形状制約モデルによる顧客の商品選択行動の予測

２０２０年　３月

西村　直樹

形状制約モデルによる
顧客の商品選択行動の予測

西村 直樹

システム情報工学研究科
筑波大学

2020 年 3 月

概要

本研究は，EC（Electronic Commerce：電子商取引）サイトの閲覧履歴や商品選択履歴を含むページ遷移データや，実店舗の販売履歴を記録した POS（Point of Sales：販売時点）データを用いた顧客の商品選択行動のモデル化と，商品選択予測の精度向上を対象とする．商品選択行動の推定精度が向上すれば，顧客に対する商品推薦や商品配置の決定などのマーケティング施策による効果が増加する [10]．また単純にどの商品が選択されるかのみでなく，商品が選択されるまでの過程に関する知見を得ることも，これらの施策を行う実務的観点から重要である [30]．

岩永ら [32] は，ページ遷移データの解析に基づき，顧客セグメンテーションの方法である RFM 分析 [31] の指標としても知られる最新度 (Recency) と頻度 (Frequency) という二つの指標と商品選択との関係に着目した．そして，最新度と頻度に関する単調性，凸性，凹性を仮定し，制約条件として課して商品選択確率を推定する手法を提案した．この形状制約モデルは，ページ遷移データの分析に対する形状制約回帰の応用として新しい，有効な手法である．

本研究でも，岩永ら [32] のアプローチと同様に，まず顧客の商品選択行動の解析のため，顧客の実際の商品選択行動を記録したデータを解析する．そして，顧客の商品選択に影響を与える特徴量を探索する．その特徴量の観察から，顧客の商品選択に関する新たな性質を仮定し，仮定を形状制約として課してモデルのパラメータを推定することで，選択予測モデルの精度の向上を試みる．さらに，実際のデータから推定されたモデルのパラメータを観察することで，顧客の商品選択の過程に関する知見を得る．

まず，EC サイトの多種多様な商品類型に対する顧客の商品選択行動の異質性に着目し，潜在クラスモデル [26, 39] を既存手法 [32] に適用した潜在クラス形状制約モデルを提案する．次に，閲覧履歴を最新度と頻度の二つの特徴量に縮約するのではなく，より詳細な商品選択に至る過程に関する情報を利用可能とするために，各期間の商品閲覧数の時系列（閲覧数列）に着目し，閲覧数の時系列に基づく形状制約閲覧数列モデルを提案する．最後に，継続購買が想定される商品に対する商品選択行動に着目し，累積選択回数に対する再選択確率の単調性と凹性を仮定した形状制約比例ハザードモデルを提案する．

目次

図目次

表目次

第 1 章

序論

1.1 はじめに

本研究は，EC（Electronic Commerce : 電子商取引）サイトの閲覧履歴や商品選択履歴を含むページ遷移データや，実店舗の販売履歴を記録した POS（Point of Sales : 販売時点）データを用いた顧客の商品選択行動のモデル化と，商品選択予測の精度向上を対象とする．

本章では，その端緒として，まず 1.2 節で商品選択行動予測の研究背景と既存研究，1.3 節で研究目的，1.4 節で形状制約の妥当性に対する評価方法，最後に 1.5 節で第 2 章からの第 4 章の概要を述べる．

1.2 商品選択行動予測の研究背景と既存研究

1.2.1 顧客行動データ収集技術の発展と活用

近年の情報技術の発展に伴い, 多くの企業にて事業運営に関する大量かつ多様なデータを容易に取得できるようになった．インターネット上では，商品販売やサービスを提供する EC サイトが多くの企業により運営されている [68]．EC サイトの一つの特性として，閲覧履歴や商品選択履歴を含むページ遷移データが取得可能であることが挙げられる．顧客はページ遷移などの情報が企業から取得されていることを認識しながらも，EC サイトへ訪問しており [29]，このことは様々な種類の顧客のページ遷移データが取得可能であることを示している．実際に，企業は EC サイトで収集した情報を利用して，顧客にとって有益な情報を提供しており [51]，特に，ページ遷移データは顧客の商品選択行動を理解し，適切に情報提供するために重要である [9, 10, 29, 49]．顧客のウェブサイト上での行動が定量的に分析可能であることから，ページ遷移データは顧客の商品閲覧や誘導，インターネット広告，EC サイトでの購買行動など，様々な領域での研究対象となっている [10]．

またオフラインでも，店舗が顧客 ID 付きの POS データを収集することで，顧客がいつ来店し，どのような商品を購買したか，どのような顧客が再来店を行っているか，どのような商品が繰り返し購買されているか，といった情報の解析が可能である [9]．

これらのデータを利用して顧客の商品選択行動の推定が可能になれば，店舗で行われている

商品推薦や商品陳列，割引クーポンやポイント配布といったインセンティブプログラムなどのマーケティング施策による効果が増加する [10]．また単純にどの商品が選択されるかのみでなく，商品が選択されるまでの過程に関する知見を得ることも，これらの施策を行う実務的観点から重要である [30]．

これらのことを背景に，顧客の商品選択行動をモデル化し，商品選択を推定する研究が行われている．

1.2.2　ページ遷移データを用いた購買行動の予測

ページ遷移データを用いた研究で最も活発なもののひとつとして，顧客の EC サイトでのオンラインの購買行動分析が挙げられる [10]．Moe，Fader [48] は，閲覧履歴と購買履歴の観察をもとに，オンライン購買の成約率予測のための確率モデルを提案した．また，様々な特徴量をもとにロジットモデルやプロビットモデルを用いて購買行動を予測する研究も行われている [50, 55, 62, 63, 69]．Boroujerdi ら [5] は，顧客の購買の嗜好を予測するため，いくつかの分類モデルを応用している．しかしこれらの研究では，商品購買に至る顧客の訪問を予測することに焦点をあてており，それぞれの商品に対して選択確率を推定するということは行われていない．

オンラインでの商品選択行動を解析している研究は多くあるが，ページ遷移データのような単純な形式のデータではなく，多くの特徴量を含むデータ [12]，ソーシャルメディアのユーザプロファイル [74]，商品のレビューや評価 [58] を対象としているものが多い．

これらの多様な特徴量を考慮する研究とは異なり，本研究では，ページ遷移データから得られる顧客の商品閲覧や，POS データから得られる過去の商品選択といった顧客行動と，商品選択の関係を明らかにすることにある．顧客の選択行動の解析は，オンライン上の EC サイトや，オフラインで商品を販売する店舗にとって重要な課題であるため，本研究は多くの事業者にとって価値がある．

1.2.3　形状制約回帰モデル

統計的な推定モデルはパラメトリックな手法とノンパラメトリックな手法の 2 つに大別できる．パラメトリックな手法では，推定モデルが特定の関数型に従うことを仮定する．一方で，ノンパラメトリックな手法では推定モデルが特定の関数型に従うことを仮定しないため，顧客の商品閲覧と商品選択確率の関係を柔軟にモデル化できることが利点である．

形状制約回帰 [7, 24, 27] は，ノンパラメトリックな回帰モデルであり，特定の関数型を仮定せず，単調性，凸性，凹性などの形状制約のもとで，与えられたデータへの尤度を最大化，または誤差を最小化するようにパラメータを推定する．

代表的な形状制約回帰の応用例としては，経済学における利用率，生産性，費用，利得の関数 [23, 66] や，金融におけるオプション価格関数の推定 [2] が挙げられる．また，統計やオペレーションズ・リサーチ，画像処理の分野にも応用がある [1, 28, 57]．そして，形状制約回帰の

ための様々なアルゴリズムが提案されている [8, 13, 22, 45, 47, 71]．また単調回帰 (Isotonic regression) は形状制約回帰の特別な例である [3]．

岩永ら [32] は，RFM 分析 [31] の指標としても知られる最新度 (Recency) と頻度 (Frequency) という二つの指標に関して，単調性，凸性，凹性を制約条件として課して商品選択確率を推定する形状制約モデルを提案した．この形状制約モデルは，ページ遷移データの分析に対する形状制約回帰の応用として新しい，有効な手法である．

1.2.4　2 次元確率表モデル

本項では，本研究と深い関係をもつ 2 次元確率表モデル [32] について説明する．

顧客の商品選択行動分析である RFM 分析 [31] の最新度と頻度の指標は，顧客が過去に購買した商品から再度購買される商品を予測するための重要な指標であることが実証されている [18, 19, 34, 60, 61]．2 次元確率表モデルは，最新度と頻度の指標に基づいて，顧客の EC サイトにおける閲覧商品に対して選択確率を推定するために提案された手法である．

確率表

商品閲覧の最新度と頻度は，顧客と商品の組に対して定義される．概略を述べると，顧客 u の商品 v に対する「最新度」は，顧客 u が商品 v のウェブページへの訪れた時間の直近さを表し，「頻度」は顧客 u が商品 v のウェブページへどの程度の回数，時間訪れたかを表す．

$I = \{1, 2, 3, \ldots\}$ と $J = \{1, 2, 3, \ldots\}$ をそれぞれ最新度と頻度を表す有限な正の整数の集合とする．このとき，2 次元確率表は以下の行列で表される:

$$X = (x_{ij})_{(i,j)\in I\times J} \in [0,1]^{I\times J}.$$

ここで，x_{ij} は最新度が i で頻度が j の顧客と商品の組について，再閲覧や購買などの行動により商品が顧客から選択される確率を表す．既存研究 [32] では，顧客の商品選択行動をこの確率表をもとに解析している．

過去の閲覧履歴と商品選択履歴から，それぞれの最新度と頻度の組 $(i, j) \in I \times J$ に対して閲覧数と商品選択数を集計し，商品選択確率の過去実績値である実績確率表を作成することは容易である．しかし， 集計に用いるデータ量が少ない場合，商品選択確率 x_{ij} に対する信頼性は低くなる．この問題に対して，岩永ら [32] は，顧客の商品閲覧の最新度と頻度の性質を制約条件とした最適化問題を解くことにより商品選択確率を推定する方法を提案した．

最適化モデル

まず，過去の閲覧履歴にて基準となる日を設定する．基準日において、最新度と頻度の組 $(i, j) \in I \times J$ に該当する顧客と商品の組 (u, v) の標本サイズ n_{ij}，商品選択数 q_{ij} を集計する．

確率表 X について，事象 $E := ((n_{ij}, q_{ij}); (i, j) \in I \times J)$ の発生確率は二項分布として以下

のように表現される：

$$f(E;X) := \prod_{(i,j)\in I\times J} \binom{n_{ij}}{q_{ij}} (x_{ij})^{q_{ij}} (1-x_{ij})^{n_{ij}-q_{ij}}. \tag{1.1}$$

最適化モデルでは，定数項を除いて対数尤度 $\log(f(E;X))$ を最大化する．

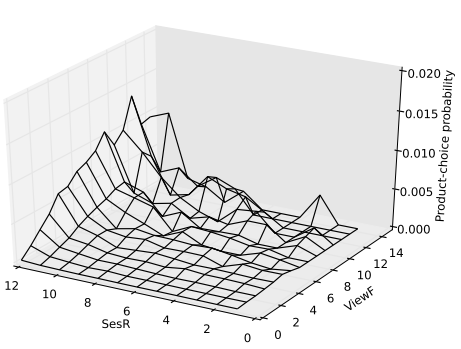
顧客の特定の商品に対する関心の増加にともない，商品選択確率の最新度と頻度は大きくなることが期待される．つまり，そのような最新度と頻度に関する単調性のもとで，単調性 (Monotonicity; Mono) モデル [32] は確率表 X を推定する：

$$\left|\begin{array}{ll} \text{maximize} & \displaystyle\sum_{(i,j)\in I\times J} (q_{ij}\log x_{ij} + (n_{ij}-q_{ij})\log(1-x_{ij})) \\ \text{subject to} & x_{ij} \le x_{i+1,j} \quad ((i,j)\in I\times J,\ i \le |I|-1), \\ & x_{ij} \le x_{i,j+1} \quad ((i,j)\in I\times J,\ j \le |J|-1), \\ & 0 < x_{ij} < 1 \quad ((i,j)\in I\times J). \end{array}\right. \tag{1.2}$$

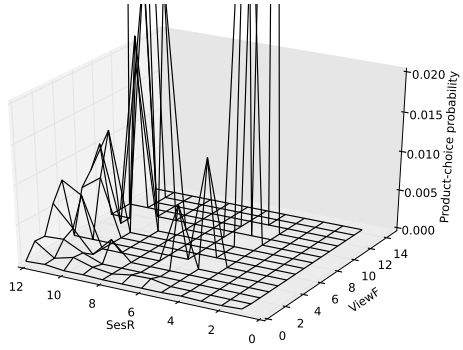
加えて，顧客の最終閲覧から時間が経過するにつれて，商品選択確率は小さくなるが，その減分は小さくなっていく．つまり，最新度が大きくなるにつれて，商品選択確率の増分も大きくなる．また，顧客の閲覧数が増加するにつれて，閲覧あたりの増分効果は小さくなっていく．つまり，頻度が大きくなるにつれて，商品選択確率の増分も小さくなる．単調性＋凸性＋凹性制約 (Monotonicity-convexity-concavity; MCC) モデル [32] は，単調性に加えて，それらの性質を凸性＋凹性制約として課す：

$$\left|\begin{array}{ll} \text{maximize} & \displaystyle\sum_{(i,j)\in I\times J} (q_{ij}\log x_{ij} + (n_{ij}-q_{ij})\log(1-x_{ij})) \\ \text{subject to} & x_{ij} \le x_{i+1,j} \quad ((i,j)\in I\times J,\ i \le |I|-1), \\ & x_{ij} \le x_{i,j+1} \quad ((i,j)\in I\times J,\ j \le |J|-1), \\ & x_{i+1,j} - x_{i,j} \le x_{i+2,j} - x_{i+1,j} \quad ((i,j)\in I\times J,\ i \le |I|-2), \\ & x_{i,j+1} - x_{ij} \ge x_{i,j+2} - x_{i,j+1} \quad ((i,j)\in I\times J,\ j \le |J|-2), \\ & 0 < x_{ij} < 1 \quad ((i,j)\in I\times J). \end{array}\right. \tag{1.3}$$

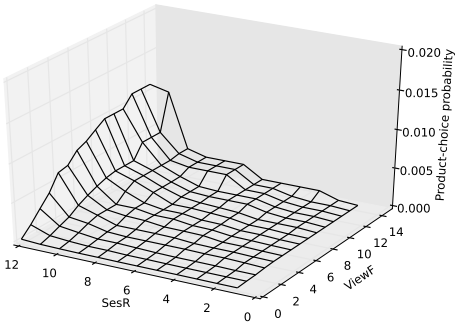
図 1.1 は，ある期間のファッション EC サイトにおける閲覧履歴と購買履歴をもとに推定された確率表を可視化した図である．最新度と頻度の定義は，それぞれ最終閲覧からの経過セッションの近さ (SesR)，閲覧回数 (ViewF) である．図 1.1 では，データ量に対する推定確率値の差異を比較をするため，顧客と商品の組のデータをすべて利用したもの（全標本）と，1% に標本抽出を行ったもの（1% 標本）を示している．実績確率表 (EMP) の図 1.1 (a),(b) では，確率表の標本サイズの小さい部分で過剰適合し凹凸が発生しているが，式 (1.2) によって図 1.1 (c),(d) のように単調性が満たされた確率表 (Mono) が推定され，式 (1.3) により図 1.1 (e),(f) のように，さらに凸性＋凹性が満たさた確率表 (MCC) が推定されていることが観察できる．また，実績値に基づく確率表では，過剰適合による凹凸は標本抽出率が 1% のときにより顕著であるが，そのような場合でも単調性と凸性＋凹性の制約条件を課すことで，全標本の確率表と形状が近くなることが観察できる．



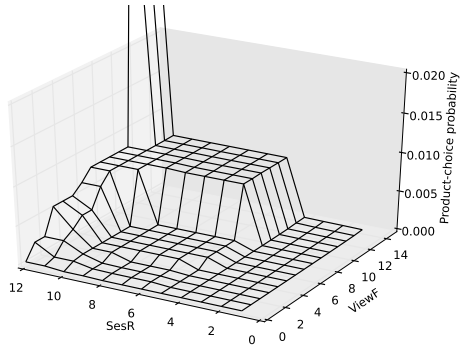
(a) EMP, 全標本



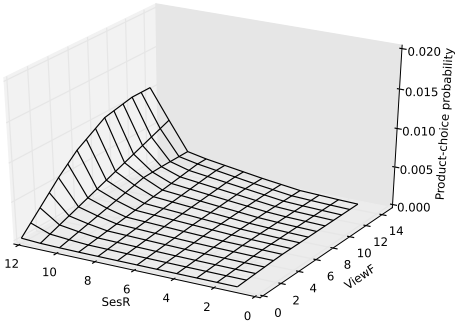
(b) EMP, 1% 標本



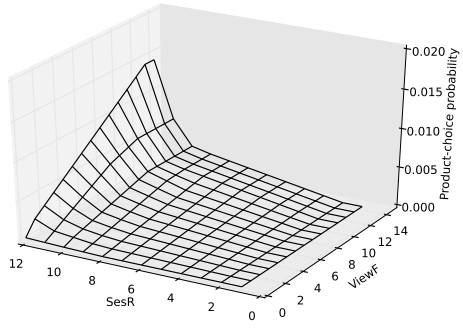
(c) Mono, 全標本



(d) Mono, 1% 標本



(e) MCC, 全標本



(f) MCC, 1% 標本

図 1.1. 実績確率表 (EMP)，単調性に基づく確率表 (Mono)，単調性＋凸性＋凹性制約に基づく確率表 (MCC) のデータ量を変化させた場合の形状の可視化

これらの制約条件を課した確率表を利用することで，同じ最新度と頻度の特徴量を利用したサポートベクターマシンとロジスティック回帰よりも商品選択に対する予測精度が向上することが報告されている [32]．またこれらの推定された確率表から，商品閲覧の回数に対する商品選択確率の増加度や，商品閲覧の無いセッション数に対する商品選択確率の減少度に関しての示唆を得ることができる．このように推定された確率表から，商品選択確率が低下した顧客に対して割引ポイントのインセンティブを提供することで，顧客離反を抑制する施策が実際に行われている [75]．

1.3　研究目的

本研究の目的は，1.2 節で述べたような，実務的観点での問題を解決することである．その手段として，商品選択行動の推定を行い，商品が選択されるまでの過程に関する知見を得る．この取り組みは，Lynd [42] によると，実践的な問題解決のための科学の適用である工学的取り組み (Engineering Endeavor) である．

具体的な取り組みとして，岩永らのアプローチと同様に，まず顧客の商品選択行動の解析のため，顧客の実際の商品選択行動を記録したデータを解析する．そして，顧客の商品選択に影響を与える特徴量を探索する．その特徴量の観察から，顧客の商品選択に関する新たな性質を仮定し，仮定を形状制約として課してモデルのパラメータを推定することで，選択予測モデルの精度の向上を試みる．さらに，実際のデータから推定されたモデルのパラメータを観察することで，顧客の商品選択の過程に関する知見を得る．

1.4　形状制約の妥当性に対する評価

本節では，以降の章で説明する提案手法での形状制約の妥当性の評価について述べる．

Burke [11] は，定量的な研究の妥当性の評価方法として，内的妥当性 (Internal validity) と外的妥当性 (External validity) を評価することを挙げている．

内的妥当性とは，観察した関係に因果関係があると判断できる確度を表しており，因果関係が強いほど内的妥当性が高い．本研究では，各章で提案する形状制約に対する内的妥当性の検証のため，顧客の商品選択行動を記録したデータを用意して数値実験を行った．数値実験では，まずデータを学習データとテストデータに分割する．そして，学習データを利用して，形状制約を課してパラメータを推定したモデル，形状制約を課さずにパラメータを推定したモデルを作成し，それぞれについてテストデータに対する商品選択の予測精度を検証する．この検証において，形状制約モデルが形状制約を課さないモデルと比較して予測精度が高い場合，そのデータにおいては，形状制約の内的妥当性が高いと評価する．

外的妥当性とは，研究から得られた結果が他の対象，状況，時間に対しても一般化可能な程度を表しており，一般化できる程度が強いほど外的妥当性が高い．本研究では，各章で提案する形状制約に対して，それぞれ顧客の商品選択行動を記録した 1 種類のデータについてのみ数値実験を行ったため，どのような対象，状況，時間の商品選択行動を記録したデータに対して

も必ず提案手法の形状制約を課すことで予測精度が向上するとは言い難い．Burke [11] は，完全に無作為なデータを用意し検証することが任意の対象，状況，時間に対する一般化を示す方法だが，すべての対象，状況，時間を考慮して完全に無作為なデータを用意することが不可能なため，数値実験をもとにした定量的研究では得ることのできる結果の外的妥当性は実質的に常に低い，と述べている．Stake [64] は，大まかな一般化の方法として，ある結果を示した研究と類似性の高い対象や状況であるほど，研究の結果の対象や状況への一般化を強く正当化できる，と述べている．これらの議論から，本研究では形状制約の外的妥当性を高めるため，まず数値実験で利用するデータの統計量や商品閲覧に関する特徴量の情報を記述する．またデータに対して複数の形状制約モデルからどのモデルを適用するのがよいかについて評価するため， 5.2 節にて各章で提案する形状制約モデルについての整理とモデル選択の指針を示す．数値実験に利用したデータの情報や形状制約モデルの選択の指針を提示することで，新たなデータに対して形状制約モデルを利用することの妥当性をある程度評価できると考える．

1.5 各章の構成

本節では，第 2 章から第 4 章の各章について，その内容と主要な結果を述べることにより論文全体の構成を説明する．なお，第 2 章「商品類型の異質性に基づく潜在クラス確率表モデル」は学術誌論文 [52] として掲載済み，第 3 章「商品閲覧数の時系列に基づく形状制約閲覧数列モデル」の一部は国際会議論文 [53] として報告済み，第 4 章「反復選択商品に対する形状制約比例ハザードモデル」は学術誌論文 [78] として掲載済みである．

1.5.1 第 2 章の概要

第 2 章では，EC サイトの多種多様な商品に対する顧客の商品選択行動の異質性に着目する．顧客の商品に対する選択行動は，商品の種類や価格帯によって異なるといわれている [4]．このような商品の異質性を考慮することで，既存研究 [32] よりも商品選択をより高精度に予測できる可能性がある．

そこで第 2 章では，潜在クラスモデル [26, 39] を 2 次元確率表モデルに適用した潜在クラス形状制約モデルを提案する．提案手法では，まず商品の各潜在クラスに対する帰属度を算出し，各潜在クラスの確率表をその商品の帰属度で重み付けして用いることで，顧客の商品に対する商品選択確率を推定する．そして，提案モデルのパラメータを推定するための EM アルゴリズム [15, 46] を提案する．また推定された各潜在クラスの確率表をもとに，顧客の各商品類型に対する選択行動の差異を分析する．

第 2 章での貢献は以下の通りである：

- 商品選択確率を推定するための潜在クラスに基づく新しい手法を提案し，そのパラメータを推定するための EM アルゴリズムを提案した．
- 数値実験により，提案手法が既存手法の MCC モデル [32] や潜在クラスロジスティック回帰モデル [21, 25, 35] よりも高い予測精度であることを検証した．

- 推定された商品選択確率をもとに，各潜在クラスに属する商品類型に対する顧客の商品選択行動を観察した．

第 2 章の構成は以下の通りである．2.2 節では，潜在クラス回帰の関連研究と，顧客の商品類型に対する商品選択行動の異質性に関する仮説について述べる．2.3 節では，提案手法である潜在クラス確率表モデルのパラメータ推定に利用する EM アルゴリズムについて説明する．2.4 節では提案手法の有効性を数値実験によって検証する．最後に，2.5 節では第 2 章のまとめと課題を述べる．

