膨大な種類のアイテムを対象とした消費者購買行動の分析モデルに関する研究

情報数理応用研究　　5218C039-0　安井一貴
指導教員　後藤正幸

An Analytical Model of Consumers Purchasing Behavior for Huge Kinds of Products

YASUI Kazuki

1. 研究背景・目的

近年の情報技術の発展に伴い，消費者の購買履歴データを大量に取得・蓄積することが可能となった．これらのデータを用いて消費者の購買行動を把握し，消費者それぞれの特性に合わせたマーケティング施策を講じることは，多くの小売企業の売上向上のために有効な手段である．例えば，各消費者の購買履歴データから，消費者ごとの購買特性を確率モデルで特定できれば，特性が類似した消費者グループの抽出や消費者 1 人 1 人の購買傾向や嗜好の分析など，様々なマーケティング活動に有用となる．しかし，存在するアイテムの種類数が膨大になると，すべてのアイテムを考慮して各消費者の購買行動モデルを構築することは現実的ではない．本研究では，アイテムの種類数が膨大な場合における，個々の消費者の購買行動分析を目指す．

通常，消費者の購買履歴データの分析にはいくつかの統計モデルが使用される．パラメトリックな統計モデルを導入するためには，取得されている観測データを訓練データとし，そのデータが正規分布などの既定のモデルに当てはまることを仮定して，それらのパラメータを推定する必要がある．代表的なパラメトリック統計モデルのパラメータ推定法としては，最尤推定法 [1] やベイズ推定法 [2] が挙げられる．最尤推定法では，尤度関数を最大化するようにパラメータが推定される．例えば，各消費者のアイテムの購入確率が多項分布に従うとして，パラメータの最尤推定を行う場合，今までに購入したアイテム (以下，購入アイテム) しか考慮できず，今までに購入されていないアイテム (以下，未購入アイテム) の購入確率の推定値は 0 となってしまう．しかし，アイテムの種類が膨大な場合，有限の学習データ内では未購入アイテムが多く存在してしまい，予測モデルとしては使えなくなってしまうという問題がある．一方，ベイズ推定法では存在する全てのアイテムを考慮した事前確率分布を仮定する．事前確率分布を導入することによって，未購入アイテムの推定購入確率を 0 にすることなく推定を行うことができる．しかし，事前確率の合計は 1 でなければならないという制約があるため，アイテムの種類数が膨大になると事前確率が小さくなってしまい適切な推定ができない恐れがある．

そこで本研究では，アイテムの種類数を無限大に拡張し，ノンパラメトリックな統計モデルによってモデル化することにより，膨大な種類のアイテムを考慮した購入分析モデルを提案する．具体的には，無限のアイテム集合を仮定し，可算の無限アイテム集合の購入確率分布を考慮することで，多種多様なアイテムの購入確率のモデル化を行う．提案モデルを用いることにより，あまり購入されていない多くのアイテムを考慮した分析モデルを構築することが可能となる．提案モデルの有効性を示すため，提案モデルを株式会社マクロミル提供の実購買データに適用し，得られる結果について考察を与える．

2. 準備

2.1. 問題設定

本研究では，異なりアイテム数が極めて多く，さらにそれが時間経過とともに増加していくような消費者の購入履歴データ (例えば，消費者の日々の全購入アイテムを記録したパネルデータなど) を扱うものとする．このようなデータに対し，アイテム数が無限であると仮定することで，様々なアイテムの購入確率を各消費者に対してモデル化する．アイテムの確率分布をノンパラメトリックに拡張することにより，あまり購入されていない多くのアイテムを考慮した分析モデルの構築が可能となる．

2.2. Chinese Restaurant Process

Chinese Restaurant Process (以下，CRP) [3] は離散確率過程モデルの 1 つであり，可算無限の事象集合上の確率過程を少ないパラメータで定義している点が特徴である．原理の説明のため，あるレストランに客が 1 人ずつ来店し，順番にテーブルに着席していく状況を考える．n 番目のデータ x_n を n 番目に来店した客，その客が着席するテーブルを離散事象と解釈している．CRP のイメージを以下の図 1 に示す．

