，潜在クラスモデルによってオンラインマーケットセグメンテーションを行う方法を示している [8]．その他，閲覧履歴や購買履歴データを活用した消費者の購買行動に関する分析や顧客行動予測システムに関する研究は多岐にわたって行われている [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24]．

これに対し，本研究では，過去のデータに基づき，中古ファッションアイテムの購買価格を予測するモデルの構築を対象とする．本研究で対象とするファッションアイテムを取り扱う EC サイト A では，ユーザから中古ファッションアイテムを買取り，値付けを行い再販売，出品を行っている．この EC サイトでは，売れ残りを防ぐため，出品アイテムが一定の期間消費者から購入されない場合，あるアルゴリズムで段階的に値下げを行う仕組みを採用している．すなわち，EC サイト側で決定した出品価格でアイテムを出品したとしても，そのとおりの値段で買い手が付くとは限らず，最終的に購入者が現れたときに，販売価格が決定される．このようなビジネスモデルにおいて，出品されたアイテムが最終的にいくらで購入されるかといった販売価格の予測を行うことは，値付けシステムの構築や運営戦略を考える際に大変重要である．近年では，人工知能技術を活用してアイテムの価格設定を行うことで，利益の最大化を図ろうとする動きはあるものの [25]，中古ファッションアイテムを対象とした場合，「アイテムの種類が非常に多く，まったく同じアイテムが出品されるケースが少ない」，「流行や商品の程度にも価格が左右される」などの理由によりその販売価格予測は難しい課題の 1 つである．また，EC サイト A で取り扱われるアイテムは膨大な種類となるため，回帰分析やニューラルネットワーク，ランダムフォレストのようなすでに実績のある機械学習手法 [26], [27] をそのまま援用したとしても，精度の高い予測を行うことは難しいという課題がある．

以上より，本研究では EC サイト A の特徴を活用した高精度な出品アイテムの販売価格予測モデルの構築を目的とする．アイテムが購入されず，一定期間が経過すると自動的に値下がりするという EC サイト A の特徴を考慮するため，出品された時点の価格を出品価格，最終的にユーザに購入される価格を販売価格と定義し，出品価格と他の属性情報から販売価格を予測するモデルを構築する．その際，出品価格から販売価格までの値下がり幅をオフ率（%）と定義し，オフ率の情報をモデル構築に活用することで，より精度の高い予測モデルの構築を目指す．しかし，オフ率は目的変数である販売価格を用いて計算されるため，説明変数として直接予測に用いることはできない．そこで，本研究では，オフ率を用いて傾向の類似したアイテムをクラスタリングし，それぞれのクラスタで予測モデルを構築することでオフ率を活用しつつ，予測段階ではこの情報を用いずに予測ができるモデルの構築を行う．具体的には，混合回帰モデル [28] を本研究で対象とする事例に援用し，入力データの特徴量および季節ごとのオフ率（中古販売品の値下げ率）の背後にある潜在的な構造をもとにしたクラスタリングを行う．その後，データの各クラスタへの所属確率を用いてクラスタごとに回帰式を構築，その混合を行うものとする．これらの情報を活用することでオフ率を計算することができない新規アイテムに対しても予測可能なモデルを構築する．分析モデルを当該 EC サイトの購買データに適用することで販売価格の予測モデルとして有効であること，ならびに構築したモデルのパラメータを解釈することで得られる知見について示す．

## 2. 事前分析

一般に，ファッションアイテムは流行や季節に敏感な商材であることが知られている．EC サイト A の出品アイテムの販売価格を予測するうえで，そのアイテムが値引きされやすいアイテムなのかを把握することは非常に重要である．すなわち，EC サイト A の出品アイテムに対し，季節ごとのオフ率の傾向で出品アイテムのクラスタリングが可能であるならば，その傾向に応じて異なる販売価格の予測モデルを構築することは，その予測を行ううえで有効な手段であると考えられる．

そこで以下では，EC サイト A の実販売データをもとに，各アイテムカテゴリに対し，月ごとのオフ率の傾向を分析

した後に，そのデータに対して $k$-means 法を適用[*1]し，季節ごとのオフ率の傾向で出品アイテムの分類が可能か分析を行う．

いま，$N$ 種類からなるアイテムカテゴリ集合を $\mathcal{I} = \{i_n : 1 \leq n \leq N\}$ とし，アイテムカテゴリ $i_n$ に対し，一年間を $M$ 期に区切ったときの $m$ 期（$1 \leq m \leq M$）における50%以上のオフ率で販売された数量の割合を $q_{nm}$ とする．季節ごとのアイテムのオフ率の傾向を分析するために，各アイテムカテゴリ $i_n$ を，この $q_{nm}$ を要素とする $M$ 次元のベクトル $\boldsymbol{q}_n = (q_{n1}, \ldots, q_{nm}, \ldots, q_{nM})^{\mathrm{T}}$ で表し，これらに $k$-means 法を適用する．ここでは，1 年間を 12 カ月に区切り（$M = 12$）分析することを考える．このとき，得られる各クラスタの中心ベクトルを $\boldsymbol{\nu}_k = (\nu_{k1}, \ldots, \nu_{km}, \ldots, \nu_{kM})^{\mathrm{T}}$（$k = 1, 2, \cdots, K$）と定義すると，$\boldsymbol{\nu}_k$ の傾向をもとに各クラスタへの季節ごとのオフ率の傾向について解釈を与えることができる．事前分析では，複数の $K$ について実験を実施したが，いずれの場合も「オフ率の季節傾向」が類似しているアイテムのクラスタが得られる傾向となった．そのため，ここでは解釈のしやすい $K = 6$ としたときの各クラスタの中心ベクトル $\boldsymbol{\nu}_k$ を**図 1** に，各クラスタに所属するアイテムの季節ごとのオフ率の傾向を解釈した結果を**表 1** に示す．

これにより，クラスタ 1 には年間を通じて低いオフ率を持ったアイテムが，クラスタ 6 には冬に高いオフ率を持ったアイテムが所属するといったように，季節ごとのオフ率の傾向をもとにアイテムを分類できることが明らかになった．したがって，販売価格を予測する際に，あらかじめアイテムに対し上記のような傾向でハードクラスタリングを行い，クラスタ別に回帰式を構築することでより高い精度のモデルが得られることが示唆される．

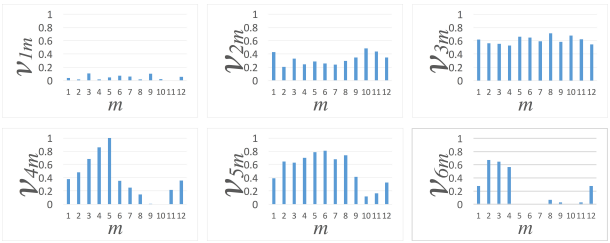


**図 1** $K = 6$ の $k$-means で得られた各クラスタの中心ベクトル $\boldsymbol{\nu}_k$
**Fig. 1** The central vectors of each cluster $\boldsymbol{\nu}_k$ given by applying $k$-means, whose $K$ is 6.