1.5.2　第 3 章の概要

第 3 章では，EC サイトにおける顧客の商品閲覧数の時系列と商品選択の関係に着目する．既存研究 [32] では，商品閲覧履歴から各顧客の閲覧商品に対する「最新度」と「頻度」を数量化し，形状制約のもとで最新度と頻度の組に対して閲覧商品が購買される確率を推定する手法が提案された．しかし，この手法では顧客の閲覧履歴が最新度と頻度の 2 次元に縮約されるため，閲覧履歴に関する多くの情報が失われてしまう．そこで，閲覧履歴を二つの特徴量に縮約するのではなく，各期間の商品閲覧数の時系列（閲覧数列）を利用することで，商品選択に至る過程のより詳細な情報が利用可能になり，商品選択を高精度に予測できる可能性がある．

そこで第 3 章では，顧客の商品に対する閲覧数列に対して商品選択確率を推定する閲覧数列モデルを提案する．また，各閲覧数列に対する選択確率を推定するための学習データの不足により生じる過剰適合を抑制するために，閲覧数列の順序関係に基づく推定値の補正方法を提案する．その際，単純に閲覧数列の順序関係を全て列挙すると，その数は膨大となってしまう．閲覧数列の順序関係については，ハッセ図で隣接する閲覧数列の組のみを考慮すれば十分であり，これにより計算の効率化が可能になる．数値実験によって冗長な閲覧数列の順序関係を取り除くことによる計算効率化の効果と，提案手法の商品選択予測における予測精度を検証する．

第 3 章での貢献は以下の通りである：

- 各期間の商品閲覧数の時系列（閲覧数列）から商品選択確率を推定する手法を提案し，過剰適合を抑制するための閲覧数列の順序関係に基づく推定値の補正方法を提案した．
- 商品選択確率を推定する際の冗長な順序関係をハッセ図を考慮して取り除く方法を提案した．
- 数値実験により，提案手法が既存手法の MCC モデル [32] や代表的な機械学習モデルよりも高い予測精度であることを検証した．
- 推定された各閲覧数列の商品選択確率をもとに，顧客の閲覧数列と商品選択の関係を考察した．

第 3 章の構成は以下の通りである．3.2 節では，各期間の商品閲覧数の時系列である閲覧数列について定義し，観察に基づき，商品選択確率に関して閲覧数列に期待される順序関係につ

いて定義する．3.3 節では，閲覧数列の順序関係に基づき商品選択確率を推定する閲覧数列モデルの最適化問題としての定式化と，閲覧数列モデルの推定を効率的に行うための冗長な制約条件の削除方法について述べる．3.4 節では，閲覧数列モデルの商品選択確率の予測精度の評価と，冗長な制約条件の削除の効果について評価する．最後に，3.5 節で第 3 章のまとめと課題を述べる．

1.5.3 第 4 章の概要

第 4 章では，継続購買が想定される商品に対する商品選択行動に着目する．顧客の商品選択行動の性質として，一度選択すると別の商品に切り替える場合に新たな探索や，その商品への習熟のためのコストが生じるため，次回の選択機会に顧客が同じ商品を選択する可能性が高くなる，というものがある [37]．このような顧客が過去に選択した商品を反復して選択する性質を利用することで，商品選択確率を高精度で予測できる可能性がある．

本章では特に，新たな店舗を探索するコストや，ヘアスタイルの好みや髪質などを説明するコストが生じるために，同じ店舗に継続して来店する可能性が高いヘアサロンの再来店の予測を対象とする．分析の対象としたヘアサロンでは売上構成において大部分を再来店顧客が占めるため，顧客の再来店確率を予測することが安定した店舗経営に繋がるという実務的な利点もある．また，各店舗の再来店顧客の来店状況を定量的に把握することが可能になれば，再来店率の高い店舗と低い店舗の差異を捉え，店舗が継続顧客を獲得するための知見を得ることができる．

再来店予測の方法として，累積来店回数に対する再来店率の単調性と凹性を考慮した形状制約比例ハザードモデルを提案する．また，推定されたモデルのパラメータをもとに，各店舗における顧客の再来店状況の推移を分析する．

第 4 章での貢献は以下の通りである：

- 累積来店回数と再来店率に関する単調性と凹性を考慮した形状制約比例ハザードモデルを提案した．
- 数値実験により，形状制約を課すことで，比例ハザードモデルの予測性能が向上し, 特に来客数が少ない小規模店舗でその効果が大きいという結果を得た．
- 比例ハザードモデルにより推定したパラメータを主成分分析により縮約し可視化することで，店舗における再来店率改善のための示唆を得た．

第 4 章の構成は以下の通りである．4.2 節では，再来店予測に利用する比例ハザードモデルの関連研究と，顧客の再来店の選択行動に関する仮説について述べる．4.3 節では，ヘアサロンの顧客の再来店を，累積来店回数に対する再来店率の単調性と凹性を考慮した比例ハザードモデルにより推定する方法を提案する．4.4 節では数値実験を通して，提案手法の有効性に関して検証する．また提案手法によって得られたモデルのパラメータを主成分分析により縮約し可視化することで，各店舗の来店状況を分析する．最後に 4.5 節で第 4 章のまとめと課題を述べる．

第 2 章

商品類型の異質性に基づく潜在クラス形状制約モデル

2.1 はじめに

本章では，EC サイトの多種多様な商品に対する顧客の商品選択行動の異質性に着目する．顧客の商品に対する選択行動は，商品の種類や価格帯によって異なるといわれている [4]．このような商品の異質性を考慮することで，既存研究 [32] よりも商品選択をより高精度に予想できる可能性がある．そこで本章では，潜在クラスモデル [26, 39] を 2 次元確率表モデルに適用した潜在クラス形状制約モデルを提案する．

本章の構成は以下の通りである．2.2 節では，潜在クラス回帰の関連研究と，顧客の商品類型に対する商品選択行動の異質性に関する仮説について述べる．2.3 節では，提案手法である潜在クラス確率表モデルのパラメータ推定に利用する EM アルゴリズムについて説明する．2.4 節では提案手法の有効性を数値実験によって検証する．最後に，2.5 節では本章のまとめと課題を述べる．

2.2 潜在クラス回帰

潜在クラス回帰（または混合クラス回帰）[26, 72] は潜在クラスモデルのひとつである [26, 39]．潜在クラス回帰では，学習に利用する標本を潜在クラスに分類し，同時にそれぞれの潜在クラスに対して回帰モデルを推定する．潜在クラス回帰モデルの尤度関数の最大化のためのアルゴリズムとして，Newton-Raphson 法 [17] や EM アルゴリズム [15, 46] が知られている．

潜在クラスモデルはマーケティングの分野で顧客の異質性を表現するために利用される．本章では，潜在クラスモデルを EC サイトの多種多様な商品の分類に利用することを考える．商品の異質性を考慮する動機として，例えば，水やペットフードのような日常的に消費される商品と，ファッションや家具のような，一度購買すると一定期間利用する商品について考える．水やペットフードは，いつも購買している銘柄があったり，価格が安価なため，あまり比較検

討されずに購買に至ることが考えられる．一方で，ファッションや家具は，日常的に消費されるものに比べて高価なため，複数の商品間で比較検討されたのちに購買に至ることが考えられる．このような商品選択に至る過程が異なる商品については，既存研究 [32] で利用された閲覧に関する特徴量である最新度や頻度が同じであったとしても，商品選択確率が異なることが予想される．具体的には，水やペットフードは少ない閲覧回数で購買に至る可能性が高く，頻度が小さくても他の類型に比べて選択確率が高いと考えられる．ファッションや家具は，時間をかけて比較検討されたのちに購買に至る可能性が高く，最新度が大きい場合と小さい場合の差が，他の類型に比べて小さいと考えられる．

このような商品類型ごとの選択確率の差異を捉える方法として，商品類型ごとに異なるモデルを推定することが考えられる．しかし，素朴にすべての類型で異なるモデルを作成してしまうと，それぞれのモデルに割り当てられる標本サイズが少なくなってしまい，小さな標本サイズのデータに対して過剰適合が生じる可能性がある．

そこで，標本サイズの小さいデータに対する過剰適合抑制のため，すべての類型で異なるモデルを作成するのではなく，顧客の閲覧に関する特徴をもとに複数の類型を潜在クラスとしてまとめ，潜在クラスに対してモデルを推定することを考える．

2.3 潜在クラス確率表モデル

本節では，商品選択確率を推定するための提案手法である潜在クラス確率表モデルについて述べる．また提案手法のパラメータを推定するための EM アルゴリズムについて説明する．

2.3.1 定式化

商品類型の集合を K とし，最新度 i，頻度 j，商品類型 k の組 $(i,j,k) \in I \times J \times K$ を考える．ある日を起点として組 (i,j,k) に該当する顧客と商品の組の標本サイズを n_{ijk} とし，q_{ijk} をそのなかで購買された標本サイズとする．

これらの商品類型を考慮するための単純な方法は，類型ごとに異なる確率表を作成することである．しかし，類型ごとに確率表を作成したとき，標本サイズの小さい類型については，確率値が信頼できる値とならない．本章では，この問題を解決するため，潜在クラス回帰 [26, 72] と同様に，類似の商品類型を集約することを考える．

特に，$|S| < |K|$ のもとで潜在クラスの集合 S を導入し，クラス $s \in S$ の構成割合を π_s とすると，以下の関係を満たす：

$$\sum_{s \in S} \pi_s = 1 \quad \text{かつ} \quad \pi_s > 0 \quad (s \in S). \tag{2.1}$$

各クラス $s \in S$ の 2 次元確率表は以下のように定義される：

$$X_s = (x_{ijs})_{(i,j) \in I \times J} \in [0,1]^{I \times J}.$$

ここで，x_{ijs} はクラス s の確率表における，最新度と頻度の値が i と j の商品に対する顧客の商品選択確率を示す．

式 (1.1) と同様に，事象 $E_k := ((n_{ijk}, q_{ijk}); (i,j) \in I \times J)$ の条件付き確率は以下のようにかける：

$$f(E_k; X_s) = \prod_{(i,j)\in I\times J} \binom{n_{ijk}}{q_{ijk}} (x_{ijs})^{q_{ijk}} (1 - x_{ijs})^{n_{ijk}-q_{ijk}}.$$

これらの確率値を重み π_s で混合することで，事象 E_k の確率は以下のようになる：

$$\sum_{s\in S} \pi_s f(E_k; X_s).$$

類型 k がクラス s に属する事後確率はベイズの定理により以下のように与えられる：

$$\frac{\pi_s f(E_k; X_s)}{\sum_{s\in S} \pi_s f(E_k; X_s)}. \tag{2.2}$$

2.3.2　EM アルゴリズム

本項では潜在クラス確率表モデルのパラメータを推定するための EM アルゴリズムを説明する．まず，類型 k がクラス s に属するとき $z_{ks} = 1$，そうでないとき $z_{ks} = 0$ となる帰属変数 $z_{ks} \in [0,1]$ を導入する．そして，完全データの対数尤度関数は以下のように表される：

$$\begin{aligned} &\log \left(\prod_{(k,s)\in K\times S} (\pi_s f(E_k; X_s))^{z_{ks}} \right) \\ &= \sum_{(k,s)\in K\times S} z_{ks} \log f(E_k; X_s) + \sum_{(k,s)\in K\times S} z_{ks} \log \pi_s. \end{aligned} \tag{2.3}$$

EM アルゴリズムは $\hat{z}_{ks}$ を帰属変数の初期推定値として開始する．そして対数尤度関数 (2.3) を最大化するために，E ステップ（Expectation ステップ）と M ステップ（Maximization ステップ）を繰り返す．

M ステップ

M ステップでは $\hat{z}_{ks}$ を対数尤度関数 (2.3) に代入し，他の変数について最大化する．特に，潜在クラスのサイズは制約条件 (2.1) のもとで $\sum_{(k,s)\in K\times S} \hat{z}_{ks} \log \pi_s$ を最大化することで決定する．ラグランジュの未定乗数法により，$s \in S$ について以下を算出する：

$$\hat{\pi}_s \leftarrow \frac{\sum_{k\in K} \hat{z}_{ks}}{|K|}. \tag{2.4}$$

商品選択確率は $\sum_{(k,s)\in K\times S} \hat{z}_{ks} \log f(E_k; X_s)$ を単調性，凸性，凹性のもとで最大化することで決定される．この問題の最適解 $\hat{X}_s = (\hat{x}_{ijs})_{(i,j)\in I\times J}$ は，それぞれの $s \in S$ について分割し

た以下の問題を解いて得ることができる：

$$
\left|\begin{array}{ll}
\text{maximize} & \displaystyle\sum_{(i,j,k)\in I\times J\times K} \hat{z}_{ks}\Big(q_{ijk}\log x_{ijs} + (n_{ijk}-q_{ijk})\log(1-x_{ijs})\Big) \\
\text{subject to} & x_{ijs} \le x_{i+1,j,s} \quad ((i,j)\in I\times J,\ i\le |I|-1), \\
& x_{ijs} \le x_{i,j+1,s} \quad ((i,j)\in I\times J,\ j\le |J|-1), \\
& x_{i+1,j,s} - x_{ijs} \le x_{i+2,j,s} - x_{i+1,j,s} \quad ((i,j)\in I\times J,\ i\le |I|-2), \\
& x_{i,j+1,s} - x_{ijs} \ge x_{i,j+2,s} - x_{i,j+1,s} \quad ((i,j)\in I\times J,\ j\le |J|-2), \\
& 0 < x_{ijs} < 1 \quad ((i,j)\in I\times J).
\end{array}\right. \tag{2.5}
$$

この最適化問題は問題 (1.3) と同様に，線形の制約条件のもとでの凹関数の最大化である．したがって，一般的な非線形最適化ソフトウェアにより最適解を得ることができる．

E ステップ

E ステップでは，他の変数を固定して各帰属変数の期待値を算出する．事後確率 (2.2) は，$(k,s)\in K\times S$ について，以下の式で算出される：

$$
\hat{z}_{ks} \leftarrow \frac{\hat{\pi}_s f(E_k;\hat{X}_s)}{\sum_{s\in S}\hat{\pi}_s f(E_k;\hat{X}_s)} \tag{2.6}
$$

EM アルゴリズムによるパラメータ推定

これら E ステップ，M ステップを終了条件を満たすまで繰り返す．提案手法の EM アルゴリズムによる潜在クラス単調性＋凸性＋凹性モデルのパラメータ推定は，以下のステップで行う：

Step 0 (初期化) $(k,s)\in K\times S$ について $\hat{z}_{ks}$ を初期値として設定する．Step 2 へすすむ．

Step 1 (E ステップ) $(k,s)\in K\times S$ について $\hat{z}_{ks}$ を式 (2.6) に基づき更新する．

Step 2 (M ステップ) $s\in S$ について式 (2.4) をもとに $\hat{\pi}_s$ を更新する．$s\in S$ について問題 (2.5) を解き $\hat{X}_s$ を更新する．

Step 3 (終了条件) 終了条件を満たしているならばアルゴリズムを停止する．そうでなければ，Step 1 へ戻る．

2.4 数値実験

本節では，商品選択確率を推定するための提案手法である潜在クラスモデルの有効性について評価する．

2.4.1 ページ遷移データ

実験は，EC サイトを運営する株式会社リクルートライフスタイル[*1]より提供されたページ遷移データを用いた．ページ遷移データは 2015 年 8 月から 10 月の期間で収集されたもの

[*1] http://www.recruit-lifestyle.co.jp/

を利用した．対象とした EC サイトでは，上記の期間で 100 商品以上購買された商品類型が 2.4.7 節で示す 56 種類あり，本章ではこの 56 種類の類型の商品について扱う．各データは商品ページの閲覧か商品の購買を表し，時間，顧客 ID，商品 ID などの情報が含まれている．

図 2.1 は，無作為抽出した 50,000 人の顧客について，商品ページ閲覧回数と購買数の関係を示した散布図である．図から，10 商品以上購買する顧客は少ないが，一方で商品を購買しないにもかかわらず多くの商品を閲覧する顧客が存在することがわかる．

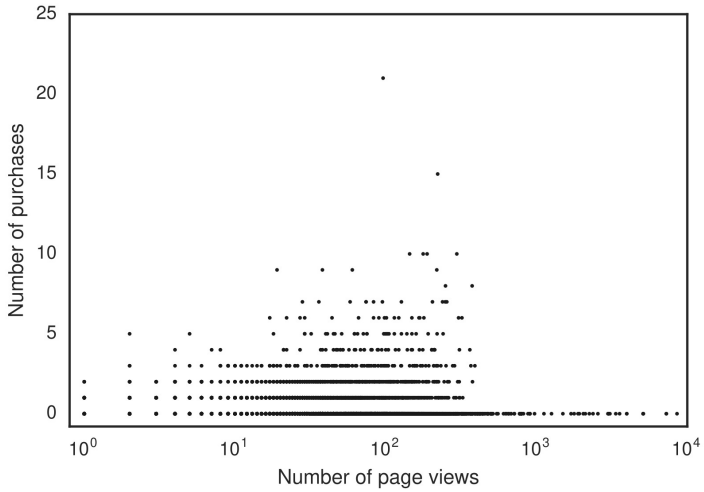


図 2.1. 無作為抽出した顧客 50,000 人の閲覧数と購買数の散布図

本実験では，ページ閲覧回数，ページ閲覧セッション数（セッション数），ページ閲覧日数（日数）をもとにした，以下の 6 つの最新度，頻度の組合せの特徴量を用意した：

ViewR　ページ閲覧回数をもとに算出した最新度 ($|I| = 24$)
SesR　セッション数をもとに算出した最新度 ($|I| = 12$)
DayR　日数をもとに算出した最新度 ($|I| = 24$)
ViewF　ページ閲覧回数をもとに算出した頻度 ($|J| = 16$)
SesF　セッション数をもとに算出した頻度 ($|J| = 8$)
DayF　日数をもとに算出した頻度 ($|J| = 8$)

例えば，m (≥ 1) 日前に最終閲覧をした顧客と商品の組については，DayR の値は $i = \max\{25 - m, 1\}$ とする．n (≥ 1) 回閲覧をした顧客と商品の組については，ViewF の値は $j = \min\{n, 16\}$ とする．これらの最新度，頻度の値は基準日より前の 4 週間のページ遷移データから算出される．また本実験において，閾値 $|I|$，$|J|$ の設定は，最新度，頻度の値について，閾値以上となる標本サイズが全標本サイズの 5% 以下となるように設定した．

2.4.2 比較モデル

数値実験では，6 つの商品選択確率の推定手法の性能を比較した．実験結果では，各モデルについて表 2.1 の略称を用いる．

表 2.1. 比較モデルの略称

略称	モデル
MCC(1)	単調性＋凸性＋凹性モデル (1.3)：全ての商品に対して 1 個の確率表を作成.
MCC(56)	単調性＋凸性＋凹性モデル (1.3)：56 個の確率表を商品類型ごとに作成.
LCMCC($\|S\|$)	潜在クラス単調性＋凸性＋凹性モデル：2.3 節を参照．$\|S\|$ は潜在クラスの数.
LCLR($\|S\|$)	潜在クラスロジスティック回帰：$\|S\|$ は潜在クラスの数.
RF	ランダムフォレスト
ANN	人工ニューラルネットワーク

LCLR($|S|$), RF, ANN では，特徴量として顧客と商品の組の最新度，頻度の 2 つの値を用いた．LCMCC($|S|$) のパラメータは 2.3.2 節で述べた EM アルゴリズムによって推定した．LCLR($|S|$) のパラメータは通常の EM アルゴリズム [26, 72] によって推定した．これらの EM アルゴリズムの終了条件は，対数尤度関数の増分が 0 となる，または反復回数が多くの場合でそれ以前に終了条件に至った 10 回となることと設定した．そして，この EM アルゴリズムをランダムにパラメータの初期値を変更しながら 10 回繰り返し，対数尤度関数が最大のものを最終的なモデルとして利用した．最適化問題 (1.3),(2.5) は株式会社 NTT データ数理システム[*2] の Numerical Optimizer v17 を利用して求解した．求解の際には，$\varepsilon = 10^{-5}$ として不等式 $0 < x_{ij} < 1$ を，$\varepsilon \leq x_{ij} \leq 1 - \varepsilon$ に置き換えて求解した．RF と ANN の計算においては，Python の機械学習ライブラリ scikit-learn[*3] (ver. 0.18.1) の RandomForestRegressor と MLPRegressor の関数をそれぞれ用いた．RF のハイパーパラメータである ‘max_depth’ は 5 分割交差確認法より 3 に設定した．計算には AMD Opteron 4133 CPU (2.80 GHz)，128 GB メモリを搭載した計算機を用いた．

学習段階：2 次元確率表の推定

まず最初のデータセットの基準日は 9 月 3 日に設定し，閲覧履歴と購買履歴から，全ての顧客と商品の組について最新度，頻度の値と購買の有無を集計する．同様にしてデータセットの基準日を 9 月 3 日から 30 日まで 1 日ずつ動かして集計し，28 種類のデータセットを作成した．学習データは顧客と商品の組を 1 要素として，購買の標本サイズは 43,566，非購買の標本サイズは 96,495,418 であった．学習データの標本サイズと予測精度の関係について検証するため，元の学習データから 1%，10% とランダムに抽出して作成した学習データを作成した．これらの学習データをそれぞれ 1% 標本 (1%-sampled)，10% 標本 (10%-sampled)，また元の

[*2] http://www.msi.co.jp
[*3] http://scikit-learn.org

学習データを全標本 (all-samples) とよぶ．2 次元確率表はこれらの学習データから推定する．

テスト段階：2 次元確率表の予測精度の評価

最初の基準日は 10 月 1 日として，商品選択確率は最新度，頻度の値を確率表で参照して求めた．そして各顧客に対して商品選択確率の降順に N 商品を選択する．ここで，商品選択確率が同じ場合には ViewR により順位づけした．4 週間で N 種類以下の商品しか閲覧していない場合には，すべての閲覧商品が選択される．この操作を 10 月 1 日から 10 月 28 日まで繰り返し，28 種類のテストデータについて検証する．これらのテストデータでは平均して，購買の標本サイズが約 1,214，非購買の標本サイズが約 3,316,018 がであった．

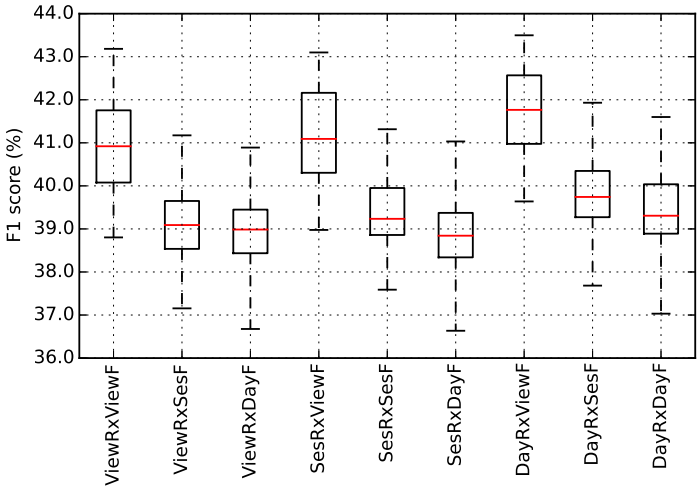
それぞれのテストデータで，各顧客に対して以下の指標を計算した：

$$\begin{aligned}
&\text{再現率 (Recall)} = \frac{\#(\text{予測され購買された商品})}{\#(\text{購買された商品})},\\
&\text{適合率 (Precision)} = \frac{\#(\text{予測され購買された商品})}{\#(\text{購買された商品})},\\
&\text{F1 値 (F1 score)} = \frac{2\cdot(\text{再現率})\cdot(\text{適合率})}{(\text{再現率})+(\text{適合率})}.
\end{aligned}$$