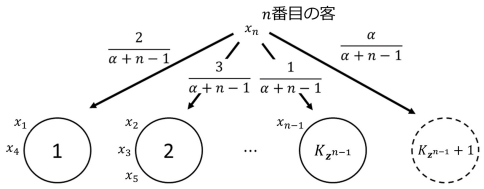

図 1: CRP のイメージ

レストランには 1, 2, ... , とラベル付けされた無数のテーブルの存在を仮定する．レストランに来店した客は以下のプロセスに従いテーブルを選択する．ここで，x_n が着席したテーブルの番号を z_n とし，n 人の客それぞれが着席したテーブルの番号をまとめて $\boldsymbol{z}^n = (z_1, z_2, ..., z_n)$ と記述する．

Step 1: 1 番目に来店した客 x_1 は，$z_1 = 1$ を選択する．

Step 2: n 番目 (>1) に来店した客 x_n は，今まで来店した客が座っているテーブルの状況に依存し，以下の式 (1) に従い既存の (他の客が着席している) テーブルか新規の (他の客が着席していない) テーブルを選択するかを決定する．

$$p(z_n = k|\boldsymbol{z}^{n-1}) = \begin{cases} \frac{c_k}{\alpha+n-1} & (k = 1, 2, ..., K_{\boldsymbol{z}^{n-1}}) \\ \frac{\alpha}{\alpha+n-1} & (k = K_{\boldsymbol{z}^{n-1}} + 1) \end{cases} \quad (1)$$

ここで，c_k は $\boldsymbol{z}^{n-1}$ の中で k が現れた回数，$K_{\boldsymbol{z}^{n-1}}$ は $\boldsymbol{z}^{n-1}$ 中で着席されているテーブルの数を表す．そして，α は集中度パラメータであり，この値が大きいほど客は新しいテーブルに着席する確率が高くなり，小さいほど既存のテーブルに着席する確率が高くなる．このように，最初は必ずテーブル 1 が選ばれるが，n の増大とともに次第に使われるテーブル数 $K_{\boldsymbol{z}^n}$ が増えていくことになる．この特徴を用いて，ディリクレ過程混合モデルなどの手法に使用されている．

3. 提案モデル

3.1. 概要

本研究の目的は，存在するアイテムの種類数が非常に多い場合における消費者 1 人 1 人の購買傾向や嗜好の違いを分析することである．すなわち，消費者ごとに各アイテムを購入する購入確率分布を推定する問題を考える．現実的な状況として，存在する膨大なアイテムの中には，個々の消費者にとっては購入する可能性がない「考慮する必要のないアイテム」が存在すると考えられる．しかし，パラメトリックな統計モデリングに基づくと，全ての消費者に対して膨大に存在する全てのアイテムの購入確率を多項分布パラメータとして定義するため，適切なパラメータ推定が難しくなるといった問題が生じてしまう．

そこで本研究では，アイテムの種類数を無限大に拡張してノンパラメトリックな統計モデルによってモデル化することにより，膨大な種類のアイテムを考慮した購入分析モデルを提案する．具体的には，CRP をアイテムの種類に対して適用し，可算無限のアイテム集合を想定したモデルを構築する．このとき，CRP における集中度パラメータ α を推定することにより，各消費者の購入アイテムと未購入アイテムの購入しやすさを推定することが可能となる．

3.2. 提案モデルの定式化

CRP の考えを応用して，アイテムの種類数が無限大をとり得るノンパラメトリックな統計モデルによるモデル化を行う．ここで，消費者集合を $\mathcal{U} = \{u_i : 1 \leqq i \leqq U\}$，アイテム集合を $\mathcal{P} = \{p_l : 1 \leqq l < \infty\}$ とする．

CRP では観測データ x_n を n 番目の客，テーブルをクラスタとしていた．本提案モデルでは，消費者 u_i が n 番目に購入したアイテムを x_{in}，テーブルをアイテムの種類と解釈して分析を行う．また，消費者 u_i が購入したアイテムの総数を N_i，n 番目に購入したアイテムに割り当てられるテーブル番号を z_n^i とし，$\boldsymbol{z}_i^{N_i} = (z_1^i, z_2^i, ..., z_{N_i}^i)$ と定義する．すなわち，新しい種類のアイテムが購入されると新しいテーブルが追加されることになる．