**表 1** $K = 6$ の $k$-means で得られた各クラスタに所属するアイテムのオフ率の傾向
**Table 1** The interpretations of each cluster given by applying $k$-means, whose $K$ is 6.

| $k$ | 傾向 |
|---|---|
| 1 | 年間を通じて低いオフ率 |
| 2 | 年間を通じて一定のオフ率 |
| 3 | 年間を通じて高いオフ率 |
| 4 | 春先に高いオフ率 |
| 5 | 秋に低いオフ率 |
| 6 | 冬に高いオフ率 |

[*1] 本稿では，本章で述べる $k$-means 法を用いたクラスタリング，ならびに確率的潜在クラスモデルを用いたクラスタリングの 2 種類のクラスタリング手法を用いている．これらを明確に区別するため，以降では，前者をハードクラスタリング，後者をソフトクラスタリングと呼ぶものとする．

## 3. 従来手法

### 3.1 重回帰分析

データ数を $L$ 件とし，$y_l$ を $l$ 番目のデータの目的変数，$\boldsymbol{x}_l = (1, x_{l1}, \ldots, x_{ld}, \ldots, x_{lD})^{\mathrm{T}}$ を $l$ 番目のデータの説明変数としたとき，重回帰分析は以下の式 (1) によりデータの関係性を推定し，予測を行う．

$$y_l = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_l + \varepsilon_l \tag{1}$$

$$\varepsilon_l \sim \mathcal{N}(0, \sigma^2) \tag{2}$$

ただし，$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d, \ldots, \beta_D)^{\mathrm{T}}$ は回帰係数と呼ばれ，$\varepsilon_l$ は平均 0，分散 $\sigma^2$ の正規分布に従う誤差項とする．パラメータ $\boldsymbol{\beta}$ は，$L$ 件のデータに対する二乗誤差の最小化を行うことにより求められ，パラメータ $\boldsymbol{\beta}$ に着目することで，各説明変数が与える影響を定量的に分析できるとして，様々な適用事例が知られている．

### 3.2 確率的潜在クラスモデル

潜在クラスモデルは，観測されたデータの背後に観測できない潜在的な変数の存在を仮定したモデルである．潜在的な変数の仮定は，様々な異質なデータが混在している現実的な複雑な問題に対して有効であることが示されている [29], [30]．潜在クラスモデルでは，観測データが各々の潜在クラスに所属する確率を推定することができるため，潜在クラスへの所属確率を用いたクラスタリングが可能となる（本稿ではこれをソフトクラスタリングと呼ぶ）．本研究ではこの特性を活かすことで入力データの特徴量および季節ごとのオフ率の背後にある潜在的な構造をもとにしたソフトクラスタリングを行う．以下本節では，潜在クラスモデルの基本的な手法の 1 つである Aspect Model [31], [32] について述べた後，潜在クラスモデルを回帰問題に適用した混合回帰モデル [28], [33] について説明を与える．

#### 3.2.1 Aspect Model

Aspect Model（以下，AM）は，Hofmann により文書と単語の関係性を表現するモデルとして提案された [31], [32]．いま，$K$ 個の潜在クラスからなる潜在クラス集合を $\mathcal{Z} = \{z_k : 1 \leq k \leq K\}$ とすると，単語 $w_i$（$1 \leq i \leq I$）と文書 $d_j$（$1 \leq j \leq J$）の同時確率 $P(w_i, d_j)$ は式 (3) で表される．

$$P(w_i, d_j) = \sum_{k=1}^{K} P(z_k) P(w_i | z_k) P(d_j | z_k) \tag{3}$$

AMは，前述の単語と文書の共起関係をその他の類似した要素に置き換えることにより様々な問題に適用が可能となり，協調フィルタリングによるレコメンデーション [29] をはじめとする様々な適用例が示されている [10], [30], [34]．また，AMは観測できない変数である潜在クラスを仮定しているため，陽にパラメータを求めることができない．このため，学習データに対する尤度関数を最大化するようEMアルゴリズム [35] を用いてパラメータ推定が行われる．

#### 3.2.2 混合回帰モデル

混合回帰モデル [28] は，目的変数 $y_l$ と説明変数 $\boldsymbol{x}_l = (1, x_{l1}, \ldots, x_{ld}, \ldots, x_{lD})^{\mathrm{T}}$ 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．このモデルは各潜在クラスに対し異なる回帰モデルを仮定しており，それらの混合により表現される．AMと同様に，潜在クラス集合を $\mathcal{Z} = \{z_k : 1 \leq k \leq K\}$ とし，潜在クラス $z_k$ における回帰モデルのパラメータを $\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{dk}, \ldots, \beta_{Dk})^{\mathrm{T}}$，補助変数を $\boldsymbol{v}_l = (v_{l1}, v_{l2}, \cdots, v_{lM})^{\mathrm{T}}$ としたとき，データ $(\boldsymbol{x}_l, y_l)$ に対する混合回帰モデルは式 (4)–(6) で表される．

$$P(y_l|\boldsymbol{x}_l) = \sum_{k=1}^{K} P(z_k|\boldsymbol{v}_l) P_k(y_l|\boldsymbol{x}_l) \tag{4}$$

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

$$f_k(\boldsymbol{x}_l) = \boldsymbol{\beta}_k^T \boldsymbol{x}_l \tag{6}$$

ここで，$P_k(y|\boldsymbol{x})$ は潜在クラス $z_k$ の回帰モデルにおける $y$ の確率密度であり，平均 $f_k(\boldsymbol{x})$，分散 $\sigma_k^2$ の正規分布に従うことを仮定している．また，本モデルのパラメータは，潜在クラスごとの各潜在クラス $z_k$ への所属確率で重み付けされた二乗誤差の最小化を目的関数とし，EMアルゴリズム [35] を用いて推定される．混合回帰モデルを応用した研究として，幾何学的に解析を行った研究 [36] や，マーケティングセグメンテーションへの応用を議論した研究 [37] など，様々な応用研究が報告されている．また，永森ら [38] は，就職ポータルサイトの被エントリ数分析モデルに混合回帰モデルを適用し，実データの分析を通じて，その有効性を示している．

## 4. 分析モデル

### 4.1 モデルの概要

前述のとおり，ECサイトAには様々な特徴を持ったアイテムが出品されている．このため，単一の重回帰モデルを適用しても，高い精度の予測販売価格を得ることは難しい．また，2章で述べた分析より，出品アイテムの季節ごとのオフ率の傾向を分析すると，秋にオフ率が高くなりやすいアイテムや，年間を通じて一定のオフ率が維持されやすいアイテムといったように，季節によるオフ率の傾向の違いにより，アイテムのグルーピングが可能であることが明らかになった．アイテムのカテゴリ，色や素材といった特徴量に加え，季節ごとのオフ率の傾向をもとに潜在クラスモデルを用いたソフトクラスタリングを行う．さらに，得られた潜在クラスごとにそれぞれ販売価格を目的変数とした回帰モデルを構築することにより，各潜在クラス[*2]に所属しているアイテムの特徴に応じた販売価格の予測を行う．