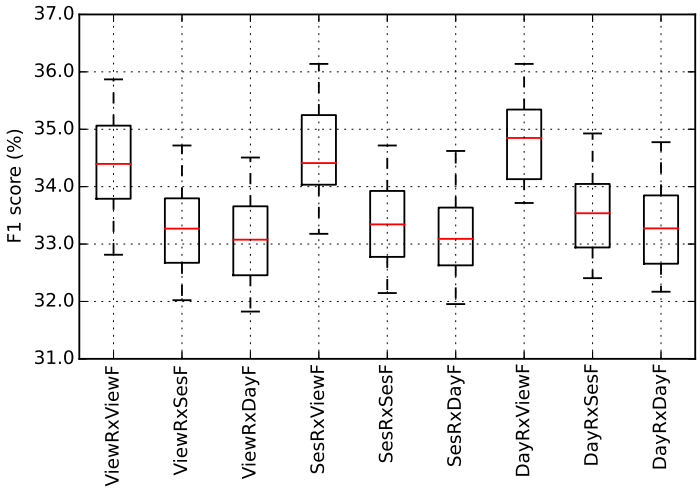
ここで $\#(\cdot)$ は商品数を表す．また購買予測の対象商品は，各顧客に過去 4 週で閲覧された商品のみを対象とする．それぞれの指標は 28 種類のテストデータについて平均した値とする．後の節では，平均値の 95% 信頼区間を図中に示す．

2.4.3　最新度と頻度の特徴量の組合せ

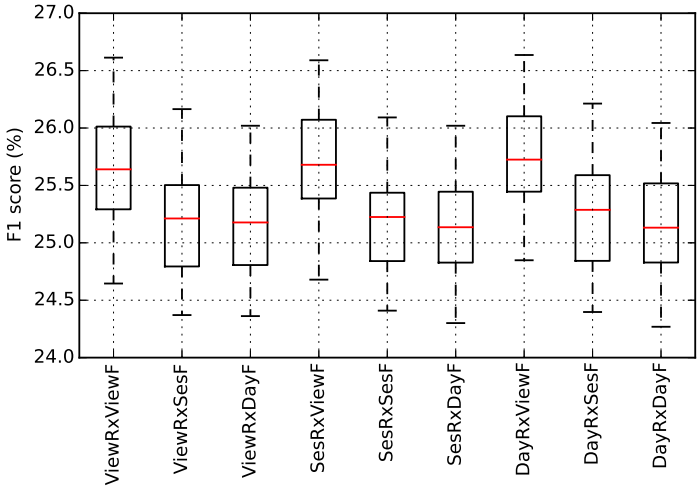
まず，商品選択確率の推定のために，最新度と頻度の組合せによる予測精度の変化について検証する．図 2.2 の箱ひげ図は，商品選択数を $N \in \{3, 5, 10\}$ としたときの 28 種類のテストデータに対する MCC(1) の F1 値の分布を表す．本実験では，最新度の特徴量 {ViewR, SesR, DayR} と頻度の特徴量 {ViewF, SesF, DayR} からなる 9 種類の組合せについて検証した．図 2.2 から，頻度について ViewF を利用することで F1 値が高く，$N \in \{3, 5, 10\}$ において DayR×ViewF の組合せが最良であることが確認できる．この結果から，後の数値実験では DayR×ViewF の特徴量の組合せを用いて検証する．



(a) $N = 3$



(b) $N = 5$



(c) $N = 10$

図 2.2. 最新度と頻度の特徴量の 9 種類の組合せに対する MCC(1) の F1 値

2.4.4　潜在クラス確率表モデルの有効性

本項では，MCC モデルの潜在クラスモデルへの拡張の有効性について検証する．図 2.3 [*4] は商品選択数を $N \in \{3, 5, 10\}$ として，学習データの標本抽出率を 1%，10%，全標本と変化させたときの MCC(1)，LCMCC(4)，LCMCC(16)，MCC(56)，RF，ANN の F1 値を示している．

MCC(1) は 1 つの確率表のみを用いるため，標本サイズが小さい場合にも過剰適合が生じづらい．このことから，学習データの標本サイズに依存して F1 値は大きく変化しない．一方で，MCC(56) は商品類型ごとに 56 種類の確率表を作成する．その結果，MCC(56) の予測精度は 1% 標本に対しては大きく悪化するが，標本サイズを増加させるにつれて改善する．RF はいずれの標本抽出率の学習データを用いたときも比較的高い予測精度を達成しているが，ANN は 1% 標本の学習データを利用した場合に予測精度が他のモデルと比較して大きく悪化している．

そして，4 種類の潜在クラス確率表を作成する LCMCC(4) では，1% 標本に対して最も良い予測精度を達成している．LCMCC(4) は 10% 標本に対して MCC(1) と MCC(56) よりも予測精度が良い．全標本に対しては，MCC(1) よりも予測精度が高いが，MCC(56) とほぼ同等となる．

16 種類の潜在クラス確率表を作成する LCMCC(16) では，1% 標本に対しては予測精度が低いが，全標本のときには，最も予測精度が高い．これらの結果は，提案手法である潜在クラス確率表は，56 種類の商品類型をより少ない潜在クラスに集約することで，既存手法である MCC モデルよりも予測性能を向上させることができることを示している．

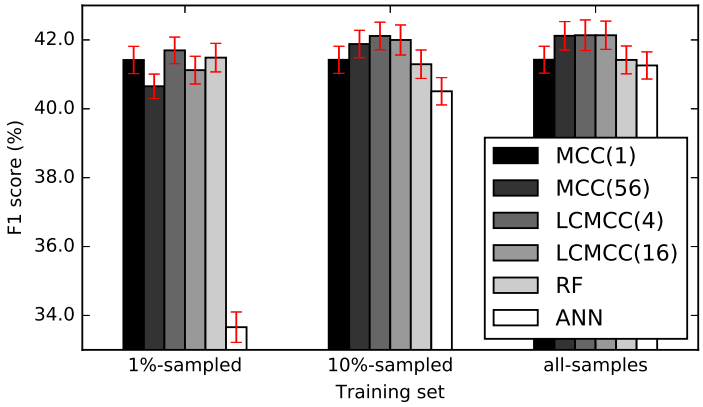
2.4.5　潜在クラスロジスティック回帰との比較

提案手法の潜在クラス確率表モデルと，既存手法の潜在クラスロジスティック回帰の予測精度の比較する．図 2.4 は 1% 標本，10% 標本，全標本と標本抽出率を変化させたときの LCMCC($|S|$) と LCLR($|S|$) の F1 値を示している．実験では商品選択数を $N \in \{3, 5, 10\}$ として，潜在クラスの数を $|S| \in \{1, 4, 8, 12, 16\}$ と設定して精度を比較した．

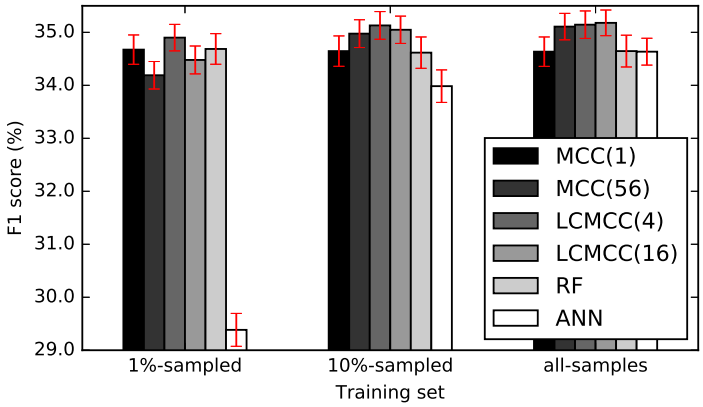
図 2.4 から，いずれの潜在クラス数 $|S|$ についても，LCMCC($|S|$) の予測精度が LCLR($|S|$) よりも高いことがわかる．以下では，それぞれの標本抽出率の学習データを用いた結果の詳細について確認する．

まず，1% 標本の結果に着目する．このとき，学習データの不足から潜在クラスの数 $|S|$ が多くなるにつれて LCLR($|S|$) の予測精度は悪化している，一方，LCMCC(4) と LCMCC(8) の F1 値は LCMCC(1) よりも向上している．言い換えると，LCMCC モデルは小さい標本サイズの学習データでも予測精度を改善する可能性がある．

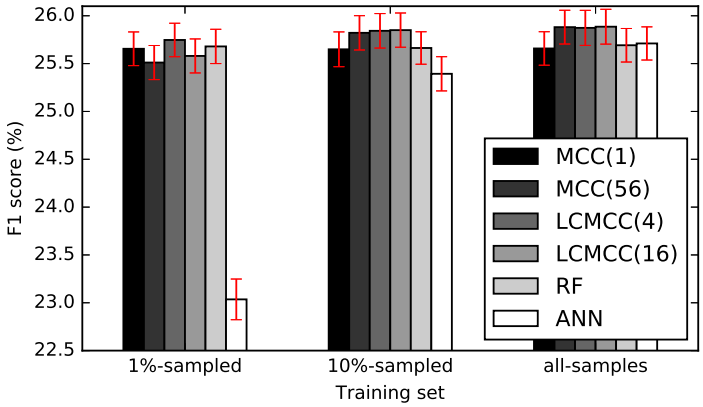
[*4] 図 2.3, 2.5 では基線が 0 でないことに留意されたい．



(a) $N = 3$



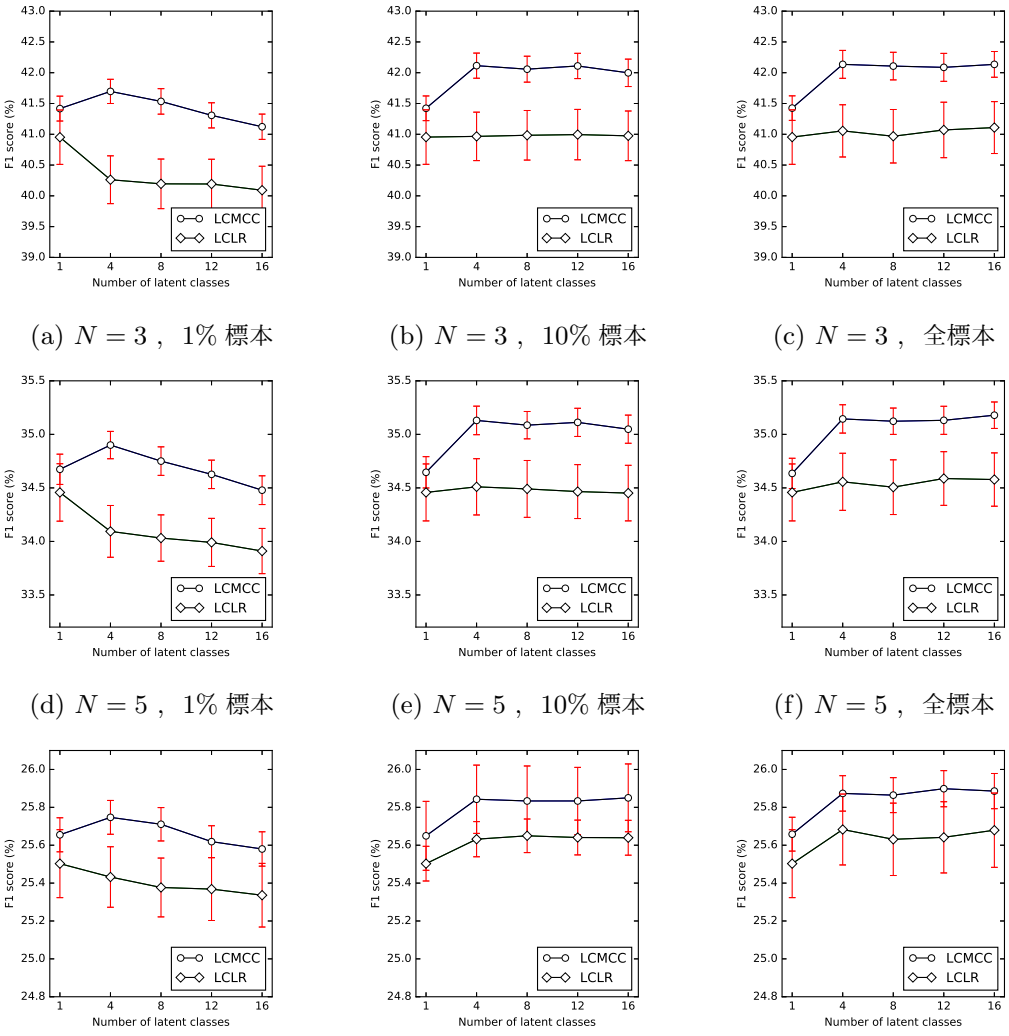
(b) $N = 5$



(c) $N = 10$

図 2.3. 標本抽出した学習データを利用した各手法の F1 値

10% 標本の学習データの場合，図 2.4(c)，2.4(f) から，LCMCC(4) の F1 値は LCMCC(1) よりも高いが，LCLR(4) と LCLR(1) の F1 値の差は小さい．また，10% 標本の学習データを利用した場合，1% 標本の学習データを利用したときに比べて全体的に予測精度が改善している．全標本のときにも 10% 標本の学習データを利用した場合と同様の結果が得られたが，図 2.4(e)，2.4(f) にみられるように，LCMCC(16) が LCMCC(8) と比較して予測精度はわずかに改善している．



(a) $N = 3$，1% 標本　(b) $N = 3$，10% 標本　(c) $N = 3$，全標本

(d) $N = 5$，1% 標本　(e) $N = 5$，10% 標本　(f) $N = 5$，全標本

(g) $N = 10$，1% 標本　(h) $N = 10$，10% 標本　(i) $N = 10$，全標本

図 2.4. 標本抽出した学習データに対する LCMCC，LCLR の F1 値

表 2.2 は LCMCC($|S|$) と LCLR($|S|$) について，潜在クラス数を $|S| \in \{1, 4, 8, 12, 16\}$ として，標本抽出率を 1%，10%，全標本と変化させたときの EM アルゴリズムによるパラメータの推定のための前処理時間，学習時間，予測時間，1 時間あたり RAM MB を示している．それぞれの値は，パラメータの初期値をランダムに変化させて 10 回計算を行い，その平均の時間を示している．また，前処理はページ遷移データからすべての $(i, j, k) \in I \times J \times K$ に対

して (n_{ijk}, q_{ijk}) を集計により計算するための時間で，LCMCC($|S|$) と LCLR($|S|$) に共通して必要な計算時間である．

表 2.2 より，潜在クラス数が増加するにつれて計算時間が増加することがわかる．一方で，各潜在クラス数 $|S| \in \{1, 4, 8, 12, 16\}$ について LCMCC($|S|$) と LCLR($|S|$) の計算時間の差は小さい．

表 2.2. LCMCC と LCLR におけるパラメータ推定と予測の計算時間と 1 時間あたり RAM MB

学習データ	$\|S\|$	前処理時間 [秒]	学習時間 [秒]		予測時間 [秒]		1 時間あたり RAM MB	
			LCMCC	LCLR	LCMCC	LCLR	LCMCC	LCLR
1% 標本	1	4.0	45.9	39.8	3.1	3.1	0.4	0.4
	4		1,626.3	2,203.8	4.9	5.8	26.9	38.8
	8		4,842.8	5,227.7	6.6	7.3	104.5	82.1
	12		9,626.5	9,599.7	9.1	9.2	218.9	138.7
	16		16,090.4	14,431.6	11.0	11.2	379.0	194.4
10% 標本	1	33.8	46.6	46.0	3.2	3.3	0.4	0.9
	4		2,413.3	2,771.6	4.9	5.0	49.3	56.5
	8		5,273.1	6,463.1	6.8	6.7	128.2	108.4
	12		10,343.9	11,418.0	8.9	9.2	253.4	193.0
	16		16,913.9	15,899.2	10.9	11.9	484.9	294.7
全標本	1	338.6	55.8	65.5	4.2	3.4	0.9	0.9
	4		2,100.0	2,953.7	5.1	4.5	47.4	66.6
	8		6,006.1	6,463.9	6.7	6.5	138.4	125.9
	12		11,206.1	11,409.7	9.2	9.9	283.3	210.7
	16		17,701.4	15,741.7	11.6	13.2	513.6	234.9

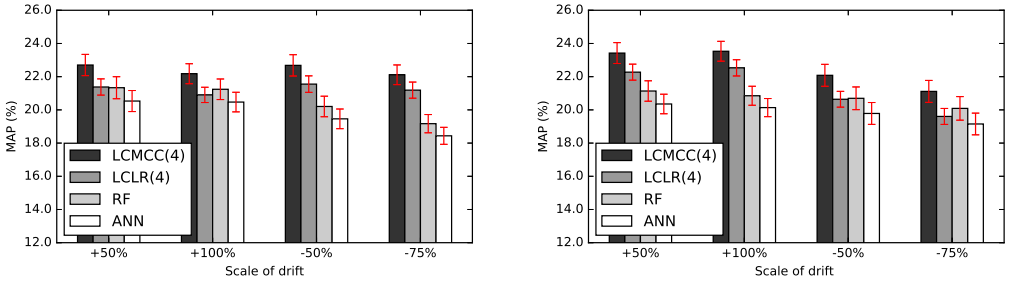
2.4.6 コンセプトドリフトに対する頑健さ

コンセプトドリフトとは，出力データと入力データの関係が，時間や外部影響によって変化することである．商品の流行や特売などの施策の影響による変化によって，学習したモデルが陳腐化してしまうことへの対処が課題とされている [59].

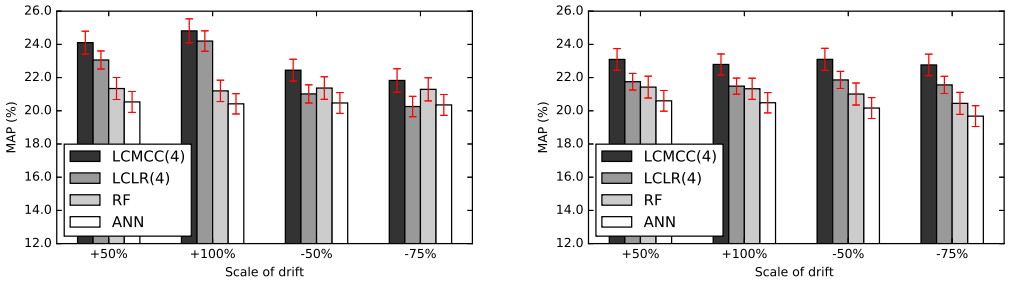
そこで本項では，テストデータに対して人工的にコンセプトドリフトを生成し，提案手法の頑健さについて検証する．検証では，販売数が上位である，レディースファッション/水/食品/靴の 4 種類の商品類型に着目し，これらの商品類型に対してテストデータで，ランダムに顧客と商品の購買の組を増加/減少させる．

図 2.5 は人工的にコンセプトドリフトを加えたテストデータに対する LCMCC(4), LCLR(4), RF, ANN の平均適合率 (Mean Average Precision: MAP) を表している．平均適合率は適合率・再現率曲線の下部の面積の平均を表し，商品選択数に依存しない予測精度の指標である (詳細は Manning et al. [43] を参照).

それぞれのモデルは全標本の学習データを利用してパラメータを推定している．図 2.5(a) では例えば，ドリフトが +50% のとき，テストデータでのレディースファッションの商品の販売数が 50% 増加した状況を表している．平均適合率の値はそれぞれのコンセプトドリフトによって異なるが，いずれの商品類型のコンセプトドリフトに対しても LCMCC(4) が 4 つのモデルの中で常に最も高い平均適合率を達成した．



(a) レディースファッションでのドリフト　　(b) 水でのドリフト



(c) 食品でのドリフト　　(d) 靴でのドリフト

図 2.5. 全標本から推定した予測モデルのドリフトを加えたテストデータに対する平均適合率

2.4.7　商品類型の潜在クラスの分析

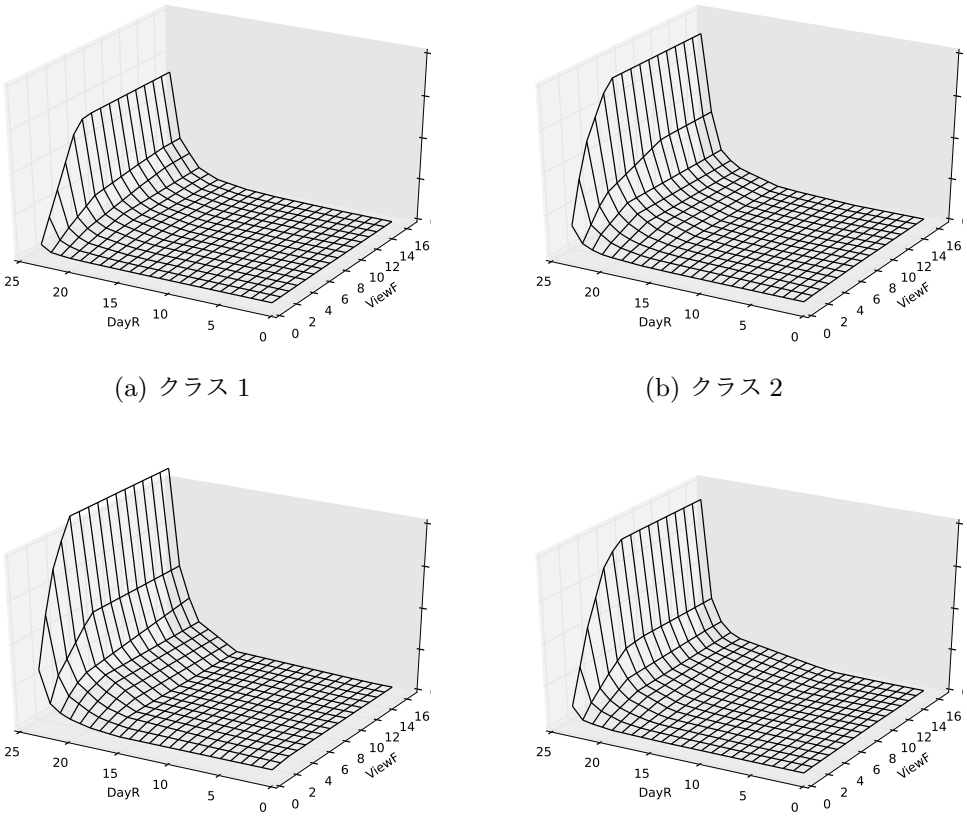
本項では，全標本の学習データで作成した商品類型の潜在クラス確率表について観察する．表 2.3 は LCMCC(4) によって作成された潜在クラスの概要である．商品類型は，最も帰属度の高い潜在クラスの欄に記載した．表 2.3 では，それぞれの潜在クラスで商品類型は商品ページ閲覧回数の降順に記載している．

図 2.6 は，LCMCC(4) によって推定された 4 つの潜在クラスの商品選択確率を表す．商品選択確率は，単調性制約を課すため，最新度と頻度に関する 2 変数の非減少関数によって表される．また最新度は凸性，頻度は凹性の制約条件を課して推定される．商品選択確率は，最新度の値が大きい部分で増分が大きいことがわかる．

4 つの確率表の差異を確認するため，図 2.7 では図 2.6 の確率表について最新度と頻度の一方を固定して商品選択確率を可視化している．表 2.3 と，図 2.6，2.7 から，商品類型の 4 つの潜在クラスの特徴について以下のように考察することができる．

表 2.3. LCMCC(4) により推定された潜在クラスと帰属度の高い商品類型

クラス	π_s	商品類型
$s=1$	0.268	レディースファッション，靴，スポーツ，バッグ，アクセサリー，家具， アウトドア，キッズ用品，メンズファッション，ジュエリー, ゴルフ， 収納家具，自転車，時計，インテリア.
$s=2$	0.161	食品，スイーツ，ソフトドリンク/コーヒー/お茶類，コスメ， PC アクセサリ，ヘルスケア，ダイエット用品，CD/DVD，猫用品/猫.
$s=3$	0.232	水，本，米/シリアル，コンタクトレンズ，ビール，ワイン，漫画 ， サプリメント，薬，ドッグフード，キャットフード，酒類 リキュール，焼酎.
$s=4$	0.339	日用雑貨，スマートフォンアクセサリ, 家電，下着，文具，ベビー， おもちゃ，キッチン用品，カー用品，寝具，食器， 犬用品/犬，DIY，手芸，花，ギフト，ガーデニング用品，ゲーム.