提案モデルの確率計算ではまず，消費者 u_i の購入したアイテムを時系列に並べる．次に，時系列に並べたアイテムを順次確認していく．このとき，アイテム x_{in} がそれ以前に購入した $n-1$ 個のアイテム $x_{i1}, x_{i2}, ..., x_{in-1}$ に含まれる $K_{\boldsymbol{z}_i^{n-1}}$ 種類の中に存在しているかを判定する．存在していればアイテム $x_{in} = p_k$ の購買確率を $\frac{c_k^i}{\alpha_i + n - 1}$ とし，存在していない (新規アイテム) であれば，そのアイテム x_{in} の購買確率を $\frac{\alpha_i}{\alpha_i + n - 1}$ とする．この操作を N_i 回繰り返す．ここで，c_k^i は $\boldsymbol{z}_i^{n-1}$ の中で k が現れた回数，α_i は消費者 u_i の未購入アイテムの購入しやすさを表すパラメータである．

その結果，アイテムの確率分布をノンパラメトリックモデルに拡張することが可能となる．これにより，存在する膨大なすべてのアイテムを対象とした確率推定の必要がなくなるため，消費者 u_i の各アイテムの購入確率が小さくなり過ぎることを防ぐことができる．さらに，新しいアイテムが追加されるたびに最初から学習をやり直す必要がなくなる．

以上の議論より，消費者 u_i がアイテム $x_{i1}, x_{i2}, ..., x_{iN_i}$ を順に購入する (つまり，テーブルが $z_1^i, z_2^i, ..., z_{N_i}^i$ の順に選ばれる) 確率は以下の式 (2) で定義することができる．

$$\begin{aligned} p(\boldsymbol{z}_i^{N_i}|\alpha_i) &= p(z_{N_i}^i|\boldsymbol{z}_i^{N_i-1}, \alpha_i) p(\boldsymbol{z}_i^{N_i-1}|\alpha_i) \\ &= p(z_{N_i}^i|\boldsymbol{z}_i^{N_i-1}, \alpha_i) p(z_{N_i-1}^i|\boldsymbol{z}_i^{N_i-2}, \alpha_i) p(\boldsymbol{z}_i^{N_i-2}|\alpha_i) \\ &= \cdots \\ &= \prod_{j=1}^{N_i} p(z_j^i|\boldsymbol{z}_i^{j-1}, \alpha_i) \end{aligned} \tag{2}$$

3.3. 提案モデルのパラメータ推定と予測

α_i は，以下の式 (3) を用いることで消費者 u_i の購買履歴データから推定することが可能である．上式 (2) は消費者 u_i がテーブルを $z_1^i, z_2^i, ..., z_{N_i}^i$ の順に選択する確率を示している．したがって，式 (2) の確率が最も大きくなる α_i の値 $\hat{\alpha}_i$ を推定することで，消費者 u_i が新規アイテムをどれくらい買い易いかを推定することができる．

$$\hat{\alpha}_i = \arg\max_{\alpha_i} \prod_{j=1}^{N_i} p(z_j^i|\boldsymbol{z}_i^{j-1}, \alpha_i) \tag{3}$$

$\hat{\alpha}_i$ が大きければ様々な種類のアイテムを購買する傾向にある消費者，$\hat{\alpha}_i$ が小さければ同じ種類のアイテムを購入しやすい傾向にある消費者と捉えることができる．

ここで，提案モデルを用いて購入アイテムの予測を行う際，古くから存在しているテーブル (アイテム) は長い時間をかけて消費者が購入しているアイテムと考えられるため，データが蓄積され購入数が相対的に多くなってしまい予測精度に影響すると考えられる．そこで，各テーブルに所属するアイテムの総数に対して重みを付与することで，古くから存在し大量に購入されているアイテムの影響を小さくする．本提案モデルでは，シグモイド関数を用いてテーブルに重みを付与する．消費者 u_i の k 番目のテーブルに対する重みを w_k^i と置き，以下の式 (4) で定義する．ここで，β は，シグモイド関数における係数 (ゲイン) である．その結果，u_i の既存の各テーブル選択確率 p_k と新規テーブルの選択確率 $\hat{y}_i$ は，それぞれ以下の式 (5), (6) で表される．ここでは，$\boldsymbol{z} = \boldsymbol{z}_i^{N_i}$ と簡略化して表記している．

$$w_k^i = \frac{1}{1 + e^{-\beta \frac{k}{K_{\boldsymbol{z}}+1}}} (k = 1, 2, ..., K_{\boldsymbol{z}} + 1) \tag{4}$$

$$p_k = \frac{w_k^i c_k^i}{\hat{\alpha}_i + N_i} (k = 1, 2, ..., K_{\boldsymbol{z}}) \tag{5}$$

$$\hat{y}_i = \frac{w_k^i \hat{\alpha}_i}{\hat{\alpha}_i + N_i} (k = K_{\boldsymbol{z}} + 1) \tag{6}$$