このようなデータの特性を考慮しつつ，高精度な予測を行うため，本研究では潜在クラスモデルを用いた混合回帰モデルを用いる[*3]．ただし，ECサイトAの保持するデータの分析に適したモデル化を行うために，パラメータ推定と新規アイテムの予測に対し，以下の点を考慮する．

まず，潜在クラスを用いたソフトクラスタリングを行う際に，当該データの大きな特徴であるオフ率を特徴量として加えることで考慮する．これにより，オフ率の傾向が類似しているアイテム群を潜在クラスとしてまとめることができ，出品価格と販売価格の関係性が異なるアイテムを異なる潜在クラスへと分離し，別々の回帰モデルを構成したのちに混合するというモデル化が可能となる．

一方，オフ率は混合回帰モデルの目的変数である販売価格と出品価格の比により計算されるため，販売価格が未知であるアイテムに対してはオフ率を計算することができず，この変数を用いた予測を行うことはできない．しかしながら，前述のソフトクラスタリングを用いてモデル化を行っているため，オフ率以外の変数を用いた潜在クラスの条件付き確率を求めることができる．このため，オフ率が未知であったとしても混合回帰モデルによる販売価格の予測値を得ることができる．

また，実応用を考えた場合，モデルの学習に用いられた過去の出品データのみではなく，販売価格が未知の新規出品データに対しても高い精度の予測販売価格を得られる必要がある．そこで，新規出品データに対し，学習で得られた各潜在クラスへの所属確率と，各潜在クラスにおける回帰式の出力を算出し，これらを混合することで，新規出品データの予測販売価格を推定することを考える．

以上から，分析モデルはアイテム属性や季節ラベルを用いた潜在クラスモデルによるソフトクラスタリングとそこで得られた潜在クラスを用いた混合回帰モデルの2段階で構成される．以降ではこれらについて詳細を述べる．

### 4.2 分析モデルの定式化

#### 4.2.1 潜在クラスモデルによるクラスタリング

以下ではファッションアイテムに特有の特性であるア

[*2] すでに述べているとおり，それぞれの潜在クラスはクラスタと解釈できることに注意されたい．個々のデータは，これらのクラスタに確率的に所属するモデルとなっている．

[*3] このようなデータに対し，潜在クラスモデルを用いずにモデル化を行っているものの，適切な説明力を持つモデルは得られず，過学習を起こしやすいことが確認されたため，潜在クラスモデルを用いた混合回帰モデルを構築するものとした．

イテム属性や季節ラベルを用いた潜在クラスモデルによるソフトクラスタリングの定式化について述べる．ここで，季節ラベルとはその商品が出品された季節や月を表すものと定義する．いま，全 $L$ 件の出品履歴データに出現する $M$ 種類の季節ラベルを $\mathcal{S} = \{s_m : 1 \leq m \leq M\}$ とする．さらにアイテムの色や素材といった $j$（$\leq J$）番目の補助情報の要素集合を $\mathcal{A}_j = \{a^j_{v_j} : 1 \leq v_j \leq V_j\}$ とする．たとえば，ある $j$ において $\mathcal{A}_j$ をアイテムの色の集合とすると，$V_j$ は色の種類数であり，$a^j_{v_j}$ は何色かを表す．出品アイテムの $J$ 種類の補助変数を表すために，$J$ 次元のベクトル $\boldsymbol{o} = (o_1, \ldots, o_j, \ldots, o_J)^{\mathrm{T}}$（$o_j \in \mathcal{A}_j$）を定義する．また，$\mathcal{R}^+$ を正の実数集合とし，各アイテムの出品価格を $b \in \mathcal{R}^+$，オフ率を $c \in \mathcal{R}^+$ とする．分析モデルでは，アイテムを季節ごとのオフ率の傾向とその属性によりソフトクラスタリングを行うために，1 つの出品データをこれらの共起 $(i_n, s_m, \boldsymbol{o}, b, c)^{\mathrm{T}}$ ととらえ，それらの間に潜在クラスを仮定する．いま，$K$ 個の潜在クラス集合を $\mathcal{Z} = \{z_k : 1 \leq k \leq K\}$ としたとき，分析モデルの確率モデルは式 (7) で表される．

$$F(i_n, s_m, \boldsymbol{o}, b, c) = \sum_{k=1}^{K} P(z_k)P(i_n|z_k)P(s_m|z_k) \cdot P(b|z_k)P(c|z_k)\prod_{j=1}^{J}\prod_{v_j=1}^{V_J} P(a^j_{v_j}|z_k)^{\delta(o_j, a^j_{v_j})} \tag{7}$$

なお，$\delta(x, y)$ は $x = y$ のとき 1，それ以外は 0 をとる指示関数とする．いま，各潜在クラス $z_k$ のもとでのアイテムの出現確率 $P(i_n|z_k)$，季節ラベルの出現確率 $P(s_m|z_k)$，$j$ 番目の補助情報の出現確率 $P(a^j_{v_j}|z_k)$ にはそれぞれ多項分布，出品価格 $b$ の出現確率密度 $P(b|z_k)$，オフ率 $c$ の出現確率密度 $P(c|z_k)$ には，それぞれ平均 $\mu_k$，$\lambda_k$，分散 $\sigma^2_k$，$\varphi^2_k$ の正規分布を仮定する．すなわち，$\mu_k$ は潜在クラス $z_k$ に所属するデータの出品価格の平均値，$\lambda_k$ はオフ率の平均値を指す．この式 (7) は，アイテムカテゴリ，季節ラベル，補助情報，出品価格，オフ率を特徴量とするデータに対し，その背後には潜在的な構造が存在することを仮定し．そのもとでデータが生起する確率を示している．

#### 4.2.2 潜在クラスを用いた混合回帰モデル

次に，前項で得られた潜在クラスを用いた混合回帰モデルの定式化について述べる．回帰式で用いる出品価格やアイテムカテゴリなどをダミー変数で表した説明変数を $\boldsymbol{x} = (1, x_1, \ldots, x_d, \ldots, x_D)^{\mathrm{T}}$ としたとき，分析モデルでは，潜在クラス $z_k$ ごとに異なる回帰係数 $\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{dk}, \ldots, \beta_{Dk})^{\mathrm{T}}$ を仮定する．さらに，各潜在クラスの回帰式の出力 $\boldsymbol{\beta}^{\mathrm{T}}_k\boldsymbol{x}$ をソフトクラスタリングの際に得られるデータの各潜在クラスへの所属確率 $P(z_k|i_n, s_m, \boldsymbol{o}, b, c)$ で重みを付けて混合することで，販売価格 $y$ を予測するモデルを構成する．

$$y = \sum_{k=1}^{K} P(z_k|i_n, s_m, \boldsymbol{o}, b, c)\boldsymbol{\beta}^{\mathrm{T}}_k\boldsymbol{x} + \varepsilon \tag{8}$$

ただし，$\varepsilon$ は，平均 0，分散 $\sigma^2$ の正規分布に従う誤差項とする．

式 (8) におけるパラメータ $P(z_k|i_n, s_m, \boldsymbol{o}, b, c)$ は式 (7) の導出の際に得られるパラメータを用いることが可能である．また，パラメータ導出方法については次節，および付録 A.1 で述べる．