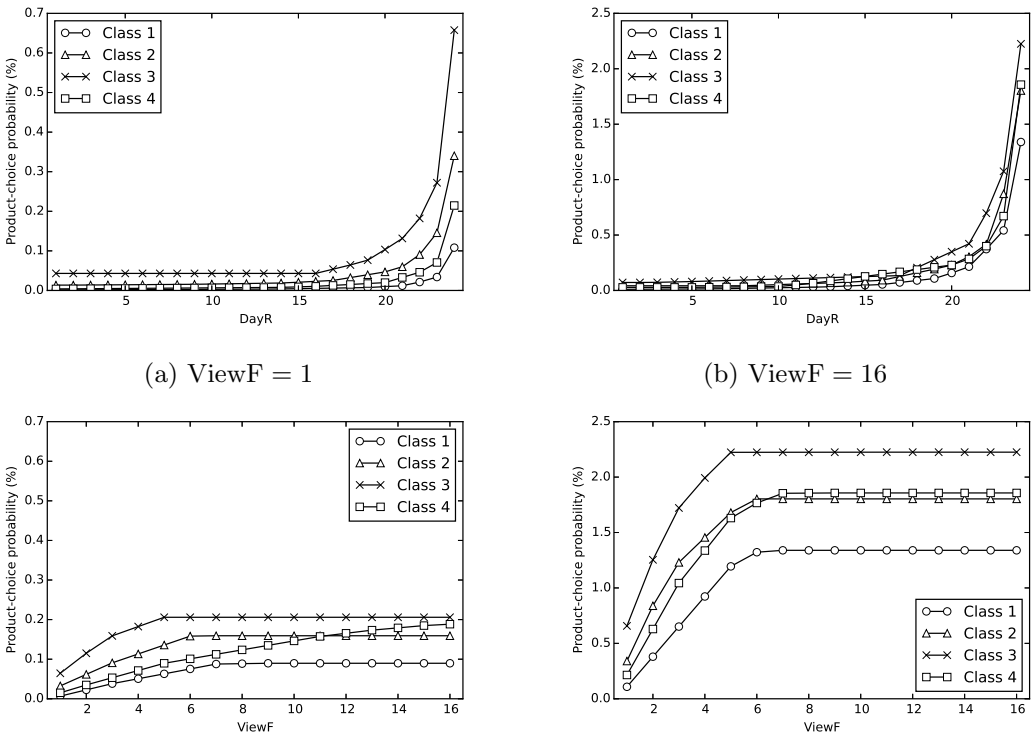
(a) クラス 1　(b) クラス 2

(c) クラス 3　(d) クラス 4

図 2.6. LCMCC(4) で推定された各潜在クラスの 2 次元確率表の可視化

クラス 1　このクラスは表 2.3 に示すように，衣類や家具の商品類型から構成される．図 2.6 や 2.7 から，クラス 1 の商品選択確率は他のクラスよりも低い．その要因として，このクラスに属するレディースファッション，靴，バッグ，アクセサリーのような商品は，繰り返し閲覧され比較された後に購買される商品であることが考えられる．加えて，これらの商品類型の商品については，顧客は EC サイトで商品を確認した後，実際の店舗でこれらの商品を実際に確認して購買する，ということも考えられる．

クラス 2　このクラスは食品，ソフトドリンク/コーヒー/紅茶，化粧品，PC アクセサリなどが含まれる．これらの商品は高価な商品ではない．そのため，閲覧頻度があまり高くなかったとしても購買に至ることがある．加えて，図 2.7(a) に見られるように，クラス 2 の商品選択確率は ViewF = 1 のときに 2 番目に高い．さらに，図 2.7(c)， 2.7(d) からわかるように，クラス 2 の商品選択確率はクラス 4 に比べて ViewF の値が小さいときのみ高い．



(a) ViewF = 1　　(b) ViewF = 16

(c) DayR = 18　　(d) DayR = 24

図 2.7. LCMCC(4) で推定された各潜在クラスの最新度と頻度ごとの商品選択確率の可視化

クラス 3　このクラスは主に，水，米/シリアル，コンタクトレンズ，ドッグフードなどの日常的に利用するような商品からなる．図 2.7 から，クラス 3 の商品選択確率は 4 つのクラスの中で常に最も高いことがわかる．このことから，顧客はこのクラスに含まれる好みの商品について，詳細に見ることなく日常的に繰り返し購買していることが示唆される．

クラス 4 このクラスは家電，文房具，キッチン用品，カー用品などの高価格な商品からなる．日常雑貨の類型には，高価な耐久消費財である電動歯ブラシ，スーツケース，高級タオルなども含まれている．これらの商品は頻繁には購買されないので，顧客は詳細を繰り返し閲覧した後に購買をする可能性がある．加えて，図 2.7(c) を見るとクラス 4 の商品選択確率は ViewF に関して単調に増加し，ViewF = 12 のときにクラス 2 よりも高くなっていることがわかる．つまり，クラス 4 の商品は商品ページ閲覧回数が増加するにつれて購買されやすくなる傾向がクラス 2 の商品よりも強いことがわかる．

2.5 まとめ

本章では，各顧客の過去の閲覧履歴の最新度と頻度の特徴量から商品選択確率を推定する，潜在クラス形状制約モデルを提案した．提案手法は，商品をいくつかの潜在クラスに分類し，それぞれのクラスについて最新度と頻度に関する単調性，凸性，凹性の制約条件を満たすように 2 次元の確率表を推定する．また，潜在クラス確率表を推定するための EM アルゴリズムも本章で提案した．さらに，実際のページ遷移データを用いて作成した潜在クラスモデルから，商品類型ごとの顧客の商品選択行動について分析した．

本章の貢献は，ページ遷移データ分析のための潜在クラスモデルの提案である．特に，既存手法の MCC モデルに潜在クラスモデルを適用することで予測精度を改善した．また，EC サイトの閲覧の最新度と頻度をもとにした商品のクラスタ分析の新しい方法を提案した．

数値実験の結果，提案手法は既存の潜在クラスロジスティック回帰モデルよりも予測精度が高かった．この結果は，形状制約と潜在クラス回帰を組み合わせることの有効性を示しており，マーケティングなどの実務の領域でのさらなる潜在クラス形状制約回帰の活用の可能性を示している．

推定された商品選択確率は顧客の商品に対する嗜好を表している．そのような嗜好に関する情報は，顧客の利便性を向上させるための商品推薦システムにとって有用である．そのため，商品選択確率の高精度の推定は，EC サイトの商品推薦システムの性能向上に寄与する．加えて，一度 EM アルゴリズムによって潜在クラスごとの確率表を推定すれば，顧客と商品の組について集計された最新度と頻度を確率表で参照するだけで利用することができるという実用的な利点もある．また，潜在クラスモデルはそれぞれの商品類型に対する販売促進戦略を計画するためにも有用な示唆を与えることができる．

EC サイトにおける購買予測では，閲覧される商品と比べて購買される商品がかなり少ないため，不均衡なデータから学習を行わなければならない．そのような不均衡データによって，多数側のデータに偏ったモデルの学習が行われてしまうため，データの前処理やアルゴリズムの観点から研究が行われている [20, 38]．2.4.2 節で述べたように，今回実験で利用したデータは，負例が正例と比較して約 2000 倍存在し，商品選択確率の推定を難しくしている．しかし本章での提案手法では，形状制約のもとでの回帰によって不均衡データをより適切に扱うことで予測精度を向上させた．

顧客の商品閲覧データは絶え間なく発生し利用可能になるため，事業者はモデルを適切なタイミングで更新する必要がある．特にデータにおけるコンセプトドリフトの問題に対処するため，4 つの方法（コンセプトドリフトの検知，学習データの期間の設定，オンライン学習，アンサンブル学習）が提案されている [59]．潜在クラス確率表モデルは，直接的にコンセプトドリフトを考慮していないが，形状制約の効果によって小さな標本サイズの学習データの場合にも高い予測精度を達成した．そのため直近の期間のデータを利用してモデルを更新する場合にも，データの変化に対応することができる．データの変化に対しては，EM アルゴリズムのオンライン学習 [40] も有効であると考えられる．

本章の研究の今後の課題として，形状制約モデルにおいて商品類型に加えて，顧客の潜在クラスを考慮することが挙げられる．予備実験では，顧客属性の潜在クラスを考慮することによるモデルの予測精度向上は確認できなかった．しかし，顧客の異質性の考慮は多くの統計的マーケティング研究における課題となっており，顧客属性の潜在クラスの考慮についてはさらなる分析の余地が考えられる．

第 3 章

商品閲覧数の時系列に基づく形状制約閲覧数列モデル

3.1　はじめに

本章では，EC サイトにおける顧客の商品閲覧数の時系列と商品選択の関係に着目する．既存研究 [32] では，1 章で述べたように，商品閲覧履歴から各顧客の閲覧商品に対する「最新度」と「頻度」を数量化し，形状制約のもとで最新度と頻度の組に対して閲覧商品が購買される確率（商品選択確率）を推定する手法が提案された．しかし，この手法では顧客の閲覧履歴が最新度と頻度の 2 次元に縮約されるため，商品閲覧に関する多くの情報が失われてしまう．閲覧履歴を二つの特徴量に縮約するのではなく，各期間の商品閲覧数の時系列（閲覧数列）を利用することで，より詳細な商品選択に至る過程に関する情報が利用可能になり，商品選択を高精度で予測できる可能性がある．

そこで本章では，顧客の商品に対する閲覧数列に対して商品選択確率を推定する閲覧数列モデルを提案する．

本章の構成は以下の通りである．3.2 節では，各期間の商品閲覧数の時系列である閲覧数列について定義し，観察に基づき，商品選択確率に関して閲覧数列に期待される順序関係について定義する．3.3 節では，閲覧数列の順序関係に基づき商品選択確率を推定する閲覧数列モデルの最適化問題としての定式化と，閲覧数列モデルの推定を効率化するための冗長な制約条件の削除方法について述べる．3.4 節では，閲覧数列モデルの商品選択確率の予測精度の評価と，冗長な制約条件の削除の効果について評価する．最後に，3.5 節で本章のまとめと課題を述べる．

3.2　閲覧数列

本節では，まず各期間の商品閲覧数の時系列である閲覧数列について定義する．そして観察に基づき，商品選択確率に関して閲覧数列に期待される順序関係について定義する．

3.2.1　閲覧数列の定義

顧客と商品の組に対して，i 期前の閲覧数を v_i，考慮する期間数を k とするとき，閲覧数列は k 次元ベクトルとして以下のように表される：

$$\boldsymbol{v} := (v_1, v_2, \ldots, v_k).$$

また，1 期間の閲覧数の上限を $v_{\max}$ とするとき，可能なすべての閲覧数列の集合は $\Gamma := \{0, 1, \ldots, v_{\max}\}^k$ となる．

例として $\boldsymbol{v} = (0, 1, 2, 0)$ なる閲覧数列を考える．これは 2 期前と 3 期前にそれぞれ 1 回と 2 回の閲覧があり，他の期間は閲覧がない顧客と商品の組のパターンである．2 次元確率表では，この閲覧数列 $\boldsymbol{v}$ は「最終閲覧が 2 期前」で「総閲覧数が 3 回」のような最新度と頻度の 2 次元のデータに縮約され，他の閲覧数列 $(0, 3, 0, 0)$ や $(0, 1, 1, 1)$ と同一視される．区別して扱うことのできる（閲覧数ゼロを除く）パターン数で比較すると，2 次元確率表は $kv_{\max}$ 個，閲覧数列は $(1 + v_{\max})^k - 1$ 個となり，大きな差がある．

閲覧数列は 2 次元確率表と比べて考慮できるパターン数が多いため，商品選択確率の予測精度が高くなることが期待される．一方で，閲覧数列のパターン数が多くなった場合，それぞれの閲覧数列のパターンに対して十分な量の学習データを確保することは難しくなってしまう．

各パターンに対して学習データが少ない状態で商品選択確率を推定することの問題として，その少ないデータに対して過剰適合した推定が行われてしまうことがある．そこで本研究では，2 次元確率表と同様に，閲覧数列に対して形状制約を導入することで過剰適合を抑制し推定精度を向上させることを考える．

3.2.2　閲覧数列の順序関係

既存研究の 2 次元確率表モデルでは「閲覧回数が多い商品ほど選択確率が高い」という頻度の単調性と，「直近に閲覧した商品ほど選択確率が高い」という最新度の単調性という 2 つの性質に基づき商品選択確率の補正している．本項では，最新度と頻度に基づく 2 次元確率表モデルの単調性を，閲覧数列に対して拡張することを試みる．

本研究で提案する閲覧数列の集合 Γ 上の順序関係を $\succeq$ とし，以下で述べる観察に基づいて定義する．

観察 1：閲覧数列の頻度に基づく単調性

二つの閲覧数列において，ある一つの期間の閲覧数のみ異なる場合，その期間に多く閲覧されているもののほうが選択確率が高くなることが期待される．形式的には，

$$(\ldots, v_i, \ldots) \to (\ldots, v_i + 1, \ldots.) \tag{3.1}$$

なる操作（閲覧数列のある成分の値を 1 単位増加させること）を繰り返して $\boldsymbol{v}$ から $\boldsymbol{v}'$ へ変換できるとき，$\boldsymbol{v}' \succeq \boldsymbol{v}$ とする．たとえば $(1, 2, 1) \succeq (1, 1, 1) \succeq (0, 1, 1)$ の関係が成り立つことを

仮定する．

観察 2-1：閲覧数列の最新度に基づく単調性 1

閲覧数の総和が等しい二つの閲覧数列を比較した場合，直近により多く閲覧されているもののほうが選択確率が大きくなることが期待される．形式的には，

$$(\ldots, v_i, \ldots, v_j, \ldots) \to (\ldots, v_i + 1, \ldots, v_j - 1, \ldots) \tag{3.2}$$

なる操作（閲覧数列の成分の値を直近の期間へ 1 単位移動させること）を繰り返して $\boldsymbol{v}$ から $\boldsymbol{v}'$ へ変換できるとき，$\boldsymbol{v}' \succeq \boldsymbol{v}$ とする．たとえば $(2,1,0) \succeq (1,2,0)$ や，$(1,2,0) \succeq (1,1,1)$ の関係が成り立つことを仮定する．

観察 2-2：閲覧数列の最新度に基づく単調性 2

観察 2-1 では，たとえば $(1,2,0) \succeq (1,1,1)$ について，閲覧数の総和がともに 3 であり，3 期前の閲覧数を 2 期前に移動させることで変換可能なため，閲覧数列 $(1,2,0)$ は $(1,1,1)$ よりも商品選択確率が大きいと仮定をする．しかし，閲覧数列 $(1,2,0)$ と $(1,1,1)$ について，$(1,2,0)$ は 2 期前から 1 期前にかけて閲覧数が減少しているため商品に対する関心が薄れている，$(1,1,1)$ は 3 期間にわたり閲覧数が一定のため関心が持続している，と考え $(1,2,0) \succeq (1,1,1)$ の関係は成立しないと考えることもできる．一方で，閲覧数列 $(1,2,0)$ と $(1,0,2)$ については，この観点に矛盾しない．つまり，二つの期間の閲覧数が交換された形である二つの閲覧数列を比較した場合に，直近により多く閲覧されているもののほうが商品選択確率が大きくなることを期待する．形式的には，

$$v_j > v_i, \quad (\ldots, v_i, \ldots, v_j, \ldots) \to (\ldots, v_j, \ldots, v_i, \ldots) \tag{3.3}$$

なる操作（閲覧数列の遠い期間の大きい閲覧数の成分と近い期間の小さい閲覧数の成分を交換する）を繰り返して $\boldsymbol{v}$ から $\boldsymbol{v}'$ へ変換できるとき，$\boldsymbol{v}' \succeq \boldsymbol{v}$ とする．

本章では，上記の観察を踏まえ，Γ 上の順序関係 $\succeq_1, \succeq_2$（それぞれを特に区別しない場合には単純に $\succeq$ と記す）をそれぞれ次のように定義する．

閲覧数列の順序関係 1

$$\boldsymbol{v}' \succeq_1 \boldsymbol{v} \iff \begin{cases} \textbf{(a)}\ \boldsymbol{v} = \boldsymbol{v}'，\text{または，} \\ \textbf{(b)}\ \text{式 (3.1), (3.2) の操作を繰り返して } \boldsymbol{v} \text{ を } \boldsymbol{v}' \text{ に変換可能．} \end{cases} \tag{3.4}$$

閲覧数列の順序関係 2

$$\boldsymbol{v}' \succeq_2 \boldsymbol{v} \iff \begin{cases} \textbf{(a)}\ \boldsymbol{v} = \boldsymbol{v}'，\text{または，} \\ \textbf{(b)}\ \text{式 (3.1), (3.3) の操作を繰り返して } \boldsymbol{v} \text{ を } \boldsymbol{v}' \text{ に変換可能．} \end{cases} \tag{3.5}$$

そして順序関係 $\boldsymbol{v}' \succeq \boldsymbol{v}$ が成り立つ場合には，閲覧数列 $\boldsymbol{v}$，$\boldsymbol{v}'$ の商品選択確率をそれぞれ $x_{\boldsymbol{v}}$，$x_{\boldsymbol{v}'}$ として，商品選択確率の単調性 $x_{\boldsymbol{v}'} \geq x_{\boldsymbol{v}}$ を要請する．

3.3　閲覧数列モデル

本節では，閲覧数列の順序関係に基づき商品選択確率を推定する閲覧数列モデルの最適化問題としての定式化と，閲覧数列モデルの推定を効率化するための冗長な制約条件の削除方法について述べる．

3.3.1　定式化

閲覧数列モデルでは，順序関係 $\succeq$ に基づく単調性制約の下で，閲覧数列 $\boldsymbol{v}$ の標本サイズ $n_{\boldsymbol{v}}$ によって重み付けられた残差 2 乗和が最小となるように，商品選択確率 $x_{\boldsymbol{v}}$ を推定する．この問題は以下のような凸 2 次最適化問題として定式化できる：

$$\left|\begin{array}{lll} \text{minimize} & \displaystyle\sum_{\boldsymbol{v}\in\Gamma} n_{\boldsymbol{v}}(x_{\boldsymbol{v}}-\hat{x}_{\boldsymbol{v}})^2 & \\ \text{subject to} & x_{\boldsymbol{v}'} \geq x_{\boldsymbol{v}}, & (\boldsymbol{v},\boldsymbol{v}'\in\Gamma \text{ かつ } \boldsymbol{v}'\succ\boldsymbol{v}), \\ & 0\leq x_{\boldsymbol{v}}\leq 1, & (\boldsymbol{v}\in\Gamma). \end{array}\right. \tag{3.6}$$

ただし $\hat{x}_{\boldsymbol{v}}$ は学習データの集計に基づく選択確率の推定値であり，学習データ数 $n_{\boldsymbol{v}}$ が少ない場合には値の信頼性が低いと考えられる．

また $\hat{x}_{\boldsymbol{v}}$ については，単純な集計方法である「(選択された商品数)/(閲覧された商品数)」の計算により算出された値以外にも，機械学習モデルの予測値などを使用することもでき，その場合にも閲覧数列の順序関係を満たすように予測値の補正が行われる．後の数値実験では，機械学習モデルの予測値に対して，閲覧数列の順序関係を満たすように補正した場合の効果についても検証する．

式 (3.6) では，閲覧数列の順序関係が必ず満たされると仮定して，単調性制約つきの最適化問題を求解するが，学習データ数の多い信頼性の高い部分については，それが真の値であると仮定し，順序関係を満たさないことを許容するモデルも考えられる．Tibshirani ら [67] は，推定値の単調性を制約条件として課すのではなく，単調性の違反に対する罰則項として目的関数に加える（正則化する）モデルを提案した．

閲覧数列モデルについても，Tibshirani ら [67] のモデルのように，閲覧数列の順序関係を制約条件として課すのでなく，学習データ数 $n_{\boldsymbol{v}}$ が多い実績確率値の信頼性の高い部分は単調性を違反することを許容するように拡張し定式化することができる：

$$\left|\begin{array}{lll} \text{minimize} & \displaystyle\sum_{\boldsymbol{v}\in\Gamma} n_{\boldsymbol{v}}(x_{\boldsymbol{v}}-\hat{x}_{\boldsymbol{v}})^2 + \lambda \sum_{\boldsymbol{v},\boldsymbol{v}'\in\Gamma \text{ かつ } \boldsymbol{v}'\succ\boldsymbol{v}} (x_{\boldsymbol{v}}-x_{\boldsymbol{v}'})_+ & \\ \text{subject to} & 0\leq x_{\boldsymbol{v}}\leq 1, & (\boldsymbol{v}\in\Gamma). \end{array}\right. \tag{3.7}$$

ただし λ は順序関係の違反に対してどの程度罰則を課すかを表す正則化パラメータ，$(x_{\boldsymbol{v}}-x_{\boldsymbol{v}'})_+$ は $x_{\boldsymbol{v}}-x_{\boldsymbol{v}'}>0$ の場合は $x_{\boldsymbol{v}}-x_{\boldsymbol{v}'}$ であり，それ以外の場合は 0 である項を表す．式 (3.7) において，正則化パラメータ λ の値を 0 とすると推定値は学習データの集計に基づく選択確率の推定値と一致し，値を大きくするにつれて閲覧数列の順序関係を満たすように推定

値が補正される．本研究では，式 (3.7) のように単調性などの性質を制約条件でなく罰則項として目的関数に加えるモデルを，形状正則化モデルとよぶ．

形状正則化モデルの利点として，正則化パラメータ λ をうまく調整することで，推定値の信頼性が低い部分は順序関係を満たす，信頼性の高い部分は補正を弱くし順序関係を違反することを許容する，というようにモデルのパラメータを柔軟に推定できることが挙げられる．一方で欠点として，正則化パラメータ λ の値を適切に決定しなければならないことが挙げられる．

3.3.2 閲覧数列モデルの冗長な制約条件の削除

上記の最適化問題は，素朴にすべての順序関係を制約条件として課した場合，制約条件の数が膨大になってしまい最適化計算が困難になってしまう．そこで，本研究では半順序の簡約表現であるハッセ図の構造を利用して，有限半順序集合 $(\Gamma, \succeq)$ があるとき，閲覧数列 $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3 \in \Gamma$ について，$\boldsymbol{v}_1 \succ \boldsymbol{v}_2$ で，かつ $\boldsymbol{v}_1 \succ \boldsymbol{v}_3 \succ \boldsymbol{v}_2$ となるような $\boldsymbol{v}_3$ が存在しない場合にのみ $x_{\boldsymbol{v_1}} \geq x_{\boldsymbol{v_2}}$ の制約条件を課す．このように冗長な順序関係を削除したとき，任意の有限な半順序に対してハッセ図が一意に定まることが知られているため，最終的に課される制約条件の組も一意に定まる．

この工夫によって，最適化計算の大幅な効率化が可能になる．たとえば期間数と閲覧数の上限 $(k, v_{\max})$ がそれぞれ $(3, 2)$ のとき，閲覧数列の順序関係 1 では，すべての式 (3.1),(3.2) の変換に対応する閲覧数列の組に対して単調性制約を課すと図 3.1 のように 90 組の順序関係が課されるが，冗長な制約条件を削除することで図 3.2 のように，42 組まで削減できる．図のグラフはそれぞれ，頂点は閲覧数列を，辺は赤い辺が閲覧数列の頻度の単調性を，黒い辺が閲覧数列の最新度の単調性を表す．同様にして，閲覧数列の順序関係 2 では，すべての式 (3.1),(3.3) で変換可能な閲覧数列の組に対して単調性制約を課すと図 3.3 のように 81 組の順序関係が課されるが，冗長な制約条件を削除することで図 3.2 のように，46 組まで削減できる．期間数 k や閲覧数上限 $v_{\max}$ が大きくなるほど，この方法による効率化の効果は大きくなり，期間数と閲覧数の上限 $(k, v_{\max})$ がそれぞれ $(3, 4)$ のとき，閲覧数列の順序関係 1 を考慮した時，図 3.5 と図 3.6 のように 540 組から 260 組に，閲覧数列の順序関係 2 を考慮した時，図 3.7 と図 3.8 のように 450 組から 316 組に削減することができる．