4. 予測精度検証実験

提案モデルの有効性を示すため，検証実験を行う．

4.1. 実験条件

提案モデルを用いて具体的に実データの分析を行う前に，提案モデルが従来手法である最尤推定法とベイズ推定法に比べて，消費者の購買傾向を精度よく把握することができているかを示すため，予測精度による検証実験を行う．実験に用いたデータは株式会社マクロミルより提供いただいた消費者購買履歴データ QPR で，2015 年の 1 月 1 日から 2015 年の 12 月 31 日までの期間にそれぞれの消費者が購入したアイテムのデータである．対象消費者数は $U = 7,870$，総アイテム種類数は 317,796 である．また，各消費者におけるアイテムの平均購入数は 1,245 個，アイテムの平均購入種類数は 513 種類である．

実験は，消費者 u_i の総購入アイテム数 N_i を学習データ数 $N_i - 10$，テストデータ数 10 に分割し，最尤推定法，ベイズ推定法，提案モデルを用いて「推定確率を用いた購入アイテム予測」と「テストデータの購買に含まれる未購入アイテムの個数予測」をそれぞれ行い，その精度を比較した．また，ベイズ推定法のアイテム購入確率の事前分布は一様分布とした．各実験において，前者の評価手法は，Top$N$ 精度を用い，以下の式 (7) で定義する．ここで，J は予測したアイテムのうち，実際に購入されたアイテム数，T は予測アイテム数であり，本実験では $T = 1, 3, 5$ とした．後者の評価手法は，Mean Absolute Error(MAE) を用い，以下の式 (8) で定義する．ここで，y_i は消費者 u_i のテストデータの中に存在する未購入アイテムの個数である．

$$\text{Top}\,N = \frac{J}{T \times U} \tag{7}$$

$$MAE = \frac{1}{U}\sum_{i=1}^{U} |y_i - 10\hat{y}_i| \tag{8}$$

また，各テーブルのアイテム数量に対する重み w_k^i のパラメータ β は事前実験より $\beta = 2.5$ とした．

4.2. 実験結果

購入アイテムと未購入アイテムの実験結果を以下に示す．

表 1: 購入アイテムの Top$N$ 精度

T	**1**	**3**	**5**
最尤推定法	0.244	0.175	0.143
ベイズ推定法	0.244	0.175	0.143
提案モデル	**0.291**	**0.229**	**0.199**

表 2: 未購入アイテムの予測個数誤差

	最尤推定法	ベイズ推定法	提案モデル
MAE	4.85	5.14	**2.47**

表 1 より，$T = 1, 3, 5$ すべての場合において提案モデルの Top$N$ 精度が良いことが分かる．また，表 2 より提案モデルの予測誤差が一番小さいことが分かる．このことから，未購入アイテムの予測購入確率は消費者の傾向を適切に捉えられていると考えられる．

以上より，予測精度という観点からの提案モデルの有効性を示すことができた．

5. 実データ分析

本章では，提案モデルをアイテムの種類数が膨大な株式会社マクロミル提供の消費者購買履歴データ QPR に適用することで，各消費者の新規アイテムの購入し易さの分析を行った結果について示す．対象期間は 2015 年の 1 月 1 日から 2015 年の 12 月 31 日，対象消費者は総購買数 100 以下の者を除外し，U=7,406 とした．

5.1. 分析結果

(a) $\hat{\alpha}_i$ に関する分析

各消費者 u_i について $\hat{\alpha}_i$ を求めた結果を以下の図 2 に示す．加えて，$\hat{\alpha}_i$ の大きさ上位 5 人と下位 5 人の詳細をそれぞれ以下の表 3, 4 に示す．