### 4.3 パラメータの学習

以下では，アイテム属性や季節ラベルを用いた潜在クラスモデルによるソフトクラスタリングと潜在クラスを用いた混合回帰モデルの両者について，それぞれのパラメータの推定方法を述べる．

まず，潜在クラスモデルによるソフトクラスタリングのパラメータ推定について述べる．$l$ 番目の出品データにおけるアイテムカテゴリを $t_l$（$\in \mathcal{I}$），出品日の季節ラベルを $u_l$（$\in \mathcal{S}$），$j$ 番目の補助情報を $w_{lj}$（$\in \mathcal{A}_j$），$\boldsymbol{w}_l = (w_{l1}, \ldots, w_{lj}, \ldots, w_{lJ})^{\mathrm{T}}$ を $l$ 番目の出品データの $J$ 種類の補助情報を表すベクトルとする．さらに，出品価格を $g_l$，オフ率を $h_l$（ともに連続値）とすると，$l$ 番目の出品データはこれらの共起 $(t_l, u_l, \boldsymbol{w}_l, g_l, h_l)^{\mathrm{T}}$ で表現できる．このとき，全 $L$ 件の出品データに対する対数尤度関数 $LL$ は以下の式 (9) で表される．

$$LL = \log\prod_{l=1}^{L}\sum_{k=1}^{K} P(z_k)P(t_l|z_k)P(u_l|z_k) \cdot P(g_l|z_k)P(h_l|z_k)\prod_{j=1}^{J} P(w_{lj}|z_k) \tag{9}$$

潜在クラスによるソフトクラスタリングにおけるパラメータは，対数尤度関数 $LL$ を EM アルゴリズムを用いて最大化をすることにより求める．

次に，潜在クラスを用いた混合回帰モデルのパラメータ推定について述べる．$l$ 番目のデータに対し，回帰式で用いる説明変数を $\boldsymbol{x}_l = (1, x_{l1}, \ldots, x_{ld}, \ldots, x_{lD})^{\mathrm{T}}$，販売価格を $y_l$ とする．このとき，各潜在クラス $z_k$ における回帰式のパラメータ $\boldsymbol{\beta}_k$ は，重み付け重回帰モデル [39] 同様に，各データの各潜在クラス $z_k$ への所属確率で重み付けされた二乗誤差を最小にするよう，以下の式 (10) で推定する．

$$\hat{\boldsymbol{\beta}}_k = \arg\min_{\boldsymbol{\beta}_k}\sum_{l=1}^{L}\alpha_{kl}(y_l - \boldsymbol{\beta}^{\mathrm{T}}_k\boldsymbol{x}_l)^2 \tag{10}$$

ただし，表記の簡素化のため，$\alpha_k = P(z_k|t_l, u_l, \boldsymbol{w}_l, g_l, h_l)$ とする．また，具体的なパラメータ更新式については付録 A.1 を参照されたい．

### 4.4 新規出品データの販売価格の予測

予測モデルの構築においては，販売価格が未知の新規出

品データに対しても高い精度の予測販売価格を得られることが望ましい．そこで，新規出品データに対し，学習により得られた各潜在クラスにおける回帰式の出力を各潜在クラスへの所属確率を用いて混合することで，予測値の算出を行う．いま，新規出品データ数を $L'$ とし，$l'$（$\leq L'$）番目の新規データのアイテムのカテゴリを $t_{l'}$（$\in \mathcal{I}$），季節ラベルを $u_{l'}$（$\in \mathcal{S}$），$j$ 番目の補助情報を $w_{l'j}$（$\in \mathcal{A}_j$），出品価格を $g_{l'}$ とする．このデータに対して，オフ率が未知であることに留意して，学習により得られた各潜在クラスへの所属確率 $P(z_k|t_{l'}, u_{l'}, \boldsymbol{w}_{l'}, g_{l'})$ を以下の式 (11) で求める．

$$P(z_k|t_{l'}, u_{l'}, \boldsymbol{w}_{l'}, g_{l'}) \propto P(z_k)P(t_{l'}|z_k)P(u_{l'}|z_k)P(g_{l'}|z_k)\prod_{j=1}^{J} P(w_{l'j}|z_k) \tag{11}$$

さらに，予測対象である $l'$ 番目の新規出品データの説明変数を $\boldsymbol{x}_{l'} = (1, x_{l'1}, \ldots, x_{l'd}, \ldots, x_{l'D})^{\mathrm{T}}$ とすると，式 (13) で示すように潜在クラス $z_k$ における回帰式の出力 $\hat{y}_{l'k}$ を $P(z_k|t_{l'}, u_{l'}, \boldsymbol{w}_{l'}, g_{l'})$ で混合することで最終的な予測販売価格 $\hat{y}_{l'}$ が得られる．

$$\hat{y}_{l'k} = \hat{\boldsymbol{\beta}}_k^{\mathrm{T}} \boldsymbol{x}_{l'} \tag{12}$$

$$\hat{y}_{l'} = \sum_{k=1}^{K} P(z_k|t_{l'}, u_{l'}, \boldsymbol{w}_{l'}, g_{l'})\hat{y}_{l'k} \tag{13}$$

## 5. 分析モデルを用いた実験

本章では，分析モデルの有効性を示すために，EC サイトに蓄積された実データを用いて，そのテストデータへの予測精度について評価を行う．

### 5.1 実験データ概要

実験データとして，2016 年に EC サイト A 上で取引された，某ファッションブランドの出品履歴データを用いる．データの件数は 67,211 件（$L = 67{,}211$）であり，販売されているアイテムカテゴリ数は 78 種類（$N = 78$）である．また，ソフトクラスタリングを行う際には，前述のとおり，季節ラベル，補助情報，アイテムカテゴリ，出品価格，オフ率を変数として用いている．季節ラベル $s_m$ にはアイテムの出品月を用いるものとし（$M = 12$），アイテムの補助情報 $A_j$ には色，素材などの 8 種類（$J = 8$）を用いた[*4]．

**表 2** 説明変数として用いる質的変数
**Table 2** Categorical variables used as explanatory variables.

| 説明変数 | カテゴリ数 | 説明 |
|---|---|---|
| アイテム | 78 | 対象アイテムの種類 |
| 出品月 | 12 | 1〜12 月 |
| 補助情報 | 84 | 色や素材など |

[*4] 8 種類の補助情報については，EC サイト A における機密事項となるため，その詳細の記載は行わない．

回帰式に用いる説明変数 $\boldsymbol{x}$ には，**表 2** に示す合計 174 次元のダミー変数と出品価格を説明変数（$D = 175$）として用いた．なお，表 2 におけるカテゴリ数とは，各変数がとる値の種類数を示している．