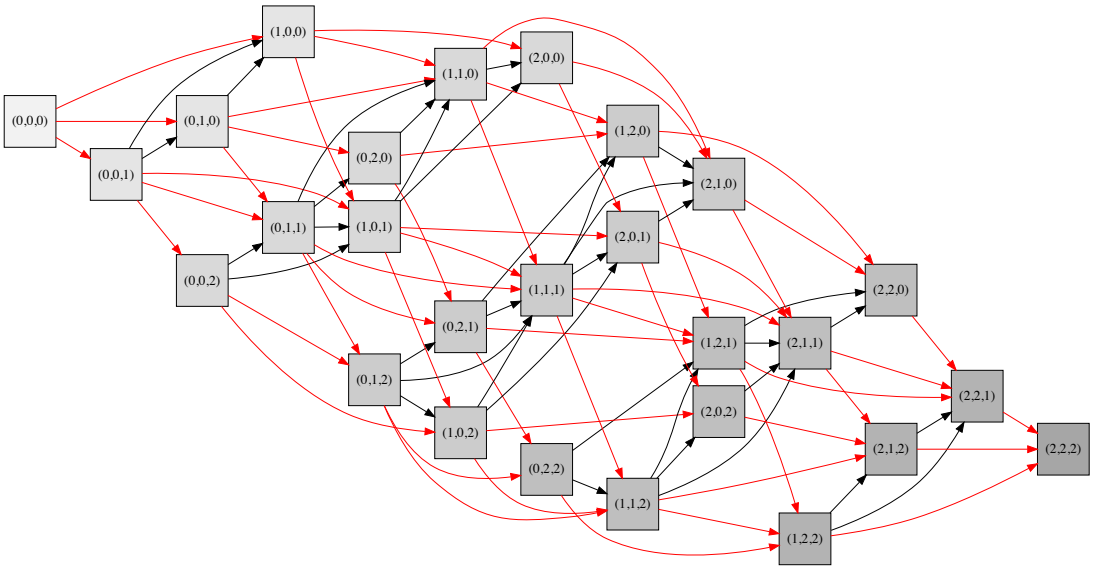


図 3.1. 閲覧数列の順序関係 1，$(k, v_{\max}) = (3, 2)$

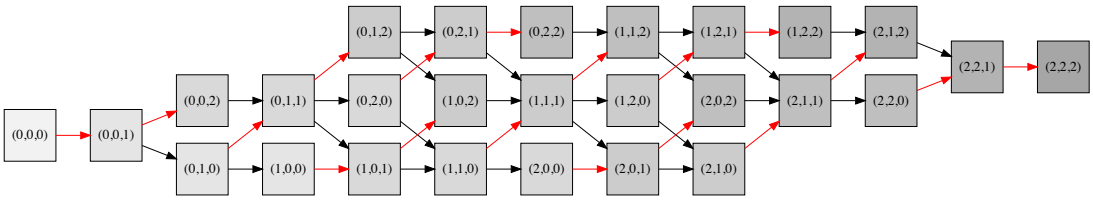


図 3.2. 閲覧数列の順序関係 1 のハッセ図，$(k, v_{\max}) = (3, 2)$

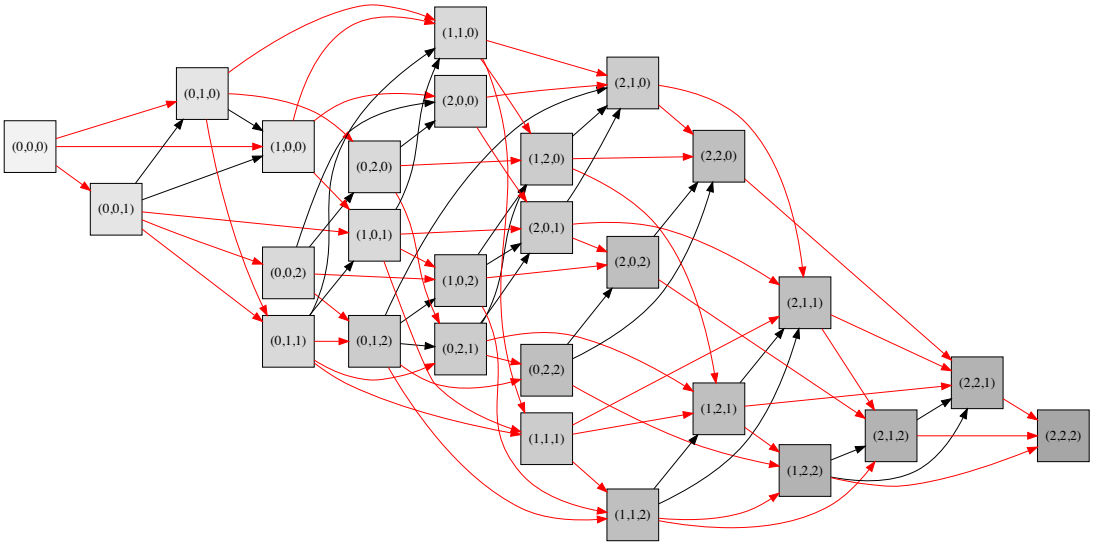


図 3.3. 閲覧数列の順序関係 2，$(k, v_{\max}) = (3, 2)$

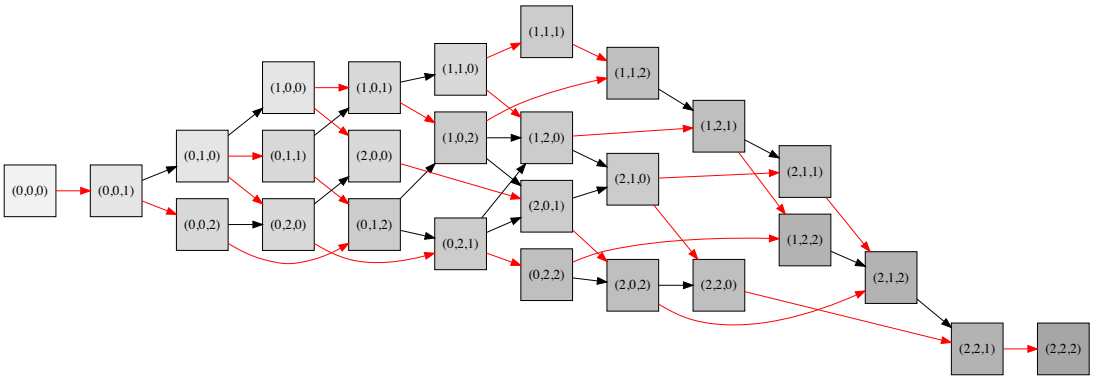


図 3.4. 閲覧数列の順序関係 2 のハッセ図，$(k, v_{\max}) = (3, 2)$

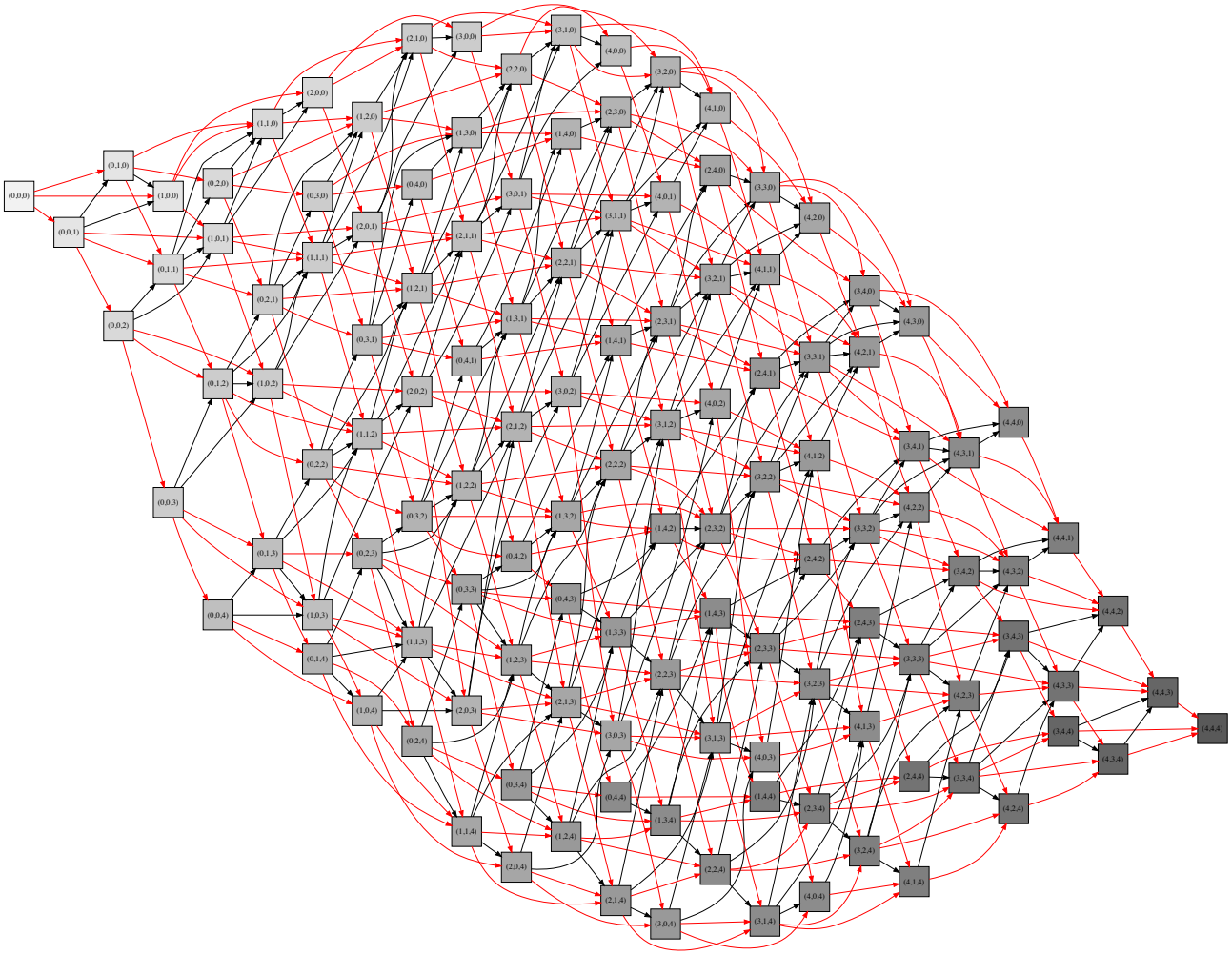


図 3.5. 閲覧数列の順序関係 1，$(k, v_{\max}) = (3, 4)$

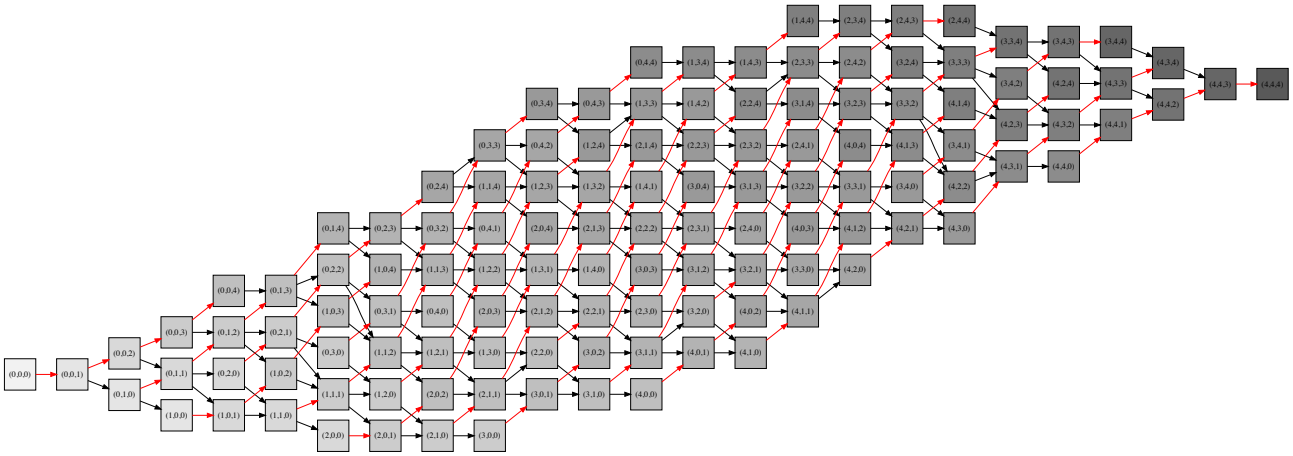


図 3.6. 閲覧数列の順序関係 1 のハッセ図，$(k, v_{\max}) = (3, 4)$

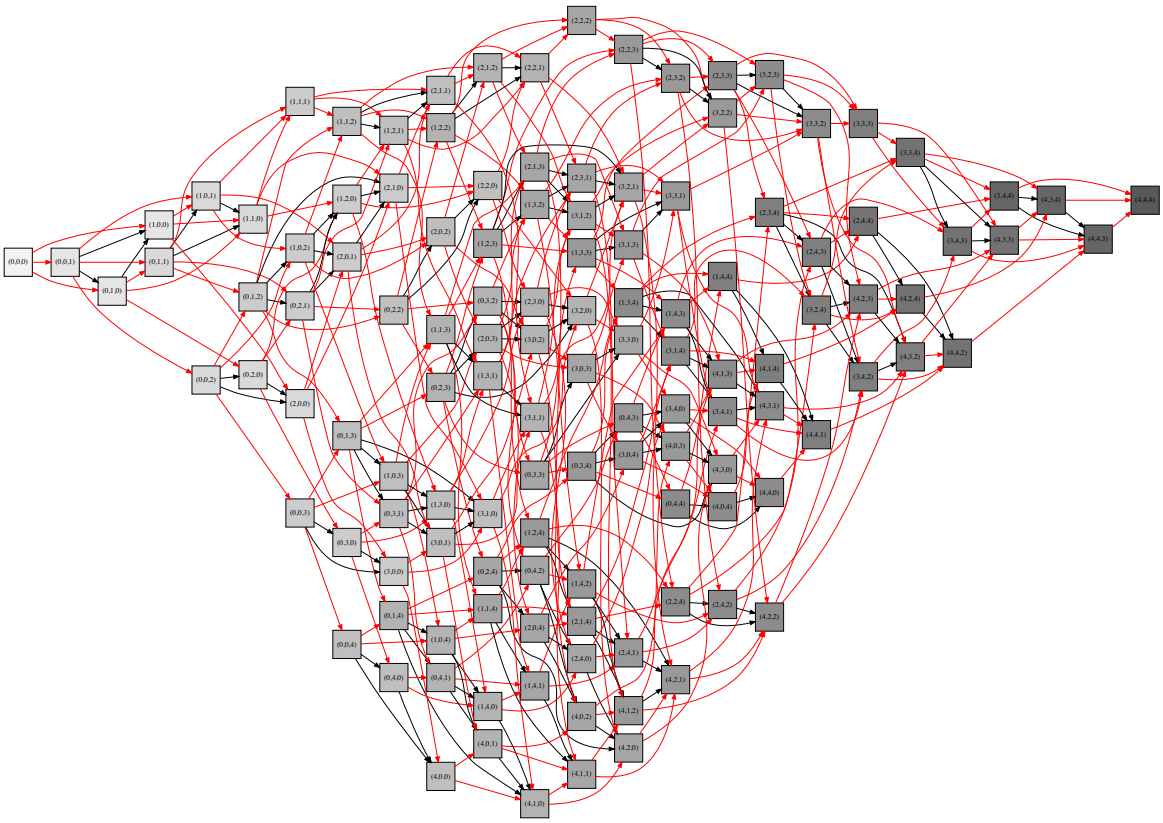


図 3.7. 閲覧数列の順序関係 2，$(k, v_{\max}) = (3, 4)$

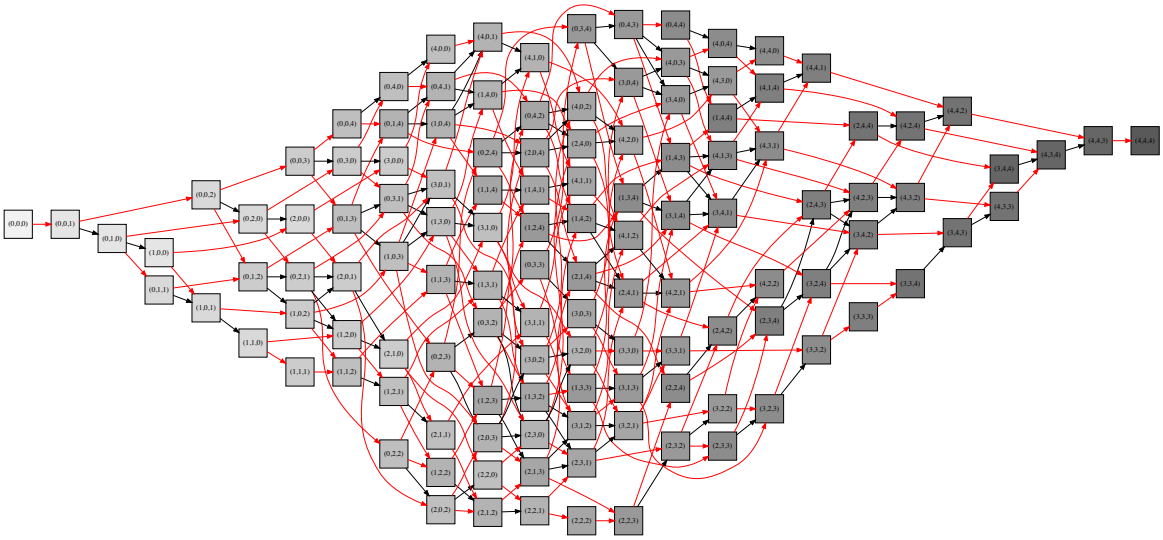


図 3.8. 閲覧数列の順序関係 2 のハッセ図，$(k, v_{\max}) = (3, 4)$

3.4　数値実験

本節では，閲覧数列モデルの商品選択確率の予測精度と，冗長な制約条件の削除の効果について評価する．

3.4.1　ページ遷移データ

本章の数値実験では，中国の EC サイト TMall のデータ [41] を用いた．このデータは TMall のデータ解析コンペティションのために公開され，TMall 内のページ遷移データが 1 年分利用できる．データの 1 レコードが，一つの商品に対する商品閲覧を表し，閲覧時刻，顧客 ID，商品 ID などの情報を含んでいる．利用データでは 422,282 通りのユニークな顧客 ID，624,221 通りのユニークな商品 ID, 28,316,459 通り の顧客 ID と商品 ID のユニークな組を含んでいる．タイムスタンプをもとに各日の閲覧数を集計し，閲覧数列を作成した．

3.4.2　比較モデル

数値実験では，5 つの商品選択確率の推定手法の性能を比較した．閲覧数列モデル，機械学習モデル（ランダムフォレスト，人工ニューラルネットワーク，L2 正則化ロジスティック回帰）では，3.4.5 項で述べるように，期間数，閲覧数上限の複数の組合せ $(k, v_{\max})$ の閲覧数列を入力として学習し検証した．閲覧商品に対する閲覧数列の算出では，閲覧日が k 日前を超えた閲覧は，k 日前の閲覧として算出した．また，ある日の閲覧数が $v_{\max}$ を超えた顧客と商品の組については，その日の閲覧数を $v_{\max}$ 回と算出した．本章では 2 次元確率表モデルとして，最適化問題 (1.2) を解く単調性モデルを利用した．2 次元確率表モデルでは最新度の最大値と頻度の最大値をそれぞれ比較する閲覧数列モデルと同様に $(k, v_{\max})$ として，例えば，m (≥ 1) 日前に最終閲覧をした顧客と商品の組については，最新度の値は $\max\{k - m + 1, 1\}$ とし，閲覧回数が n (≥ 1) の顧客と商品の組については，$\min\{n, v_{\max}\}$ とした．また実験結果では，各モデルについて表 3.1 の略称を用いる．

表 3.1. 比較モデルの略称

略称	モデル
PV-seq1	閲覧数列の順序関係 1 を用いた閲覧数列モデル
PV-seq2	閲覧数列の順序関係 2 を用いた閲覧数列モデル
2-dim	2 次元確率表モデル [32]
LR	L2 正則化ロジスティック回帰
ANN	人工ニューラルネットワーク
RF	ランダムフォレスト

PV-seq1，PV-seq2，2-dim の最適化計算では，凸二次計画ソルバーである OSQP [65] を利用した．LR と ANN と RF の計算では，Python の機械学習ライブラリである，scikit-learn (ver. 0.18.1) の, LogisticRegressionCV, MLPRegressor, RandomForestRegressor をそれぞれ用いた．機械学習モデルにはハイパーパラメータがそれぞれ存在し，それらのハイパーパラメータは論文 [56] を参考に，表 3.2 の範囲で 3 分割交差確認法により探索した．

表 3.2. 機械学習モデルのハイパーパラメータの探索範囲

ライブラリ	ハイパーパラメータ	探索範囲
LogisticRegressionCV	Cs	[0.0001, 10000] の範囲で対数的に等間隔に 10 点
MLPRegressor	activation	{‘logistic’, ‘tanh’, ‘relu’}
	solver	{‘lbfgs’, ‘adam’, ‘sgd’}
	learning_rate	{‘constant’, ‘invscaling’, ‘adaptive’}
RandomForestRegressor	n_estimators	{10, 100, 1000}
	min_weight_fraction_leaf	{0.0, 0.25, 0.5}
	max_features	{‘sqrt’, ‘log2’, None}

また各機械学習モデルでは，前処理として閲覧数列の各期間について平均 0，分散 1 に正規化した．加えて，データの正例と負例の偏りを解消し予測精度を向上させるために，各機械学習モデルでは学習データにおける負例数を減少させるアンダーサンプリング処理を行った．

3.4.3 評価方法

表 3.1 の 5 つの手法の評価では，まず 90 日間の学習期間の閲覧履歴における顧客の閲覧商品に対して，それぞれどの程度再閲覧されるかをスコア付けする．そして，スコアの大きいものから順に上位 N 個を選択し，それらが実際に再閲覧されたかどうかを F1 値で比較する．スコアが同じ商品については，より直近に閲覧されたものを選択した．TMall のデータ [41] では，90 日間の学習データと 1 日のテストデータが 5 組提供されており，実験では 4 組をモデルの学習に利用し，残り 1 組のデータをモデルの評価に利用した．また，データ量に対する頑健さを評価するために，学習データから $10\%, 1\%, 0.1\%$ の標本を無作為抽出したデータについても検証を行い，10 回の無作為抽出での試行の平均の F1 値を算出した．計算には，Intel Core i7 CPU (3.1 GHz)，16GB メモリを搭載した計算機を用いた．

3.4.4 ハッセ図の構造を利用した冗長な制約条件の削除の効果

まず，閲覧数列モデルにおいて冗長な制約条件を削除することによる効果を検証する．表 3.3 は，期間数と閲覧数の上限を変化させたときの，閲覧数列に対して頻度の単調性と最新

度の単調性の制約条件をすべて課した場合 (All1，All2)，式 (3.4) の (b)，式 (3.5) の (b) で変換可能な制約条件をすべて課した場合 (Naive1, Naive2)，ハッセ図の構造を利用して冗長な制約条件を削除した場合 (Hasse1, Hasse2) の制約条件の数を表し，表 3.4 はそれぞれの場合の最適化の計算時間を表す．表 3.4 で，OM は計算において計算機がメモリ不足となったことを示している．

冗長な制約条件を削除する操作は，Python のライブラリ NetworkX[*1]の transitive_reduction 関数を用いた．

All1 と All2，または Naive1 と Naive2 の制約条件数を比較すると，期間数と閲覧数上限の組合せが $(5,1),(1,6)$ の場合を除いて，最新度の単調性 1 を用いた場合が，最新度の単調性 2 を用いた場合よりも，制約条件数が多いことがわかる．これは，期間数 k に対して 2 つの期間の組合せは ${}_k\mathrm{C}_2$ 組存在するが，最新度の単調性 1 の「閲覧数列の成分の値を直近の期間へ 1 単位移動させる操作」が可能な組合せは，最新度の単調性 2 の「閲覧数列の遠い期間の大きい閲覧数の成分と近い期間の小さい閲覧数の成分を交換する操作」が可能な組合せよりも多いためである．Naive1 と Naive2 の制約条件の差異は図 3.1 と図 3.3，図 3.5 と図 3.7 の差がそれぞれ対応する．