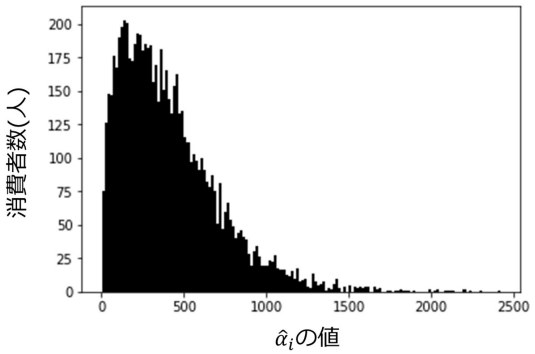

図 2: $\hat{\alpha}_i$ の大きさの分布

表 3: $\hat{\alpha}_i$ の値上位 5 人

	総購買数	アイテムの種類数	$\hat{\alpha}_i$
1	238	237	12,901.0
2	589	553	4,415.1
3	243	237	3,739.6
4	1,424	1,162	2,979.8
5	1,056	892	2,660.4

表 4: $\hat{\alpha}_i$ の値下位 5 人

	総購買数	アイテムの種類数	$\hat{\alpha}_i$
1	118	6	0.5
2	129	3	0.5
3	143	4	0.7
4	182	6	1.2
5	216	8	1.6

図 2 より，$\hat{\alpha}_i$ の分布は単峰形で，その平均は 416.8 であった．すなわち，複数の潜在的なグループは存在しないと考えられるため，平均的な $\hat{\alpha}_i$ の値を持つ消費者を一般的な購買行動をとる消費者と見なすことができる．一方で，$\hat{\alpha}_i$ の分布の裾が伸びだす $\hat{\alpha}_i = 1500$ 以上の値をとる消費者は 69 人存在した．特に，最も大きい値は $\hat{\alpha}_i = 12,901.0$ であった．この消費者について見てみると，238 個のアイテムを購入した中で，237 種類のアイテム，すなわち，同じアイテムはほとんど購入せず，様々な種類のアイテムを購入していた．このような消費者は多種多様なアイテムを好む傾向が非常に強い消費者であると考えられる．

(b) 未購入アイテムの購入確率についての分析

表 3, 4 より総購買数に対するアイテムの種類の割合「アイテム種類数/総購買数」に着目する．このとき,「アイテム種類数/総購買数」の大きい消費者ほど $\hat{\alpha}_i$ の値が大きい傾向があり,「アイテム種類数/総購買数」の小さい消費者ほど $\hat{\alpha}_i$ の値が小さい傾向が見られる．しかし，$\hat{\alpha}_i$ と「アイテム種類数/総購買数」の大小関係の傾向が逆になっている消費者がいくらか存在する．表 5 に，その一例を示す．

表 5: $\hat{\alpha}_i$ と新規アイテムの購買確率の関係

比較項目 \ 消費者	A	B
総購買数 N_i	3,603	123
アイテムの種類数	2,255	112
アイテム種類数/総購買数	0.626	**0.911**
$\hat{\alpha}_i$ の値	**2,582**	569.7
新規アイテムの購入確率	0.418	**0.822**

表 5 は N_i が大きく異なる消費者の一例である．表 5 の消費者 B は購入したアイテムがほぼ異なるため，消費者 A に比べて「アイテム種類数/総購買数」の値は大きいが，$\hat{\alpha}_i$ の値は消費者 A の方が大きいことが分かる．このことから，$\hat{\alpha}_i$ の値は「アイテム種類数/総購買数」の大きさだけでなく，消費者の総購買数に少なからず影響を受けていると考えられる．すなわち，総購買数が少ない場合，様々な種類のアイテムを購入していても $\hat{\alpha}_i$ の値は大きくなりにくい．したがって，$\hat{\alpha}_i$ の値だけですべての判断するのは適切でないと考えられる．