### 5.2 実験概要

本実験では，評価指標として 10 分割交差検定[*5]におけるテストデータに対する平均二乗誤差である $MSE$ と，モデルのあてはまりを評価する $R^2$ 値の 2 つの指標を用いて評価を行うものとした．性能を比較するための手法として，データのクラスタリングを行わない単一の重回帰分析，ランダムフォレスト回帰（以下，RF），多層パーセプトロン（以下，MLP），ならびに $k$-means 法を用いてハードクラスタリングを行い，所属クラスタの回帰式を用いるモデル（以下，比較モデル）を用いた．比較モデルにおける予測では，新規入力データのクラスタは与えられていないが，$k$-means 法の原理を考慮し，各クラスタの代表点との距離が最も近いクラスタの回帰式を用いて予測するものとした[*6]．

なお，RF における決定木の本数は，事前実験の結果から最も精度の高かった 150 とし，MLP に関しては中間層の数が 5 層で各層のニューロンの数を 50 としたモデルを採用した．

### 5.3 実験結果

**図 2**，**図 3** より，学習データへのあてはまりに対しては，RF が最も良い評価値が得られていることが分かる．一方で，**図 4**，**図 5** より，分析モデルは一定の潜在クラス数 $K$ のときに，比較手法よりも良い評価値が得られている

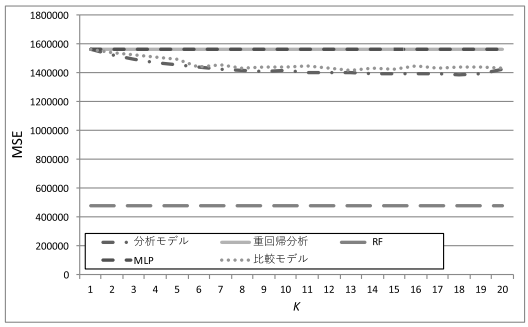


**図 2** 潜在クラス数を変化させたときの学習データに対する $MSE$
**Fig. 2** MSE score of the train data.

[*5] 10 分割交差検定は対象となるデータを 10 分割し，そのうちの 1 つを予測用のテストデータ，残りを学習データとすることでモデルの学習に用い，予測を行うという操作を 10 回繰り返すことを表す．

[*6] 比較モデルは，学習データとテストデータの次元数が同一でなければ，すなわち，テストデータにオフ率の情報がなければその予測を行うことができない．しかし，オフ率は本来，予測を行う際には未知の変数であるため，テストデータに対する特徴量として利用することができない．本研究ではクラスタリングを行ったうえで回帰式を構築する予測の性能を把握することを目的に，オフ率を用いるものとした．このため，実際の価格予測の際には当該モデルは利用できないことに注意されたい．

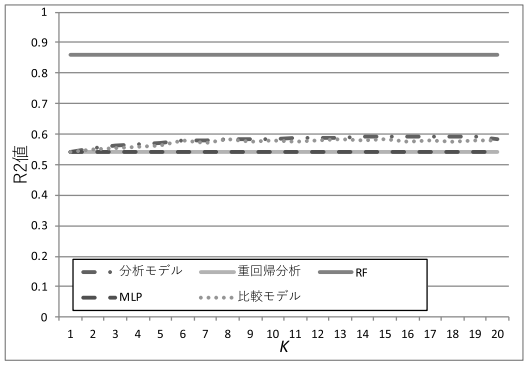


図 3 潜在クラス数を変化させたときの学習データに対する $R^2$ 値
**Fig. 3** $R^2$-value of the train data.

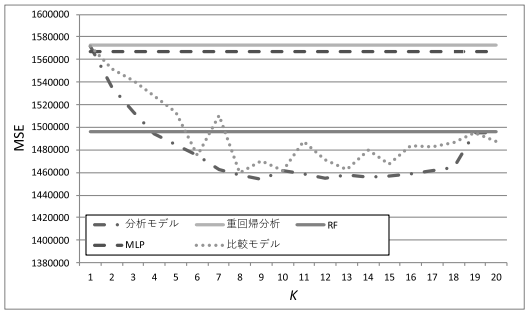


図 4 潜在クラス数を変化させたときのテストデータに対する $MSE$
**Fig. 4** MSE score of the test data.

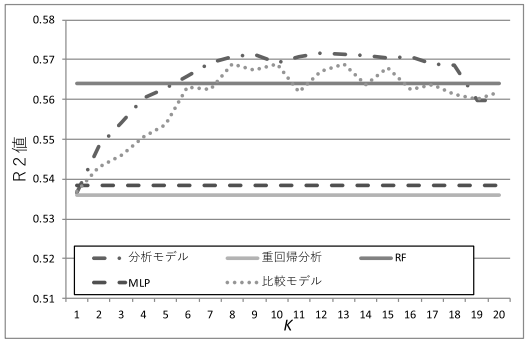


図 5 潜在クラス数を変化させたときのテストデータに対する $R^2$ 値
**Fig. 5** $R^2$-value of the test data.

ことが分かる．また，$K$ の値が大きくなった際に評価指標が低下しているのは過学習が生じたためであると考えられる．この結果より，潜在クラス数 $K$ の設定に留意すれば，分析モデルは EC サイト A におけるアイテムの販売価格を予測するモデルとして有効なモデルであることが分かる．

また，分析モデルを学習データへ適用した際の評価指標が RF よりも低いにもかかわらず，テストデータへの精度が高くなった理由として，以下が考えられる．まず，実験より，RF は学習データに対して予測精度がきわめて良好であるにもかかわらず，テストデータに対する予測精度が悪化していることが分かる．これは，RF のモデルの表現能力が高く，過学習が生じてしまっていることに起因していると考えられる．一方で，分析モデルの学習データへのあてはまりは重回帰分析や MLP を多少上回る程度であるが，潜在クラス数によってはテストデータへのあてはまりがかなり改善している．潜在クラス数が少ない場合，予測精度は RF が分析モデルよりも優れているということから複数の回帰モデルを構築して混合することの効果が見てとれる．また，混合回帰モデルを用いることでテストデータへの予測精度が改善していることは，補助変数の違いによって説明変数と目的変数の関係性が異なることが要因であると考えられる．すなわち，説明変数と目的変数の関係性は線形モデルで説明ができるものの，その関係性については季節ラベルなどの商品属性によって変化すると考えられる．

また，分析モデルと比較モデルにおける結果を比較すると，それぞれの潜在クラス数に対し，比較モデルは分析モデルと類似した傾向となっており，全体的には従来手法と比較して良い性能となっていることが分かる．このことから，本研究の問題設定のもとでは（ソフト，ハード問わず）クラスタリングを実施し，得られたクラスタに対し，回帰式を構築することの有効性が示された．さらに，分析精度はソフトクラスタリングを用いた分析モデルが優れており，潜在クラスモデルによるソフトクラスタリングを用いることで，より高精度な予測が可能となることが分かる．

## 6. 得られた結果の分析

実験の結果，分析モデルは一定の潜在クラス数の場合に比較手法よりも良い評価値が得られ．特に潜在クラス数 $K = 9$ のときに最も高い精度が得られることが明らかになった．分析モデルにより得られた各潜在クラスに所属するアイテムの特徴を把握することで，各潜在クラスの傾向に応じた異なる値付けシステムの構築への応用が期待される．