一方で，Hasse1 と Hasse2 の制約条件数の比較から，ハッセ図の構造を考慮して冗長な制約条件を削除したとき，最新度の単調性 1 を用いた場合のほうが最新度の単調性 2 を用いた場合よりも制約条件の数が少ないことがわかる．これは，「閲覧数列の成分の値を直近の期間へ 1 単位移動させる操作」では冗長な制約条件を削除すると，隣接期間の移動に対応する $k-1$ 組の制約条件のみが残り，多くの制約条件が削除されるのに比べて，「閲覧数列の遠い期間の大きい閲覧数の成分と近い期間の小さい閲覧数の成分を交換する操作」では隣接期間の移動に対応しない制約条件も残るためである．Hasse1 と Hasse2 の制約条件の差異は図 3.2 と図 3.4，図 3.6 と図 3.8 の差がそれぞれ対応する．

Naive1 と Hasse1，Naive2 と Hasse2 での最適化計算時間を比較すると，いずれも冗長な制約条件を削除することで計算が高速化していることがわかる．また制約条件数が多くなるにつれて，計算機で利用するメモリも多く必要となるため，冗長な制約条件を削除することは計算時間削減，メモリ利用量削減の両方の観点で有効である．

3.4.5　期間数と閲覧数上限の組合せに対する閲覧数列モデルの予測精度

期間数と閲覧数上限を変化させながら，実績の選択確率の集計値を用いた閲覧数列モデル (PV-seq(EMP))，閲覧数列の順序関係 1 を用いた閲覧数列モデル (PV-seq1)，閲覧数列の順序関係 2 を用いた閲覧数列モデル (PV-seq2) の予測精度を，各手法で商品選択確率の推定値が高い順に上位 3 商品を予測したときの F1 値を用いて比較する．

期間数と閲覧数上限の組は，期間数を 3 から始めて，PV-seq1 と PV-seq2 の最適化計算が 30 分以内に終了した閲覧数上限をそれぞれ選択した．閲覧数上限を 2 とした場合，最適化計

[*1] https://networkx.github.io/

算が可能な期間数の上限は 9 となった．また期間数が 8 の場合にも，閲覧数上限は 2 までしか計算ができなかったため，表 3.5,3.6 からは除外している．

表 3.6 から，PV-seq(EMP)，PV-seq1，PV-seq2 のいずれも最も F1 値が高い期間数と閲覧数上限の組は，期間数 7，閲覧数上限 3 であることがわかる．またいずれの期間数と閲覧数上限の組でも，実績確率をそのまま用いる PV-seq(EMP) に比べて，順序関係の制約条件を課した PV-seq1，PV-seq2 がいずれも予測精度を向上させている．このことから，閲覧数列の順序関係に基づき，実績の商品選択確率を補正することが有効であることが示唆される．また PV-seq1，PV-seq2 については，期間数と閲覧数上限の組により F1 値が高い手法が異なることがわかる．

この期間 k と閲覧数上限 $v_{\max}$ の組合せに関する分析に基づき，以降の検証では最も予測精度が高かった $(k, v_{\max}) = (7, 3)$ と，閲覧数の影響を詳しく調べるために上限 $v_{\max}$ を増やした $(k, v_{\max}) = (5, 6)$ について実験結果を示す．

表 3.3. 期間数と閲覧数上限を変化させたときの順序関係の制約条件数

			順序関係の制約条件数					
期間数	閲覧数上限	決定変数の数	All1	All2	Naive1	Naive2	Hasse1	Hasse2
5	1	32	430	430	160	160	48	48
5	2	243	21,383	17,945	1,890	1,620	594	634
5	3	1,024	346,374	255,260	9,600	7,680	3,072	3,546
5	4	3,125	3,045,422	2,038,236	32,500	25,000	10,500	12,898
5	5	7,776	18,136,645	11,282,058	86,400	64,800	28,080	36,174
5	6	16,807	82,390,140	48,407,475	195,510	144,060	63,798	85,272
1	6	7	21	21	6	6	6	6
2	6	49	1,001	861	120	105	78	93
3	6	343	42,903	32,067	1,638	1,323	798	1,018
4	6	2,401	1,860,622	1,224,030	18,816	14,406	7,350	9,675
5	6	16,807	82,390,140	48,407,475	195,510	144,060	63,798	85,272

表 3.4. 期間数と閲覧数上限を変化させたときの最適化計算時間

			最適化計算時間 [秒]					
期間数	閲覧数上限	決定変数の数	All1	All2	Naive1	Naive2	Hasse1	Hasse2
5	1	32	0.00	0.01	0.00	0.00	0.00	0.00
5	2	243	2.32	1.66	0.09	0.07	0.03	0.02
5	3	1,024	558.22	64.35	3.41	0.71	0.13	0.26
5	4	3,125	OM	OM	24.07	13.86	1.72	5.80
5	5	7,776	OM	OM	180.53	67.34	9.71	36.94
5	6	16,807	OM	OM	906.76	522.84	86.02	286.30
1	6	7	0.00	0.00	0.00	0.00	0.00	0.00
2	6	49	0.03	0.01	0.01	0.00	0.00	0.00
3	6	343	12.80	1.68	0.20	0.03	0.05	0.02
4	6	2,401	OM	OM	8.07	4.09	2.12	2.87
5	6	16,807	OM	OM	906.76	522.84	86.02	286.30

表 3.5. 閲覧数列モデルの順序関係の制約条件数と最適化計算時間

			順序関係の制約条件数		最適化計算時間 [秒]	
期間数	閲覧数上限	決定変数の数	PV-seq1	PV-seq2	PV-seq1	PV-seq2
9	2	19,683	83,106	86,386	244.15	256.42
7	3	16,384	67,584	76,818	96.08	254.31
6	4	15,625	62,500	76,506	62.92	209.67
5	6	16,807	63,798	85,272	86.02	286.30
4	12	28,561	99,372	142,800	198.82	539.76
3	30	29,791	84,630	118,850	86.72	241.46

表 3.6. 閲覧数列モデルの予測精度

		F1 値		
期間数	閲覧数上限	PV-seq(EMP)	PV-seq1	PV-seq2
9	2	13.07%	13.40%	13.37%
7	3	13.23%	13.52%	13.53%
6	4	13.14%	13.49%	13.48%
5	6	12.90%	13.18%	13.18%
4	12	12.68%	12.93%	12.95%
3	30	12.25%	12.40%	12.40%

3.4.6　データ量に対する閲覧数列モデルの予測精度

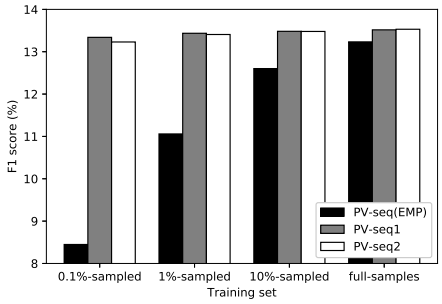
本項では，PV-seq(EMP)，PV-seq1，PV-seq2 について，学習データの 0.1% 標本 (0.1%-sampled)，1% 標本 (1%-sampled)，10% 標本 (10%-sampled)，全標本 (full-samples) を用いた場合の予測精度を検証する．また実験では，予測商品数をそれぞれ 3，5，10 と変化させて予測精度を計算した．

図 3.9 [*2]から，学習データのデータ量，予測商品数に関わらず，PV-seq(EMP) と比較して PV-seq1, PV-seq2 のほうが予測精度が高いことがわかる．特に，PV-seq(EMP) と PV-seq1, PV-seq2 の予測精度の差は学習データ量が少なくなるほど広がることがわかる．これは，2 次元確率表モデルと同様に，PV-seq(EMP) では少ない学習データに対して過剰適合が生じてしまっていたのが，閲覧数列の順序関係に基づき補正することで過剰適合が抑制されたためと考えられる．

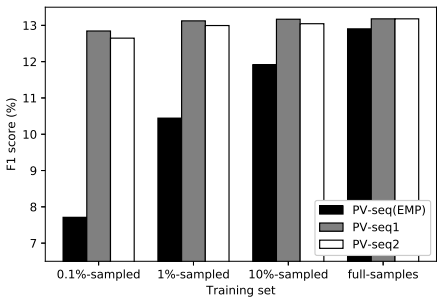
PV-seq1 と PV-seq2 の予測精度を比較すると，PV-seq(EMP) との予測精度の差と比較して差はわずかであるが，学習データが 0.1% 標本，1% 標本，10% 標本においては PV-seq1 の予測精度が高い場合が多く，全標本の場合には PV-seq2 の予測精度が高い場合もあることが

[*2] 図 3.9, 3.10, 3.11 では基線が 0 でないことに留意されたい．

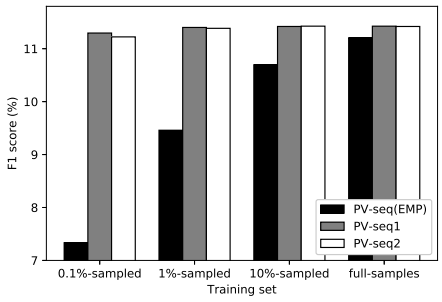
わかる．このことから，学習データが少ない場合には，閲覧数列の最新度の単調性 1 を利用するのがよく，学習データが多い場合には閲覧数列の最新度の単調性 2 を利用することで予測精度が改善する場合もあると言える．この結果から，以降の 2 次元確率表モデルとの比較，一般的な機械学習モデルとの比較では PV-seq1 を閲覧数列モデルとして利用して比較する．



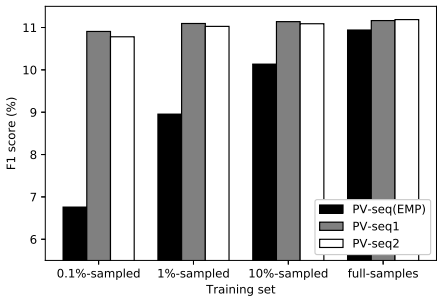
(a) $N = 3,\ (k, v_{\max}) = (7, 3)$



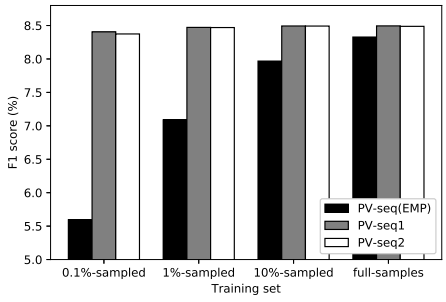
(b) $N = 3,\ (k, v_{\max}) = (5, 6)$



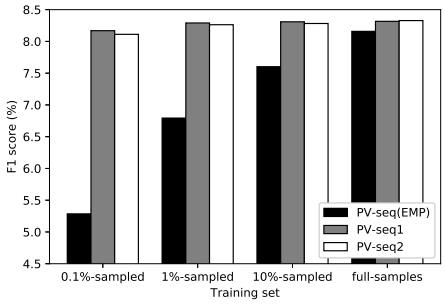
(d) $N = 5,\ (k, v_{\max}) = (7, 3)$



(e) $N = 5,\ (k, v_{\max}) = (5, 6)$



(g) $N = 10,\ (k, v_{\max}) = (7, 3)$



(h) $N = 10,\ (k, v_{\max}) = (5, 6)$

図 3.9. 学習データのデータ量の変化に対する PV-seq(EMP)，PV-seq1，PV-seq2 の予測精度

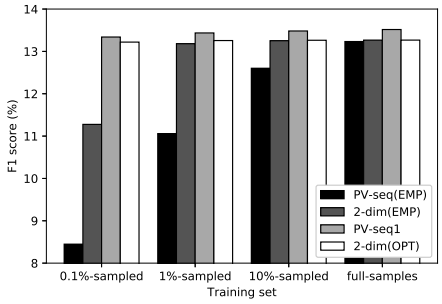
3.4.7　2 次元確率表モデルと閲覧数列モデルの比較

本項では，既存手法である 2 次元確率表モデル（2-dim）と閲覧数列モデル（PV-seq1）の予測精度を比較する．3.4.6 節と同様に，学習データを 0.1% 標本，1% 標本，10% 標本，全標本と変化させ，予測商品数をそれぞれ 3，5，10 と変化させて検証した．図 3.10 では，2 次元確率表モデル，閲覧数列モデルそれぞれについて，商品選択確率の過去の実績値をもとに予測するものを 2-dim(EMP)，PV-seq(EMP)，形状制約を課した最適化問題により補正した推定値をもとに予測するものを 2-dim(OPT)，PV-seq1 と表す．

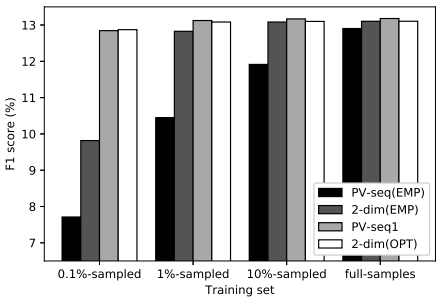
図 3.10 から，商品選択確率の過去の実績値をもとに予測を行なった 2-dim(EMP)，PV-seq(EMP) について，いずれの場合も 2-dim(EMP) が予測精度が高いことがわかる．これは 3.2.1 節で述べたように，区別可能な閲覧履歴のパターン数が，2 次元確率表は $kv_{\max}$ 個，閲覧数列は $(1+v_{\max})^k-1$ 個のため，閲覧数列のそれぞれのパターンに対して割り当てられるデータ量が少なくなってしまい過剰適合が生じたためであると考えられる．特に，2-dim(EMP) と PV-seq(EMP) の予測精度の差は，データ量が小さくなるほど大きいことがわかる．

一方で 2-dim(OPT)，PV-seq1 については，0.1% 標本のときは 2-dim(OPT) の予測精度が高い場合が多いが，1% 標本，10% 標本，全標本のときは PV-seq1 の予測精度が高い場合が多い．このことから，データ量が十分に確保できる場合には PV-seq1 の予測精度が高く，データ量が少ない場合には 2-dim(OPT) の予測精度が高い場合もあるということが示唆される．

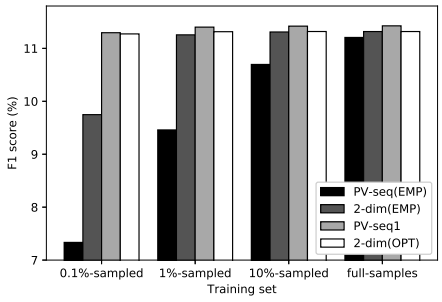
これは，2 次元確率表モデルに比べて多くの閲覧パターンを区別できる閲覧数列モデルでは，各パターンに対して十分にデータ量が確保できた場合に予測精度が高くなったと考えられ，顧客の閲覧行動を低次元に縮約せず，より高次元に扱うことで予測精度が改善する，という仮説を支持する結果と考えられる．この結果から，実績値を利用した PV-seq(EMP) においても，今回の学習データよりもさらにデータ量が大きいデータの場合には，2-dim(EMP) よりも予測精度が高くなることが期待される．



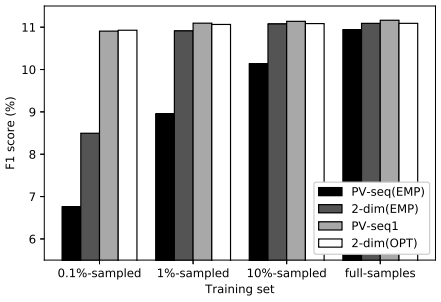
(a) $N = 3,\ (k, v_{\max}) = (7, 3)$



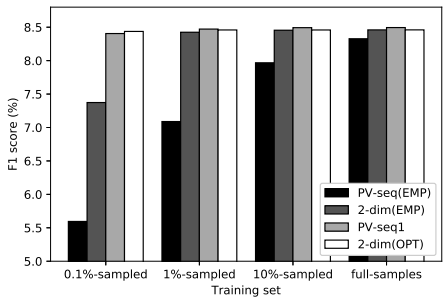
(b) $N = 3,\ (k, v_{\max}) = (5, 6)$



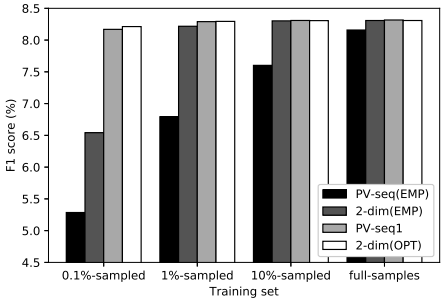
(d) $N = 5,\ (k, v_{\max}) = (7, 3)$



(e) $N = 5,\ (k, v_{\max}) = (5, 6)$



(g) $N = 10,\ (k, v_{\max}) = (7, 3)$



(h) $N = 10,\ (k, v_{\max}) = (5, 6)$

図 3.10. 学習データのデータ量の変化に対する 2-dim(EMP)，PV-seq(EMP)，2-dim(OPT)，PV-seq1 の予測精度

3.4.8 機械学習の手法と閲覧数列モデルの比較

本項では，同じ閲覧数列を特徴量として利用したときの，閲覧数列モデル PV-seq1 と表 3.1 の機械学習モデルとの予測精度を比較する．

図 3.11 では，3.4.2 項で述べた方法で学習した機械学習モデルを用いた予測 LR(EMP)，

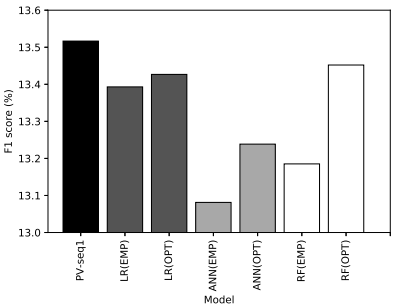
ANN(EMP)，RF(EMP) と，機械学習モデルの予測値に対して閲覧数列の順序関係 1 を考慮して補正したモデルを用いた予測 LR(OPT)，ANN(OPT)，RF(OPT) について閲覧数列モデル PV-seq1 と比較した．また，3.4.5 項で利用した $(k, v_{\max}) = (7, 3)$ と $(5, 6)$ の閲覧数列を入力データとしてそれぞれの手法で予測を行い，予測商品数をそれぞれ 3，5，10 と変化させて検証した．

まず，閲覧数列モデル PV-seq1 と，機械学習モデル LR(EMP)，ANN(EMP)，RF(EMP) の予測精度を比較すると，図 3.11 から，$(k, v_{\max}) = (5, 6)$ の予測商品数 10 の場合以外のすべての場合で，PV-seq1 が機械学習モデルに比べて予測精度が高いことがわかる．閲覧数列モデルは，閲覧数列に関する仮説を頻度の単調性，最新度の単調性として明示的に形状制約として課して選択確率を推定する方法であるが，同じ閲覧数列の入力に対して機械学習モデルを利用して学習させたものよりも概ね予測精度が高いことから，顧客の閲覧行動に関する仮説として機械学習のモデルで実績値をもとに学習されたものよりも予測精度向上の観点で妥当な仮説であったという示唆が得られる．

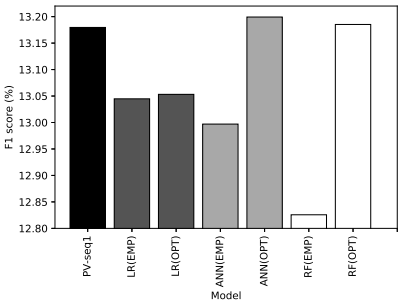
つづいて，機械学習モデルを用いて算出した予測値に対して閲覧数列の順序関係に基づき補正した LR(OPT)，ANN(OPT)，RF(OPT) との予測精度を比較すると，すべての場合で機械学習モデルを最適化モデルでの補正なしで用いた予測よりも，予測精度が向上している，または同じであることがわかる．このことから，閲覧数列の順序関係に基づく形状制約による補正は，機械学習モデルの予測に対する後処理としても，予測精度向上の観点で有効であることが示唆される．

また，形状制約による補正によって予測精度が変化しなかった $(k, v_{\max}) = (5, 6)$ の予測商品数 10 の LR(EMP) では，既にすべての閲覧数列について形状制約が満たされていた．LR の学習においては，各期間の閲覧数に対する偏回帰係数がパラメータとなっており，学習データから近い期間について偏回帰係数が大きくなるようにパラメータの推定ができれば今回仮定した閲覧数列の順序関係 1 は満たされる．$(k, v_{\max}) = (5, 6)$ の予測商品数 10 以外でも，他の機械学習モデルに比べて形状制約の補正による予測精度の改善幅が小さいのは，多くの部分で既に閲覧数列の順序関係が満たされているためであると考えられる．

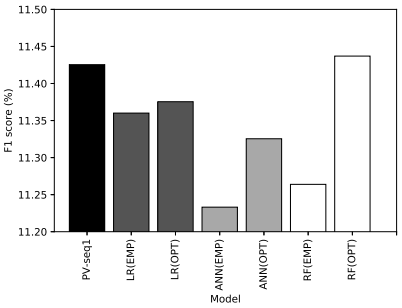
一方で，ANN と RF では LR に比べて大きく予測精度が改善している．これは ANN や RF が非線形なモデルであり，LR と比較して閲覧数列の順序関係が満たされない部分が多いためであると考えられる．機械学習モデルは一般的に予測結果の解釈が難しいことが指摘されており，直感に反する結果が出力されることもある．本章の提案手法のような形状制約を機械学習モデルの後処理として課すことによって，最終的な予測結果で仮定が満たされることを担保できるようになる．提案手法の顧客行動の仮説をもとにした形状制約による商品選択確率の推定は，そういった予測結果に対して解釈性が求められるような状況でも有用である．



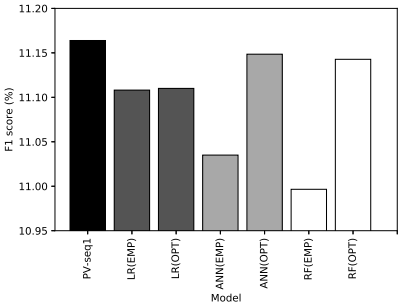
(a) $N = 3,\ (k, v_{\max}) = (7, 3)$



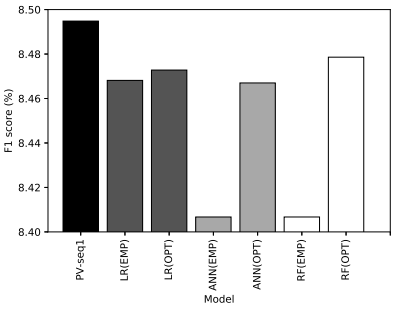
(b) $N = 3,\ (k, v_{\max}) = (5, 6)$



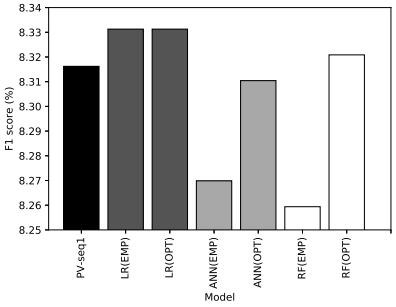
(d) $N = 5,\ (k, v_{\max}) = (7, 3)$



(e) $N = 5,\ (k, v_{\max}) = (5, 6)$



(g) $N = 10,\ (k, v_{\max}) = (7, 3)$



(h) $N = 10,\ (k, v_{\max}) = (5, 6)$

図 3.11. 閲覧数列モデル PV-seq1 と機械学習の手法の予測精度

3.4.9　形状制約モデルと形状正則化モデルの比較

本項では，単調性を制約条件として課しモデルのパラメータを推定する閲覧数列モデル（PV-seq1）と単調性の違反に対する罰則を目的関数に加えてパラメータを推定する形状正則化閲覧数列モデル（PV-seq(Reg)）を比較する．