また，$\hat{\alpha}_i$ を用いて各消費者 u_i の新規アイテムの購買確率を求めた結果，平均購買確率は図 3 と同様に単峰性の分布となり，新規アイテムの購買確率の平均は 0.259 であった．すなわち，平均的な消費者はアイテムを購入するとき，0.259 程度の確率で未購入アイテムを購入すると言える．また，未購入アイテムの購買確率が 0.5 以上の消費者が 482 人存在した．これらの消費者は，アイテムを購入するとき 2 分の 1 以上の確率で未購入アイテムを購入する嗜好の多様性が高いグループといえる．ただし，これらの消費者の中にはアイテムの総購買数が少ない消費者も存在した．これらの消費者は購入したアイテムが少ないため，次回購入アイテムが購入したことがあるアイテムと重複する可能性が低い．このような購買を始めたばかりの消費者は，未購入アイテムの購買確率が大きく推定されてしまっている．ただし，総購買数が小さければ $\hat{\alpha}_i$ の値は大きくならない傾向がある．つまり，新規アイテムの購買確率と $\hat{\alpha}_i$ の値の両方を考慮して分析を行うことが重要であると考えられる．

(c) 消費者の嗜好と $\hat{\alpha}_i$ の関係性分析

次に，消費者の購買傾向と求めた $\hat{\alpha}_i$ の関係性を分析する．まず，各消費者が購入したアイテムをいくつかのカテゴリにまとめ，各アイテムカテゴリに属するアイテムの購入数を要素とするベクトルを構成し，k-means を用いて消費者を簡易的に分類する．その結果得られた各グループにおける $\hat{\alpha}_i$ の平均値を求めた．分析結果を以下の表 6 に示す．

表 6 より，書籍を大量に購入する消費者グループは $\hat{\alpha}_i$ の値が非常に大きい傾向が見られた．また，食品を大量に購入する消費者グループの $\hat{\alpha}_i$ の値も大きい．このことは，様々な種類がある書籍や食品を購入するグループの消費者は，新規アイテムを購入する傾向が強いことを示唆する．一方で，清涼飲料や酒を多く購入する消費者グループは $\hat{\alpha}_i$ の値が小さい結果となった．このことから，消費者は清涼飲料・酒に関して，決まった好みのものが存在し，それらを多く購入している可能性が考えられる．

表 6: 消費者の嗜好と $\hat{\alpha}_i$ の関係

所属する消費者の特徴	$\hat{\alpha}_i$ の平均値
書籍を大量に購入するグループ	517.1
食品関係を多く購入するグループ	423.1
菓子・清涼飲料・パンを多く購入するグループ	318.8
酒を多く購入するグループ	281.5
清涼飲料を多く購入するグループ	149.0

6. 考察

本研究で得られた $\hat{\alpha}_i$ を活用することにより，様々なアイテムを広く購入する消費者や決まったアイテムのみを購入する傾向が強い消費者を把握することができると考えられる．本提案モデルを用いて新しいもの好きの消費者を把握することにより，新商品を発売する際の効率的なマーケティング施策に結び付くことが期待される．近年，小売企業で扱われるアイテムの数は非常に多いため，提案モデルを効果的に適用できる場面が多くあると予想される．

また，本提案モデルは，消費者数が一定でアイテム数が時間とともに徐々に増加する購買履歴データなどの分析にも有効であると考えられる．本研究で用いた QPR データにはこのような特性があるため，提案モデルを用いた分析に有効であると考えられる．

7. まとめと今後の課題

本研究では，アイテムの購入確率モデルをノンパラメトリックモデルに拡張することにより，膨大な種類のアイテム集合に対応できる購入分析モデルを提案した．そのために，消費者が購入する多種多様なアイテムを柔軟に表す CRP のアイデアを導入した．実購買履歴データに提案モデルを適用させ分析を行うことで，モデルの有効性を示した．

今後の課題として，今回の分析モデルでは $\hat{\alpha}_i$ や新規アイテムの購入確率を推定することができたが，アイテムの購入順序による傾向の分析や新規に購入するアイテムの種類までは考慮できていない．そのため，購入順序による違いや消費者の購入したアイテムの詳細も考慮可能な分析モデルが必要であると考えられる．

参考文献

[1] Le Cam, L, “Maximum likelihood: an introduction,” *International Statistical Review/Revue Internationale de Statistique*, Vol. 58, No. 2, pp.153–171, 1990.

[2] D. V. Lindley, “Introduction to Probability and Statistics from a Bayesian Viewpoint,” *Cambridge University Press*, 2009.

[3] 石井 健一郎，上田 修功，続・わかりやすい パターン認識 −教師なし学習入門−，オーム社，2014.