そこで，以下では $K = 9$ としたときに分析モデルで得られた結果について，各潜在クラスに所属するアイテムの特徴，潜在クラスごとに説明変数が販売価格に与える影響力の 2 つの観点から分析を行う．

まず，どのような特徴量を持ったアイテムが各クラスに所属しているかを分析するために，各潜在クラスのもとでのアイテムの生起確率である $P(i_n|z_k)$，および色や素材といった補助情報の生起確率である $P(a^j_{v_j}|z_k)$ を解釈した結果を**表 3** に示す．また，各クラスでどの季節（月）に出品されたアイテムが出現しやすいかを表す $P(s_m|z_k)$ を**図 6** に示す．

これらの結果を見ると，たとえばクラス 1 には春秋に出品されるデニム，スカートが高い確率で所属しているといったことが分かる．このことから，潜在クラスごとに異なる特徴量や，異なる季節ごとのオフ率の傾向を持ったアイテムが属していることが分かる．

次に，説明変数が目的変数である販売価格に与える影響

表 3 各潜在クラスに高い確率で所属するアイテムの解釈
**Table 3** Features of items belonging to each latent class.

| $k$ | 解釈 |
|---|---|
| 1 | デニムやスカート |
| 2 | メンズのカーディガンなどの上着 |
| 3 | レディースのパンツ類 |
| 4 | バッグなどの小物類 |
| 5 | レディースのサロペット・ジャケット類 |
| 6 | コート類 |
| 7 | メンズの T シャツ類 |
| 8 | レディースの高品質のカットソー |
| 9 | カットソーなどの人気商品 |

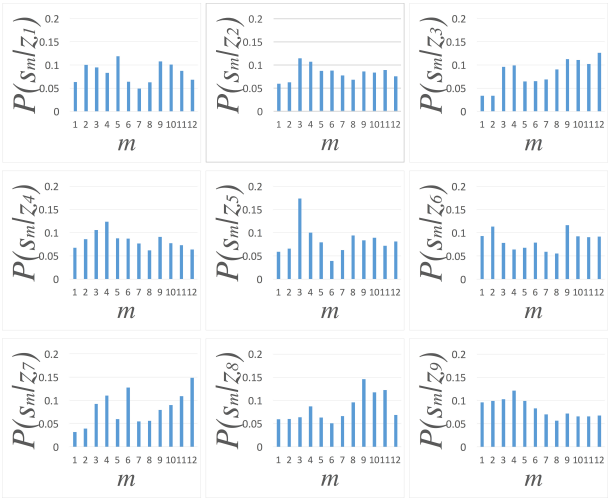


図 6 各潜在クラスにおける出品月の出現確率
**Fig. 6** The conditional probability of exhibit Month given by each latent class.

表 4 各潜在クラスの出品価格の回帰係数と $t$ 値
**Table 4** The coefficients and $t$-value of the exhibit price.

| $k$ | 出品価格の回帰係数 | $t$ 値 |
|---|---|---|
| 1 | 0.685 | 9.92 |
| 2 | 1.085 | 4.32 |
| 3 | 0.376 | 2.81 |
| 4 | 0.661 | 15.5 |
| 5 | 0.047 | 0.05 |
| 6 | 0.704 | 33.2 |
| 7 | 0.259 | 1.72 |
| 8 | 0.168 | - |
| 9 | 0.332 | - |

力の分析を行う．ここでは，各潜在クラスごとの回帰係数と，その説明変数の重要度を測る指標として用いられる回帰係数の $t$ 値の 2 つの観点から分析を行う．本稿では，説明変数の代表として出品価格に注目し，**表 4** に各潜在クラスの出品価格の回帰係数と $t$ 値を示す．なお，潜在クラス 8，9 は，同じ出品価格を持つデータのみが所属し，$t$ 値の算出ができなかったため，- と記すものとした．

表 4 より，各潜在クラスにより異なる回帰係数が得られていることが分かる．すなわち，各潜在クラスごとに目的変数である販売価格に対する各説明変数の影響が異なることが示唆される．たとえば，潜在クラス 2 では，出品価格の回帰係数が 1.085 と他クラスよりも大きな値が得られているので，出品価格が販売価格に与える影響は大きいと解釈することができる．同様に，$t$ 値についても潜在クラスにより異なる値が得られていることが分かる．一般に，重回帰分析における回帰係数 $\beta^d$ の $t$ 値は $d$ 番目の説明変数に対する回帰係数が 0 であることを帰無仮説とした場合の検定推定量となる．すなわち，$t$ 値が大きいほど，この変数が重要であると解釈することができる[*7]．$t$ 値が低くなっている潜在クラスに所属するアイテムは，販売価格の予測の際に出品価格の重要度が低いと考えることもできる．したがって，これらのアイテムでは，出品価格と販売価格の相関が低く，実際に取引されそうな価格を想定した効果的な値付けができていない可能性があると考察できる．

このように，分析モデルを用いることにより，各潜在クラスに所属するアイテムごとに異なる説明変数の販売価格への影響力を定量化できることが可能となった．

## 7. 考察

ファッション系商材は一般に，各アイテムが持つ販売価格に対する要因が，流行やトレンドに加え，季節やアイテムのカテゴリなどによって大きく異なるため，各要因の影響を定量化することや販売価格の予測は相対的に難しい．これに加え，EC サイト A では出品価格から自動で値下げを行うシステムを採用しており，最終的な販売価格の予測の難易度を高めている．このような問題に対し，本研究ではファッション系商材特有の特徴である季節ごとの傾向，ならびに EC サイト A の特徴でもあるオフ率をモデルに組み込んだことで高い精度の予測モデルを構築することができたと考えられる．

また，分析モデルは潜在クラスモデルによるソフトクラスタリングを行った後，複数の線形回帰モデルを混合することでモデルを構築しているため，6 章で述べたように各潜在クラスに所属するアイテムごとの販売価格に対する要因分析が可能となり，その適用範囲を広げることが可能となる．

本研究の最終目標は最適な出品価格の設定である．6 章で述べたように，分析モデルを用いることで各アイテムが持つ販売価格に対する出品価格の影響を定量化することが可能となり，出品価格を変化させた際の販売価格の変化を予測することも可能である．しかしながら，出品価格の引き上げにより顧客の購買行動もまた変化する可能性があ

[*7] 本研究で用いているモデルは回帰の混合により構成されるため，個々の要素の回帰モデルにおける $t$ 値の厳密な検定統計量の議論は理論的に保障されるものではない．しかし，参考値として用いることはできる．

り，その導入に関しては慎重に検討を行う必要があるといえる．

## 8. おわりに

本研究では，ファッションECサイトAにおいて，あらかじめアイテムの基本情報や季節ごとのオフ率の傾向をもとに潜在クラスモデルを用いてソフトクラスタリングを行い，潜在クラスごとに販売価格を目的関数とする回帰式を構築し，それらの値を所属確率で重み付ける予測モデルを提案した．

分析モデルは比較手法よりも新規出品データに対して高い予測精度を示すことが明らかになった．加えて，得られたパラメータを分析することで，それぞれの潜在クラスに所属するデータに対し販売価格に対する有効な要因の分析が可能であることを示した．今後の課題として，在庫期間なども考慮した出品価格の決定手法への反映や本研究で得られた知見の具体的な応用などがあげられる．