目的関数に対する形状正則化モデルの罰則項の強さを変化させるため，式 (3.7) の正則化パラメータ λ の大きさを変化させて実験した．また，λ の大きさは，学習データや閲覧数列モデルの期間数，閲覧数上限，順序関係の制約条件数に応じて以下のように調整した：

$$\lambda = \frac{\sum_{\boldsymbol{v}\in\Gamma} n_{\boldsymbol{v}}}{(\text{制約条件数})(1+v_{\max})^k}\ \gamma.$$

ここで，γ は正則化パラメータ λ の大きさを調整するためのハイパーパラメータである．形状正則化モデルでは，最適化問題の求解に要する計算機のメモリが多かったため，3.4.6 項，3.4.7 項，3.4.8 項の数値実験の閲覧数列の閲覧数上限からそれぞれ 1 減らし，$(k, v_{\max})$ が $(7,2),(5,5)$ の組合せについて検証した．γ の値は 1000 の場合に $(7,2),(5,5)$ の組合せのいずれも順序関係の制約条件をすべて満たすようになったため，1000 を基準に 0.1 倍ずつ $1000, 100, 10, 1, 0.1$ と 5 つの場合について検証した．

表 3.7 は，$(k, v_{\max})$ が $(7,2),(5,5)$ の場合の閲覧数列の順序関係 1（式 (3.4)）に基づく制約条件数と実績の閲覧数列モデル（PV-seq(EMP)），形状制約閲覧数列モデル（PV-seq1），形状正則化モデル（PV-seq(Reg)）での制約条件の違反数を示している．まず，実績の閲覧数列モデルでは順序関係の制約条件がそれぞれ半数以上満たされているが，満たされていない部分もある．形状正則化モデルでは，γ の値を 0.1 から増加するにつれて順序関係の制約条件の違反数は減少している．また形状制約モデルは明示的にすべての制約条件を満たすように最適化問題を解いているため，制約条件の違反数は 0 である．順序関係の制約条件の違反数の観点では $(k, v_{\max}) = (7,2)$ における PV-seq(Reg) の $\gamma = 100, 1000$ と PV-seq1，$(k, v_{\max}) = (5,5)$ における PV-seq(Reg) の $\gamma = 1000$ と PV-seq1 でいずれも 0 で同じであるが，順序関係を仮定しない部分の推定値の順序関係については異なることがあるため，それぞれのモデルの予測値は同じとは限らない．

表 3.7. 正則化パラメータと順序関係の制約条件の違反数

			順序関係の制約条件の違反数						
				PV-seq(Reg) $\gamma =$					
期間数	閲覧数上限	制約条件数	PV-seq (EMP)	0.1	1	10	100	1000	PV-seq1
7	2	7,290	3,072	1,884	276	5	0	0	0
5	5	28,080	13,239	2,801	2,755	425	16	0	0

表 3.8 は，$(k, v_{\max})$ が $(7,2),(5,5)$ の場合の形状制約閲覧数列モデル（PV-seq1），形状正則化モデル（PV-seq(Reg)）での最適化問題 (3.6),(3.7) の計算時間を示している．表 3.8 から，単調性を制約条件を明示的に課した形状制約モデルは，単調性の違反を罰則項として目的関数に加える形状正則化モデルに比べて計算時間が短いことがわかる．

表 3.8. 形状制約モデルと形状正則化モデルの最適化計算時間

		最適化計算時間 [秒]					
		PV-seq(Reg) $\gamma =$					
期間数	閲覧数上限	0.1	1	10	100	1000	PV-seq1
7	2	1.87	1.84	2.16	5.11	7.16	0.86
5	5	73.85	31.83	24.37	20.63	56.37	9.71

表 3.9 は，全標本のテストデータに対して，$(k, v_{\max})$ が $(7,2),(5,5)$ の場合の形状制約閲覧数列モデル，形状正則化モデルの商品選択予測精度を検証した結果である．数値実験では，予測商品数を 3,5,10 と変化させてそれぞれ検証した．

表 3.9 から，概ね順序関係の制約条件の違反数が減少するにつれて予測精度が向上していることがわかる．また，形状制約閲覧数列モデルに比べて形状正則化モデルの予測精度が高い部分もあることがわかる．

表 3.9. 形状制約モデルと形状正則化モデルの予測精度

			F1 値						
				PV-seq(Reg) $\gamma =$					
期間数	閲覧数上限	予測商品数	PV-seq (EMP)	0.1	1	10	100	1000	PV-seq1
7	2	3	13.16%	13.28%	13.32%	13.30%	13.33%	13.33%	13.32%
		5	11.26%	11.37%	11.39%	11.39%	11.40%	11.41%	11.39%
		10	8.42%	8.50%	8.50%	8.51%	8.50%	8.50%	8.51%
5	5	3	12.99%	13.03%	13.15%	13.20%	13.23%	13.23%	13.23%
		5	10.98%	11.02%	11.13%	11.19%	11.20%	11.17%	11.19%
		10	8.18%	8.22%	8.29%	8.32%	8.31%	8.30%	8.31%

3.3.1 項で述べたように，形状正則化モデルは表現力の向上により形状制約モデルよりも予測精度が高くなる部分があるという利点を確認した．一方で，正則化パラメータを設定する必要があることや表 3.8 の結果のように計算時間が増加するという欠点も確認した．

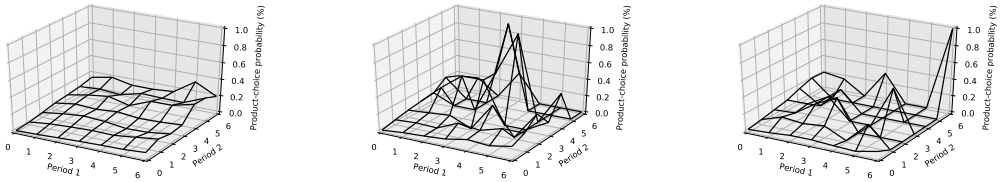
3.4.10 閲覧数列モデルの商品選択確率の可視化

本項では，PV-seq1 と PV-seq2 により推定された商品選択確率と，実績値の商品選択確率 PV-seq(EMP) を可視化し観察する．図 3.12 は，学習データを全標本利用してモデルのパラメータを推定したもの，図 3.13 は，10% 標本の学習データからモデルのパラメータを推定し

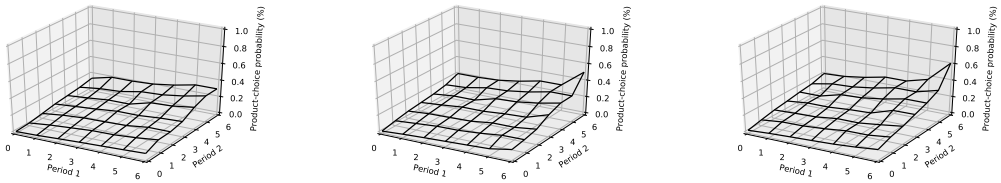
たものである．それぞれの図は $(k, v_{\max}) = (5, 6)$ の閲覧数列について，期間 $k = 4, 5$ の閲覧数 v_4, v_5 が 0，$k = 3$ の閲覧数 v_3 を 0,1,2 としたときに，期間 $k = 1, 2$ の閲覧数 v_1, v_2 をそれぞれ 0 から 6 の間で変化させた商品選択確率の推定値を表す．

まず，図 3.12 の $v_3 = 0$ の PV-seq(EMP)，PV-seq1，PV-seq2 を比較すると，v_1, v_2 がともに小さい部分に関してはいずれも同じような形状となっているが，v_1, v_2 がともに大きい部分に関しては PV-seq(EMP) は閲覧数が多くなったときに商品選択確率が減少することを示すような凹凸がみられ，閲覧数列の頻度の単調性が満たされていない部分があることがわかる．閲覧数列の頻度の単調性の違反については，PV-seq(EMP) について，$v_3 = 1, 2$ の部分で更に顕著に観察できるようになる．頻度の単調性の違反が多くなる要因としては，図 3.12 (a) が表す「3 期前に閲覧されず 1,2 期前に閲覧される商品」に比べて，図 3.12 (b)，(c) が表す「3 期前に 1 回または 2 回閲覧されて，1,2 期前に閲覧される商品」の学習データでのデータ量が少なく，その少ないデータ量をもとに集計された確率値となっているためと考えられる．同様に，10% 標本の学習データからパラメータを推定した図 3.12 では，(a)，(b)，(c) のそれぞれで図 3.12 よりも大きな凹凸が観察できる．

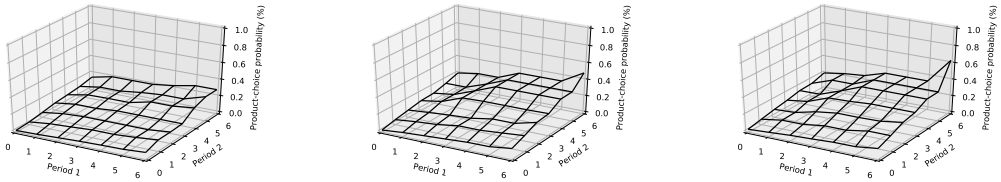
PV-seq1 と PV-seq2 については，閲覧数列の順序関係を満たすように形状制約を課しているため，ともに PV-seq(EMP) のような閲覧数列の単調性を違反するような凹凸はなくなっている．いずれの図からもデータ量によらず v_1, v_2 がともに小さい部分で商品選択確率の推定値が小さく，v_1, v_2 がともに大きい部分にかけて大きくなっていく，という単調性が満たされていることが確認できる．PV-seq1 と PV-seq2 の形状を比較すると，PV-seq1 は，閲覧数列の最新度の単調性 1 における「閲覧数列の成分の値を直近の期間へ 1 単位移動させる操作」により，図の左奥側の $v_1 = 0, v_2 = 6$ の部分から，右前側 $v_1 = 6, v_2 = 0$ にかけて単調に商品選択確率が増加している．一方で，PV-seq2 では，閲覧数列の最新度の単調性 2 での「閲覧数列の遠い期間の大きい閲覧数の成分と近い期間の小さい閲覧数の成分を交換する操作」から必ずしも右前側 $v_1 = 6, v_2 = 0$ にかけて単調に商品選択確率が増加するとは限らず，特に $v_1 = 3, v_2 = 3$ の中央辺りに商品選択確率が高くなっている部分がある．3.2.2 節でも例として挙げたように，$v_1 = 3, v_2 = 3$ のように継続して閲覧が行われていることを表す閲覧数列と $v_1 = 6, v_2 = 0$ のように直近で多く閲覧が行われている閲覧数列に関して，PV-seq1 では直近の閲覧数が多い閲覧数列に対して商品選択確率が高くなるように，明示的に形状制約を課しており，PV-seq2 ではそのような形状制約を明示的には課さず，データの集計値の表す傾向に委ねる点が PV-seq1 と PV-seq2 の差異である．そのため，データ量が十分に確保できる場合には PV-seq2 のように，明示的に形状制約を課さず予測精度が向上する場合があったが，データ量が少ない場合には PV-seq1 のように明示的に形状制約を課し，過剰適合を抑えることで予測精度が向上した，と解釈できる．



(a) PV-seq(EMP), $v_3 = 0$ (b) PV-seq(EMP), $v_3 = 1$ (c) PV-seq(EMP), $v_3 = 2$

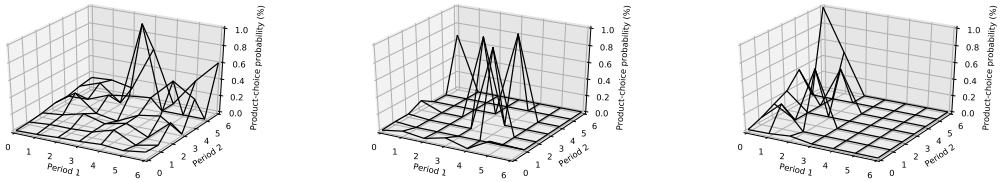


(d) PV-seq1, $v_3 = 0$ (e) PV-seq1, $v_3 = 1$ (f) PV-seq1, $v_3 = 2$

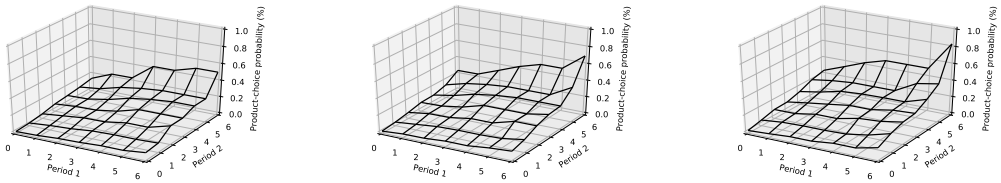


(g) PV-seq2, $v_3 = 0$ (h) PV-seq2, $v_3 = 1$ (i) PV-seq2, $v_3 = 2$

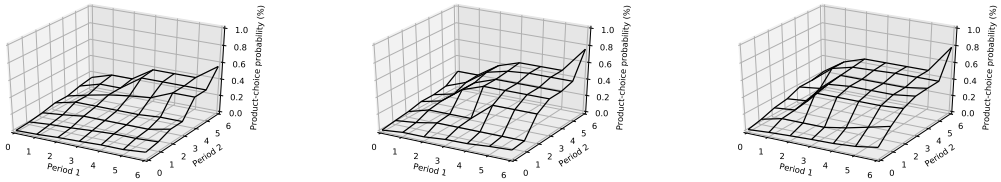
図 3.12. $(k, v_{\max}) = (5, 6)$ の全標本を学習に利用した場合の PV-seq(EMP)，PV-seq1，PV-seq2 の閲覧数列モデルの商品選択確率の可視化



(a) PV-seq(EMP), $v_3 = 0$　(b) PV-seq(EMP), $v_3 = 1$　(c) PV-seq(EMP), $v_3 = 2$



(d) PV-seq1, $v_3 = 0$　(e) PV-seq1, $v_3 = 1$　(f) PV-seq1, $v_3 = 2$



(g) PV-seq2, $v_3 = 0$　(h) PV-seq2, $v_3 = 1$　(i) PV-seq2, $v_3 = 2$

図 3.13. $(k, v_{\max}) = (5, 6)$ の 10% 標本のデータを学習に利用した場合の PV-seq(EMP)，PV-seq1，PV-seq2 の閲覧数列モデルの商品選択確率の可視化

3.5　まとめ

本章では，商品閲覧の時系列 (閲覧数列) に対して形状制約つき最適化問題を解くことで商品選択確率を推定する手法を提案した．また閲覧数列について，半順序の簡約表現であるハッセ図の構造を利用して最適化計算を効率化する方法を提案した．数値実験の結果，検証に利用したデータに対して，本章の提案手法である閲覧数列モデルは，期間数や閲覧数上限が同じとき，既存研究の形状制約モデル（2 次元確率表モデル）と比較して予測精度が高いことがわかった．そして，同一の閲覧数列の特徴量を用いて学習を行ったとき，機械学習の手法（L2 正則化ロジスティック回帰，人工ニューラルネットワーク，ランダムフォレスト）と比較して予測精度が高いことがわかった．また，閲覧数列モデルによる商品選択確率の補正は，機械学習の手法による予測値に対しても適用可能で，それによって予測精度が改善することを確認した．閲覧数列モデルの商品選択確率を推定するための最適化計算を行うとき，ハッセ図の構造を利用して冗長な制約条件を削除することで計算が高速化することがわかった．

本章では，閲覧数列に期待される性質として，閲覧数列の頻度に関する単調性と，最新度に

関する二つの単調性を利用したが，それ以外の閲覧数列の性質を探索することが今後の課題として挙げられる．また，本研究では期間数や閲覧数の閾値について，それぞれ 1 日ごと，1 閲覧ごとと設定して期間数や閲覧数の上限を設定しているが，たとえば複数日で 1 期間，複数回閲覧で 1 閲覧区間というように区分を調整することで，より適切に顧客の商品閲覧の差異を捉え，予測精度が改善する可能性がある．そして，閲覧数列以外の特徴量については，第 2 章で提案したように潜在クラスとして考慮することも可能であり，これによっても潜在クラス型確率表モデルと同様に予測精度が改善する可能性がある．

第 4 章

反復選択商品に対する形状制約比例ハザードモデル

4.1　はじめに

本章では，継続購買が想定される商品に対する商品選択行動に着目する．顧客の商品選択行動の性質として，一度選択すると別の商品に切り替える場合に新たな探索や，その商品への習熟のためのコストが生じるため，次回の選択機会に顧客が同じ商品を選択する可能性が高くなる，というものがある [37]．このような顧客が過去に選択した商品を反復して選択する性質を利用することで，商品選択確率を高精度で予測できる可能性がある．

本章では特に，新たな店舗を探索するコストや，ヘアスタイルの好みや髪質などを説明するコストが生じるために，同じ店舗に継続して来店する可能性が高いヘアサロンの再来店の予測を対象とする．再来店予測の方法として，累積来店回数に対する再来店率の単調性と凹性を考慮した形状制約比例ハザードモデルを提案する．

本章の構成は以下の通りである．4.2 節では，再来店予測に利用する比例ハザードモデルの関連研究と，顧客の再来店の選択行動に関する仮説について述べる．4.3 節では，ヘアサロンの顧客の再来店を，累積来店回数に対する再来店率の単調性と凹性を考慮した比例ハザードモデルにより推定する方法を提案する．4.4 節では数値実験を通して，提案手法の有効性に関して検証する．また提案手法によって得られたモデルのパラメータを主成分分析により縮約し可視化することで，各店舗の来店状況を分析する．最後に 4.5 節で第 4 章のまとめと課題を述べる．

4.2　比例ハザードモデルを用いた再選択予測

本節では，再選択予測に利用される比例ハザードモデルの関連研究と，顧客の再選択行動に関する仮説について述べる．

4.2.1　比例ハザードモデルを用いた再選択予測の関連研究

顧客の再選択に与える影響を推定する既存研究では，比例ハザードモデル [14] を適用した研究が数多く行われている [36, 70, 77]．小西 [77] は比例ハザードモデルを用いてヘアサロンの顧客の再来店に影響を与える要因の分析を行っている．Van，Lariviere [70] は金融機関の顧客離反に影響を与える要因について分析を行っている．Kapoor ら [36] はウェブサービスにおける顧客の再来訪時間の推定を行っている．

しかし，これらの再選択予測において，利用可能なデータが十分な量でない場合には，推定されたモデルが少ないデータに対して過剰適合し，誤った解釈が行われてしまう恐れがある．そこで，本節でも，前章までのアプローチと同様に，実際の顧客の選択行動を記録したデータを解析し，顧客の再来店に関する特徴の探索を行う．そして，探索により得られた仮説をもとに形状制約を課したモデルを提案し，予測精度を改善することを試みる．

4.2.2　利用データ

本章では経営科学系研究部会連合協議会主催, 平成 29 年度データ解析コンペティションで提供されたヘアサロンチェーンのデータを利用した．

提供された 2015 年 7 月から 2017 年 6 月までの 2 年間のデータのうち, 新規店舗や長期間の休業期間があった店舗を除いた 10 店舗の POS データを分析対象とした．この POS データのうち, 会計履歴と顧客のデモグラフィックをまとめたデータ (顧客マスタ) を用いて分析する．会計履歴では会計日, 顧客 ID, 売上, 各顧客の累積来店回数などがまとまっている．顧客マスタでは各顧客の郵便番号, 性別, 年代などがまとまっている．

4.2.3　顧客の再選択行動に関する仮説

本項では，4.2.2 項のヘアサロンの来店データの解析から，累積来店回数と再来店の関係について着目する．図 4.1 は，データ期間の前半 2015 年 7 月 1 日から 2016 年 6 月 30 日の 1 年間に上述の 10 店舗に来店した顧客を対象として, 翌 365 日以内の来店を再来店顧客と定義し, 顧客の累積来店回数と再来店率の関係について集計を行ったものである．

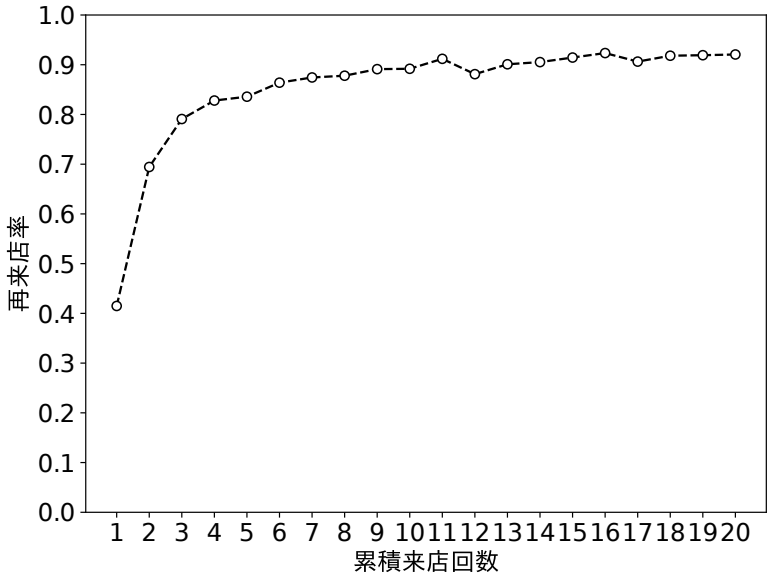


図 4.1. 顧客の累積来店回数と再来店率の関係

図 4.1 から累積来店回数と再来店率の関係について以下二つの傾向があることがわかる：

1. 単調性：累積来店回数が多い顧客ほど再来店率は高い.
2. 凹性：累積来店回数が多くなるにつれて再来店率の増分は逓減する.

本章では，比例ハザードモデルの推定における過剰適合抑制のため，来店数（Frequecy）と再来店率の関係に着目し，その関係を制約条件として課すことで過剰適合を抑制し，予測性能を向上させることを考える.

次節では, このような単調性と凹性を考慮して累積来店回数が再来店に与える影響を分析するための形状制約比例ハザードモデルについて説明する.

4.3　形状制約比例ハザードモデル

本節では Deng et al. [16] にしたがって，本章で提案する形状制約比例ハザードモデルとその推定方法について述べる.

4.3.1　比例ハザードモデル

ここでは店舗ごとの再来店傾向を顧客の再来店間隔をもとに分析するため, 比例ハザードモデル [14] を用いて再来店間隔をモデル化する. まず来店時から次に来店するまでの期間 (来店間隔) を確率変数 T とする. 来店時から時点 t まで来店しなかった顧客が次の瞬間に来店する確率 (ハザード率) を $\lambda(t)$ と表す. すなわち

$$\lambda(t) = \lim_{\Delta \to 0} \frac{1}{\Delta} \mathrm{P}\{t \leq T \leq t + \Delta \mid T \geq t\}$$

である. ここで $\mathrm{P}\{A \mid B\}$ は事象 B が生起したもとで事象 A が生起する条件つき確率を表す. 比例ハザードモデルではハザード率に対して以下の形を仮定する:

$$\lambda(t \mid x) = \lambda_0(t) \exp(\boldsymbol{\beta}^\top \boldsymbol{x}). \tag{4.1}$$

ここで $\boldsymbol{x} = (x_1, x_2, \ldots, x_M)^\top$ は M 個の説明変数であり, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_M)^\top$ は対応する偏回帰係数を表す. また $\lambda_0(t)$ は説明変数 $\boldsymbol{x}$ に依存しないハザード率を表し, ベースラインハザードとよぶ. なおハザード率とベースラインハザードの比の対数は

$$\log\left(\frac{\lambda(t \mid x)}{\lambda_0(t)}\right) = \boldsymbol{\beta}^\top \boldsymbol{x} \tag{4.2}$$

であり, $\boldsymbol{x}$ に関する線形関数であることに注意する. なお以降では簡単のため, (4.2) を対数ハザード比とよぶことにする.

本章では顧客の累積来店回数が来店間隔に与える影響を分析するため, 比例ハザードモデルの説明変数には累積来店回数と顧客属性を用いる. ここで顧客属性を説明変数として用いるのは, 累積来店回数だけでなく顧客属性も来店間隔に影響を与えているという仮定にもとづく.