**謝辞** 貴重なデータを提供いただき，また日頃から本研究のモデルや結果について実務的側面から様々なアドバイスをいただいているファッション系ECサイトAの皆様に深く感謝をいたします．

### 参考文献

[1] 中村雅章，矢野健一郎：服のインターネット・ショッピングと消費者の知覚リスクに関する実態調査研究，中京企業研究，Vol.35, pp.31–57 (2013).
[2] 中村雅章：インターネット・ショッピングと実店舗を利用したファッション衣料の購買行動，中京ビジネスレビュー=Cyukyo Business Review，Vol.12, No.1, pp.29–62 (2016).
[3] Goto, M., Mikawa, K., Hirasawa, S., Kobayashi, M., Suko, T. and Horii, S.: A New Latent Class Model for Analysis of Purchasing and Browsing Histories on EC Sites, *Industrial Engineering and Management Systems*, Vol.14, No.4, pp.335–346 (2015).
[4] Ben Schafer, J., Konstan, J. and Riedl, J.: Recommender Systems in E-commerce, *Proc. 1st ACM Conference on Electronic Commerce*, pp.158–166, ACM (1999).
[5] 岩永二郎，鍋谷昴一，梶原 悠，五十嵐健太：関心度と忘却度に基づくレコメンド手法—単調性制約付きレコメンドモデルの構築，オペレーションズ・リサーチ，Vol.59, No.2, pp.72–80 (2014).
[6] 田端佑介，堤田恭太，生田目崇：協調フィルタリングと商品の購買間隔を考慮した補正手法による商品推薦システムの提案，オペレーションズ・リサーチ，Vol.61, No.2, pp.97–106 (2016).
[7] Hou, C., Chen, C. and Wang, J.: Tree-Based Feature Transformation for Purchase Behavior Prediction, *IEICE Trans. Inf. and Syst.*, Vol.E101-D, No.5, pp.1441–1444 (2018).
[8] Dias, J.G. and Vermunt, J.K.: Latent Class Modeling of Website Users' Search Patterns: Implications for Online Market Segmentation, *Journal of Retailing and Consumer Services*, Vol.14, No.6, pp.359–368 (2007).
[9] Park, Y.-H. and Fader, P.S.: Modeling Browsing Behavior at Multiple Websites, *Marketing Science*, Vol.23, pp.280–303 (2004).
[10] 石垣 司，竹中 毅，本村陽一：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，人工知能学会論文誌，Vol.26, No.6, pp.670–681（オンライン），DOI: 10.1527/tjsai.26.670 (2011).
[11] 里村卓也：トピックモデルによる顧客データの統合的分析，オペレーションズ・リサーチ，Vol.63, No.2, pp.67–74 (2005).
[12] 杉山啓太，豊田秀樹，長尾圭一郎，磯部友莉恵，岡 律子：ファッションECサイトにおけるイノベーター検出モデル—基準変数のある多種混合の項目反応モデリング，オペレーションズ・リサーチ，Vol.63, No.2, pp.75–82 (2005).
[13] 鶴見裕之，澁谷浩太朗，村瀬明宏：小売業のカテゴリー間プロモーション・マネジメント—消費者の複数カテゴリー購買行動モデル，オペレーションズ・リサーチ，Vol.50, No.2, pp.92–98 (2005).
[14] 本橋永至，樋口知之：市場構造の変化を考慮したブランド選択モデルによる購買履歴データの解析，マーケティング・サイエンス，Vol.21, No.1, pp.37–59 (2013).
[15] 武政孝師，後藤順哉：ECサイトにおける顧客の閲覧履歴を利用した商品ランキング生成法，オペレーションズ・リサーチ，Vol.59, No.8, pp.465–471 (2014).
[16] 高野祐一，田中未来，鮭川矩義，竹山光将，神里 栄，千代竜佑，小林 健，田中研太郎，中田和秀：ファジィクラスタワイズ回帰を用いた共同購入型クーポンサイトの閲覧傾向分析，オペレーションズ・リサーチ，Vol.59, No.2, pp.81–87 (2014).
[17] 西村直樹，鮭川矩義，高野祐一，岩永二郎，水野眞治：ECサイトの商品特性を考慮した2次元確率表による購買予測，オペレーションズ・リサーチ，Vol.60, No.2, pp.69–74 (2015).
[18] 伊藤孝太朗，澤邊 剛，保坂桂佑，松下亮祐，雪島正敏：顧客のセグメンテーションと商品のスコアリングによる購買予測，オペレーションズ・リサーチ，Vol.60, No.2, pp.75–80 (2015).
[19] 山下 遥，鈴木秀男：セール品に注目した顧客の購買行動の解析—2値データのクラスタリングを考慮したロジスティック回帰分析，オペレーションズ・リサーチ，Vol.60, No.2, pp.81–88 (2015).
[20] Platzer, M. and Reutterer, T.: Ticking Away the Moments: Timing Regularity Helps to Better Predict Customer Activity, *Marketing Science*, Vol.35, pp.779–799 (2016).
[21] 白井康之，森田裕之，Cheung, S.，中元政一，高嶋宏之：商品の潜在的類似性に基づくクラスタリング手法の提案，オペレーションズ・リサーチ，Vol.61, No.2, pp.80–87 (2016).
[22] 北島良三，遠藤啓太，上村龍太郎：入力ニューロンの潜在性に着目した小売店店舗の非継続来店顧客検知モデルの作成，オペレーションズ・リサーチ，Vol.61, No.2, pp.88–96 (2016).
[23] 三好哲也：アパレルオンラインショッピングにおける消費者特性の分析：データ分析コンペティションデータの分析を通して，経営システム，Vol.27, No.2, pp.61–69 (2017)（オンライン），入手先 ⟨https://ci.nii.ac.jp/naid/40021268532/⟩.
[24] 白井康之，森田裕之，後藤祐介：商品の潜在的類似性に基づくクラスタリング手法の提案，オペレーションズ・リサーチ，Vol.62, No.2, pp.91–99 (2017).
[25] 日経コンピュータ 2018年1月18日号：特集 最適価格はAIに聞け—「値付け」変幻自在，利益最大化」(2018).
[26] Bishop, C.: *Pattern Recognition and Machine Learning*, Springer-Verlag, New York (2006).
[27] Witten, I.H. and Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Ed.,

Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2005).
[28] Faria, S. and Soromenho, G.: Fitting mixtures of linear regressions, *Journal of Statistical Computation and Simulation*, Vol.80, No.2, pp.201–225 (2010).
[29] Hofmann, T.: Latent semantic models for collaborative filtering, *ACM Trans. Information Systems* (*TOIS*), Vol.22, No.1, pp.89–115 (2004).
[30] Swait, J. and Adamowicz, W.: The Influence of Task Complexity on Consumer Choice: A Latent Class Model of Decision Strategy Switching, *Journal of Consumer Research*, Vol.28, No.1, pp.135–148 (2001).
[31] Hofmann, T.: Probabilistic latent semantic analysis, *Proc. 15th Conference on Uncertainty in Artificial Intelligence*, pp.289–296, Morgan Kaufmann Publishers Inc. (1999).
[32] Hofmann, T.: Probabilistic Latent Semantic Indexing, *Proc. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, *SIGIR '99*, pp.50–57, ACM (online), DOI: 10.1145/312624.312649 (1999).
[33] Leisch, F.: FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R, *Journal of Statistical Software*, Vol.11, No.8, pp.1–18 (online), DOI: 10.18637/jss.v011.i08 (2004).
[34] Chen, D., Wang, D., Yu, G. and Yu, F.: A PLSA-based approach for building user profile and implementing personalized recommendation, *Advances in Data and Web Management*, pp.606–613, Springer (2007).
[35] Dempster, A.P., Laird, N.M. and Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B* (*methodological*), Vol.39, No.1, pp.1–22 (1977).
[36] Lindsay, B.G.: Mixture Models: Theory, Geometry and Applications, *NSF-CBMS Regional Conference Series in Probability and Statistics*, Vol.5, pp.i–163 (1995) (online), available from ⟨http://www.jstor.org/stable/4153184⟩.
[37] Wedel, M. and Kamakura, W.: *Market segmentation: Conceptual and methodological foundations*, Kluwer Academic Publishers (1999).
[38] 永森誠矢, 山下 遥, 荻原大陸, 後藤正幸：混合回帰に基づく就職ポータルサイトの被エントリ数分析モデルに関する一考察, 情報処理学会論文誌, Vol.59, No.4, pp.1273–1285 (2018).
[39] Cleveland, W.S. and Devlin, S.J.: Locally weighted regression: An approach to regression analysis by local fitting, *Journal of the American Statistical Association*, Vol.83, No.403, pp.596–610 (1988).

# 付 録

## A.1 分析モデルのパラメータ更新式

以下では，分析に用いたモデルのパラメータの更新式について述べる．

### A.1.1 潜在クラスによるクラスタリング

式 (9) で表される対数尤度 $LL$ が収束するまで，以下の更新式を用いることでパラメータの更新を行う．

[E-step]

$$P(z_k|t_l,u_l,\boldsymbol{w}_l,g_l,h_l) \propto P(z_k)P(t_l|z_k)P(u_l|z_k)P(g_l|z_k)P(h_l|z_k)\prod_{j=1}^{J}P(w_{lj}|z_k) \tag{A.1}$$

[M-step]

$$P(z_k) \propto \sum_{l=1}^{L}\alpha_{kl} \tag{A.2}$$

$$P(i_n|z_k) \propto \sum_{l=1}^{L}\alpha_{kl}\delta(t_l=i_n) \tag{A.3}$$

$$P(s_m|z_k) \propto \sum_{l=1}^{L}\alpha_{kl}\delta(u_l=s_m) \tag{A.4}$$

$$P(a_{v_j}^j|z_k) \propto \sum_{l=1}^{L}\alpha_{kl}\delta(w_{lj}=a_{v_j}^j) \tag{A.5}$$

$$\mu_k = \frac{\sum_{l=1}^{L}\alpha_{kl}g_l}{\sum_{l=1}^{L}\alpha_{kl}} \tag{A.6}$$

$$\sigma_k^2 = \frac{\sum_{l=1}^{L}\alpha_{kl}(g_l-\mu_k)^2}{\sum_{l=1}^{L}\alpha_{kl}} \tag{A.7}$$

$$\lambda_k = \frac{\sum_{l=1}^{L}\alpha_{kl}h_l}{\sum_{l=1}^{L}\alpha_{kl}} \tag{A.8}$$

$$\varphi_k^2 = \frac{\sum_{l=1}^{L}\alpha_{kl}(h_l-\lambda_k)^2}{\sum_{l=1}^{L}\alpha_{kl}} \tag{A.9}$$

なお，上式 (A.2)–(A.9) において，式の簡素化のために，$\alpha_{kl}=P(z_k|t_l,u_l,\boldsymbol{w}_l,g_l,h_l)$ とした．

### A.1.2 潜在クラスによる混合回帰モデル

いま，以下のように $\boldsymbol{X}$，$\boldsymbol{Y}$，$\boldsymbol{B}_k$，$\boldsymbol{W}_k$ を定義する．

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1d} & \cdots & x_{1D} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{l1} & \cdots & x_{ld} & \cdots & x_{lD} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{L1} & \cdots & x_{Ld} & \cdots & x_{LD} \end{pmatrix} \tag{A.10}$$

$$\boldsymbol{Y} = (y_1,\ldots,y_l,\ldots,y_L)^{\mathrm{T}} \tag{A.11}$$

$$\boldsymbol{B}_k = (\beta_{k0},\beta_{k1},\ldots,\beta_{kd},\ldots,\beta_{kD})^{\mathrm{T}} \tag{A.12}$$

$$\boldsymbol{W}_k = \begin{pmatrix} \alpha_{k1} & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & \alpha_{kl} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & \alpha_{kL} \end{pmatrix} \tag{A.13}$$

このとき，式 (10) で表される目的関数である各データの各潜在クラス $z_k$ への所属確率で重み付けされた二乗誤差は，以下の式 (A.14) で表される $S_e$ のように書き換えられる．

これを $B_k$ に関して最小化することで式 (A.15) で表される更新式を得ることができる．

$$S_e = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}_k)^{\mathrm{T}} \boldsymbol{W}_k (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}_k) \quad \text{(A.14)}$$

$$\boldsymbol{B}_k = (\boldsymbol{X}^{\mathrm{T}} \boldsymbol{W}_k \boldsymbol{X})^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{W}_k \boldsymbol{Y} \quad \text{(A.15)}$$


仁ノ平 将人

1994 年生．2018 年早稲田大学大学院修士課程修了．在学時，機械学習手法を用いた購買データの分析に関する研究に従事．

三川 健太

1981 年生．2005 年武蔵工業大学環境情報学部環境情報学科卒業．2007 年同大学大学院修士課程修了．2016 年早稲田大学大学院博士後期課程修了．博士（工学）．2013 年早稲田大学助手．2016 年湘南工科大学工学部情報工学科講師．機械学習とその応用に関する研究に従事．IEEE，電子情報通信学会，日本経営工学会等，各会員．

後藤 正幸 （正会員）

1969 年生．1994 年武蔵工業大学大学院修士課程修了．2000 年早稲田大学大学院博士課程修了．博士（工学）．1997 年早稲田大学理工学部助手．2000 年東京大学大学院工学系研究科助手．2002 年武蔵工業大学環境情報学部情報メディア学科助教授．2008 年早稲田大学創造理工学部経営システム工学科准教授．2011 年同大教授．情報数理応用とデータサイエンス，ならびにビジネスアナリティクスの研究に従事．著書に，『入門パターン認識と機械学習』，コロナ社 (2014)，『ビジネス統計～統計基礎とエクセル分析』，オデッセイコミュニケーションズ (2015) 等．IEEE，電子情報通信学会，人工知能学会，日本経営工学会，経営情報学会等，各会員．