累積来店回数は通常一つの量的変数として扱われるが, この場合累積来店回数と対数ハザード比との間に線形の関係を仮定してしまう. しかし 4.2.3 節で累積来店回数と再来店率の間に単調性と凹性という非線形の関係が見られたように, 累積来店回数と対数ハザード比の間にも同様の非線形の関係があると考えられる. そこで本章ではその非線形な関係を捉えるため, 累積来店回数は量的変数としてではなく, 回数別のダミー変数として説明変数に用いる. いま説明変数の添え字集合を, $D = \{1, 2, \ldots, K\}, U = \{K+1, K+2, \ldots, M\}$ (ただし $1 \leq K < M$) と二つに分割し, 累積来店回数に関するダミー変数を $\boldsymbol{x}_D := (x_k)_{k \in D}$, 顧客属性に関する説明変数を $\boldsymbol{x}_U := (x_k)_{k \in U}$ とする. そして説明変数全体を $\boldsymbol{x} = (\boldsymbol{x}_D^\top, \boldsymbol{x}_U^\top)^\top$ と表す. ここで $\boldsymbol{x}_D$ は

$$x_k = \begin{cases} 1 & (\text{累積来店回数が } k \text{ の場合}) \\ 0 & (\text{そうでない場合}) \end{cases} \qquad (k \in D)$$

とする. 説明変数 $\boldsymbol{x} = (\boldsymbol{x}_D^\top, \boldsymbol{x}_U^\top)^\top$ に対応する偏回帰係数を $\boldsymbol{\beta} = (\boldsymbol{\beta}_D^\top, \boldsymbol{\beta}_U^\top)^\top$ と表す.

累積来店回数のダミー変数 x_k $(k \in D)$ に対応する偏回帰係数 β_k $(k \in D)$ に着目すると, β_k の値が大きければ累積来店回数 k 回の顧客の来店間隔が短い (すなわち再来店しやすい) ことを表す. 以降では β_k $(k \in D)$ の指数をとった値 $\exp(\beta_k)$ $(k \in D)$ を累積来店回数が k 回の顧客の定着度とよぶことにする.

4.3.2 部分尤度関数

ここでは比例ハザードモデルの推定に用いる部分尤度関数について説明する. ある一つの店舗に着目し, データの期間内で観測された N 個の来店に関する標本の集合を $\{(t^i, \delta^i, \boldsymbol{x}^i)\}_{i=1}^N$

と表す．ここで t^i は当該の来店から同一の顧客がデータ期間内に再来店するまでの期間 (来店間隔) を表す．ただしデータ期間内に再来店がなかった場合 (すなわち打ち切りデータの場合), t^i は来店時からデータ期間の最後の時点までの期間とする．δ^i はデータ期間内に再来店があったか否かを表し, データ期間内に再来店があった場合 1, そうでなければ 0 をとる．$\boldsymbol{x}^i = ((\boldsymbol{x}^i_D)^\top, (\boldsymbol{x}^i_U)^\top)^\top$ はその標本に対応する説明変数である．また各標本の来店間隔 t^i に対し, 来店間隔が t^i 以上となる標本の集合 (リスク集合) を $K(t^i)$ と表す．すなわち

$$K(t^i) = \{j \mid t^j \geq t^i\}$$

である．

いま来店間隔 t^i $(i = 1, 2, \ldots, N)$ がすべて異なる値で $t^1 < t^2 < \cdots < t^N$ と昇順で並べられているものとする．このときリスク集合 $K(t^i)$ が与えられたもとで説明変数 $\boldsymbol{x}^i$ の標本の来店間隔が t^i となる条件つき確率は

$$\frac{\lambda_0(t^i)\exp(\boldsymbol{\beta}^\top \boldsymbol{x}^i)}{\sum_{j\in K(t^i)} \lambda_0(t^i)\exp(\boldsymbol{\beta}^\top \boldsymbol{x}^j)} = \frac{\exp(\boldsymbol{\beta}^\top \boldsymbol{x}^i)}{\sum_{j\in K(t^i)} \exp(\boldsymbol{\beta}^\top \boldsymbol{x}^j)} \tag{4.3}$$

と表される．比例ハザードモデル (4.1) の部分尤度関数はデータ期間内で再来店のある標本 (すなわち $\delta^i = 1$ となる標本) について式 (4.3) を掛け合わせた

$$\prod_{i:\delta^i=1} \frac{\exp(\boldsymbol{\beta}^\top \boldsymbol{x}^i)}{\sum_{j\in K(t^i)} \exp(\boldsymbol{\beta}^\top \boldsymbol{x}^j)} \tag{4.4}$$

として定義される．したがって比例ハザードモデルの対数部分尤度関数は式 (4.4) の対数をとって以下のようになる:

$$\begin{aligned} LL(\boldsymbol{\beta}) &:= \log\left[\prod_{i:\delta^i=1} \frac{\exp(\boldsymbol{\beta}^\top \boldsymbol{x}^i)}{\sum_{j\in K(t^i)} \exp(\boldsymbol{\beta}^\top \boldsymbol{x}^j)}\right] \\ &= \sum_{i:\delta^i=1} \boldsymbol{\beta}^\top \boldsymbol{x}^i - \sum_{i:\delta^i=1} \log\left[\sum_{j\in K(t^i)} \exp(\boldsymbol{\beta}^\top \boldsymbol{x}^j)\right]. \end{aligned}$$

比例ハザードモデルの推定では対数部分尤度関数 $LL(\boldsymbol{\beta})$ を最大化する最適化問題を解いて偏回帰係数を推定する．

4.3.3　パラメータ推定

4.2 節で述べた通り, 顧客の累積来店回数と再来店率の間には単調性と凹性の関係があることが確認された．これら単調性, 凹性の関係は, 累積来店回数のダミー変数に対応する偏回帰係数 $\boldsymbol{\beta}_D$ についても同様に満たされると考えられる．そこで本章では偏回帰係数 $\boldsymbol{\beta}_D$ に対して, 累積来店回数に関する単調性と凹性の制約を課してモデルのパラメータを推定する．

まず偏回帰係数 β_k $(k \in D)$ が累積来店回数 k に対して単調増加する性質 (単調性) は以下の制約として表される:

$$\beta_k \leq \beta_{k+1} \quad (k = 1, 2, \ldots, K-1).$$

同様に, 偏回帰係数 β_k $(k \in D)$ の増分が累積来店回数 k に対して単調減少する性質 (凹性) は以下の制約として表される:

$$\beta_{k+1} - \beta_k \geq \beta_{k+2} - \beta_{k+1} \quad (k = 1, 2, \ldots, K-2).$$

これらの制約を課して対数部分尤度関数を最大化する最適化問題は以下のようになる:

$$\left| \begin{array}{ll} \text{maximize} & LL(\boldsymbol{\beta}) \\ \text{subject to} & \beta_k \leq \beta_{k+1} \quad (k = 1, 2, \ldots, K-1), \\ & \beta_{k+1} - \beta_k \geq \beta_{k+2} - \beta_{k+1} \quad (k = 1, 2, \ldots, K-2). \end{array} \right. \tag{4.5}$$

式 (4.5) は最大化する目的関数が凹関数で実行可能領域が凸集合となる凸最適化問題であり, 逐次二次計画法 [54] などを用いれば大域的最適解を得ることができる.

なお本章では来店間隔を月単位で計算しているため, モデルの推定に用いる標本の集合は来店間隔の等しい標本 (タイデータ) を含む. そこで実際にモデルを推定する際は対数部分尤度関数 $LL(\boldsymbol{\beta})$ に Breslow の近似 [6] を適用した.

4.4 数値実験

本節では, 4.3 節で説明した形状制約比例ハザードモデルの予測性能を検証する. また，推定されたモデルのパラメータを主成分分析により縮約し可視化することで，各店舗の来店状況を相対化し，実務的な示唆を提示する.

4.4.1 実験設定

提供された 2015 年 7 月 1 日から 2017 年 6 月 30 日までの 2 年間のデータを用いて, 2015 年 7 月 1 日から 2016 年 6 月 30 日, 2015 年 8 月 1 日から 2016 年 7 月 31 日というように, 1 ヶ月ずつ期間をずらした 12 ヶ月間の 12 データセットを作成した. 検証においては, それぞれのデータセットの中でデータをランダムに 3 等分し, 一つをテストデータ, 残りを学習用データとして検証を 3 回繰り返す 3 分割交差検証を行った. 評価指標としてテストデータに対する平均対数部分尤度を用いる. ここで平均対数部分尤度は対数部分尤度を標本数で割った値である. テストデータに対する平均対数部分尤度が高いモデルはモデルの推定に用いなかった未知データに対して当てはまりが良いことを表し, テストデータに対する平均対数部分尤度は比例ハザードモデルの汎化性能を評価する指標として用いられる [33].

比例ハザードモデルでは説明変数として以下のものを用いた：

- 顧客の累積来店回数：累積来店回数 1 回目から 19 回目について当該累積来店回数であれば 1，そうでなければ 0 をとるダミー変数；
- 性別：男性であれば 1, 女性であれば 0；
- 年代：「30 代」,「40 代」,「50 代以降」についてそれぞれ当該年代であれば 1，そうでなければ 0 をとるダミー変数；
- 前回来店時売上：平均 0, 分散 1 になるように正規化した連続変数；

- 店舗との距離：店舗と顧客が住所登録した郵便番号の座標間の距離が「2km 以上 5km 未満」,「5km 以上」についてそれぞれ当該距離であれば 1，そうでなければ 0 をとるダミー変数；
- 来店月：1 月から 11 月について当該来店月であれば 1, そうでなければ 0 をとるダミー変数.

なお顧客属性を表すダミー変数である性別, 年代, 店舗との距離, 来店月について, 今回使用するデータセットをすべてまとめたデータに対してそれぞれログランク検定 [44] を行ったところ, これらの説明変数を含めた場合と含めなかった場合の差はいずれも有意水準 1% のもとで 0 ではなかった.

問題 (4.5) は Python ライブラリ Scipy の optimize.minimize 関数[*1] を用いて逐次二次計画法 (method='SLSQP') により求解した.

4.4.2　単調性制約と凹性制約の有効性

ここでは本章で提案する形状制約比例ハザードモデルの有効性を検証する. 各店舗に対して, 累積来店回数を一つの量的変数として扱った場合 (線形) , ダミー変数として扱い制約を課さなかった場合 (制約なし) , 制約を課した場合 (単調性, 単調性＋凹性) それぞれの手法を全 12 個のデータセットに対して適用し, 各データセットで 3 分割交差検証を行った. 得られた平均対数部分尤度 (12 個 × 3 回 = 36 個の平均対数部分尤度) の算術平均を店舗別にまとめた結果を表 4.1 に示す.

表 4.1 を見ると, 今回検証に利用した全ての店舗について累積来店回数を量的変数として扱うよりもダミー変数として扱った場合の方が平均対数部分尤度の平均値は高いことがわかる. また制約なしの比例ハザードモデルに比べて単調性制約つきの比例ハザードモデルの方が, さらに単調性制約つき比例ハザードモデルに比べて単調性＋凹性制約つき比例ハザードモデルのほうが平均対数部分尤度の平均値は高い. この結果より, 累積来店回数と再来店率に対して非線形に加えて単調性と凹性を仮定したことで予測性能が改善したと考えることができる.

表 4.1. 各手法で計算した平均対数部分尤度の平均値

	店舗									
	A	B	C	D	E	F	G	H	K	L
線形	−5.3457	−5.4724	−4.6284	−5.0375	−4.7545	−4.6068	−5.0867	−4.3243	−4.9772	−4.1188
制約なし	−5.3279	−5.4417	−4.5751	−4.9884	−4.7071	−4.5711	−4.8515	−4.2103	−4.9259	−4.0781
単調性	−5.3257	−5.4397	−4.5736	−4.9869	−4.7047	−4.5682	−4.8490	−4.2081	−4.9240	−4.0749
単調性＋凹性	**−5.3253**	**−5.4395**	**−4.5731**	**−4.9865**	**−4.7042**	**−4.5676**	**−4.8484**	**−4.2075**	**−4.9236**	**−4.0740**

続いて店舗の来客数ごとに大規模店舗 (A, B, D) , 中規模店舗 (C, E, G, K) , 小規模店舗 (F, H, L) と 3 区分に層別して, 制約なし, 単調性制約, 単調性＋凹性制約の比例ハザードモデルの比較を行う. ここでは表 4.1 と同様, 各店舗に対して全 12 個のデータセットそれぞれで 3 分割交差検証を行って得られた各平均対数部分尤度に対して改善率を計算し, それらを店舗

[*1] https://docs.scipy.org/doc/scipy/reference/optimize.minimize-slsqp.html

の各規模で層別して箱ひげ図を描く．ここで改善率は制約つきモデルの平均対数部分尤度を MLL_a, 制約なしモデルの平均対数部分尤度を MLL_b として,

$$\text{改善率 } (\%) = \frac{MLL_a - MLL_b}{MLL_b} \times 100$$

と定義した．店舗規模別の改善率をまとめた箱ひげ図を図 4.2 に示す．

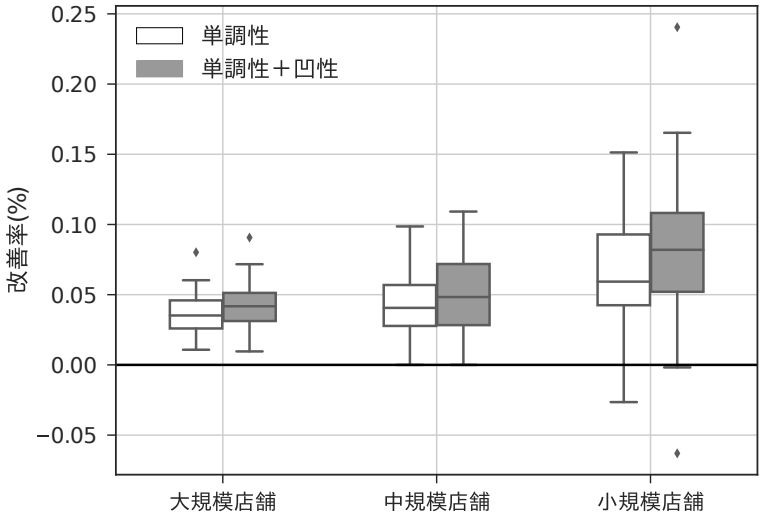


図 4.2. 店舗の規模別の平均対数部分尤度の改善率

図 4.2 からは, いずれの店舗規模でも単調性制約つきモデルと比較して単調性 + 凹性制約つきモデルのほうが改善率の中央値が大きい値をとっていることがわかる．これら二つのモデルの改善率の結果に対してウィルコクソンの符号順位検定 [73] を行ったところ, いずれの店舗規模においても有意水準 1% のもとで改善率の差は 0 ではないことが確認された．また図 4.2 からは店舗の規模が小さくなるほど制約なしのモデルと単調性制約を課したモデル, 制約なしモデルと単調性＋凹性制約を課したモデルの改善率の差が大きくなることがわかり, またその差は単調性＋凹性制約つきモデルのほうが大きいことがわかる．

図 4.3 の小規模店舗 L の定着度をみると, 単調性制約を加えることで累積来店回数間の定着度の逆転が補正され, また凹性制約を加えることで定着度の変動が抑えられることがわかる．一方, 図 4.4 の大規模店舗 A をみると, 小規模店舗と同程度に単調性制約により累積来店回数に対する定着度の逆転が補正されたが, 単調性制約を満たした時点で凹性制約もおおむね満たしており, 小規模店舗より凹性制約による補正幅が小さいことがわかる．このような結果は, 来客数の少ない小規模店舗では過剰適合が生じやすく, 単調性制約に加えて凹性制約を課すことの効果が大きく表れていたのが要因であると考えられる．

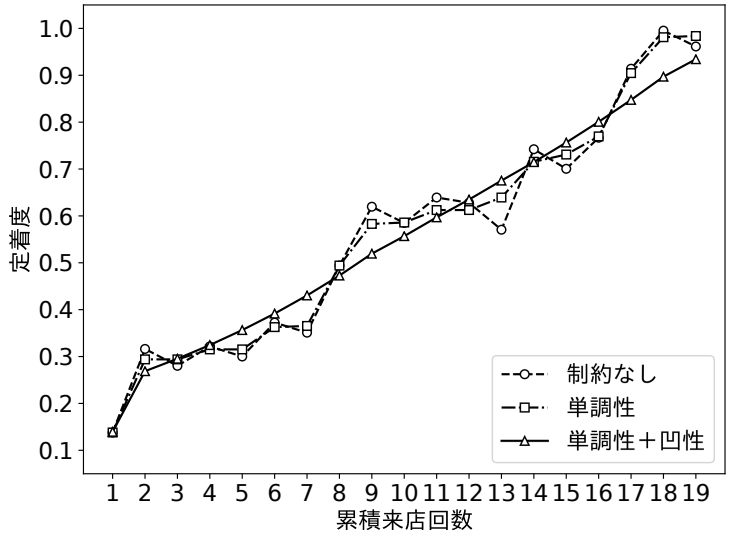


図 4.3. 小規模店舗 L での制約なし，単調性，単調性＋凹性の場合の定着度

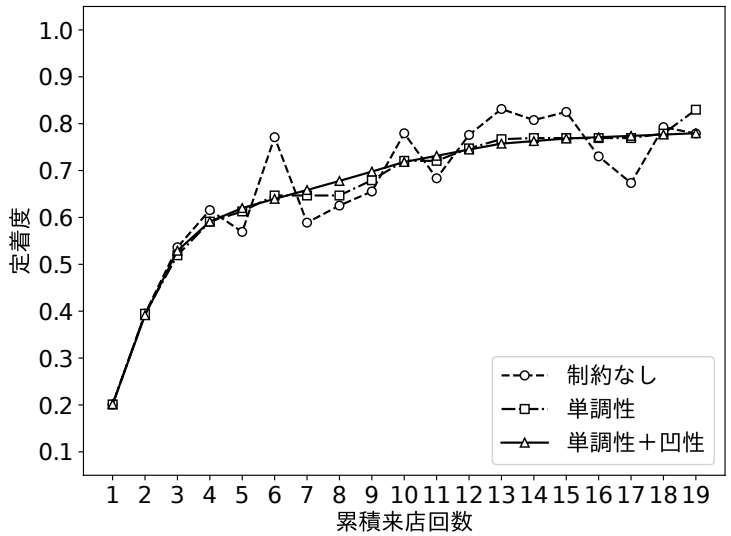


図 4.4. 大規模店舗 A での制約なし，単調性，単調性＋凹性の場合の定着度

4.4.3　主成分分析による店舗状況推移の可視化

本項では, 提案モデルを用いて推定した各月の店舗ごとの定着度に対して主成分分析を適用することで状況推移を可視化し, 各店舗の状況について考察する.

定着度に対する主成分分析の適用

4.3 節で述べた提案モデルでは, 各店舗に対して 1 回目来店の顧客, 2 回目来店の顧客など, それぞれの累積来店回数に対して定着度を推定した. しかし, 累積来店回数による顧客の再来店の差異を細かく考察したい場合, 各店舗の定着度が高次元となり解釈が困難となる.

そこで本章では, 各店舗の定着度に対して主成分分析を適用することで低次元の合成変数を作成し, 状況を解釈しやすくすることを考える. 具体的には図 4.5 のように, 各店舗に対して推定した 19 次元の定着度から 2 次元の合成変数を作成することで各店舗を 2 次元平面上へと射影する.

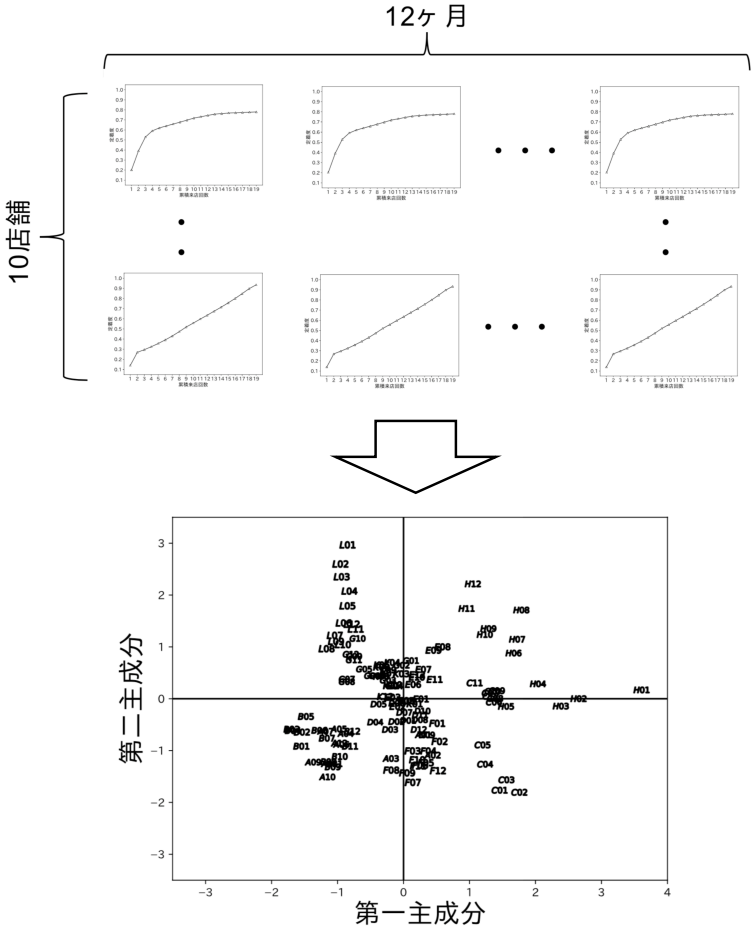
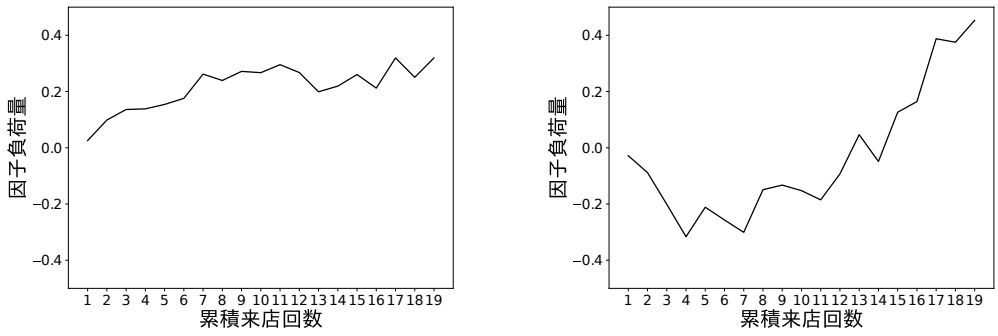
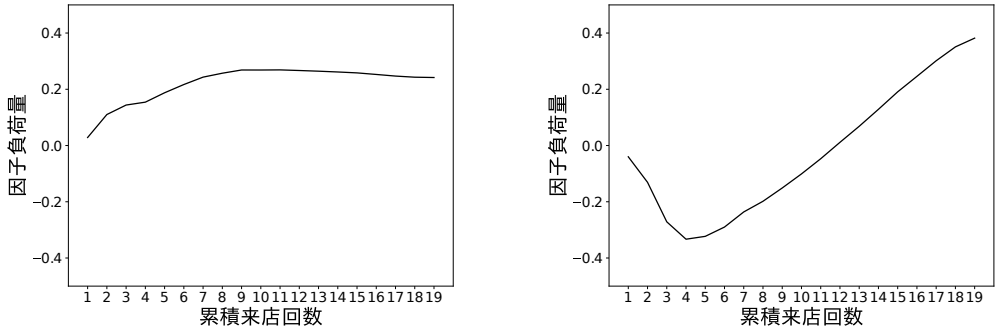


図 4.5. 各月, 各店舗の定着度の主成分分析による可視化の概念図

2016 年 7 月から 2017 年 6 月までの各店舗に推定された定着度に対して主成分分析を適用し, 第一主成分と第二主成分の因子負荷量と寄与率をそれぞれ図 4.6, 表 4.2 にまとめる. 図 4.6 より, 単調性＋凹性制約を課した場合は制約なしの場合と比較して累積来店回数に対する因子負荷量が平滑化されていることがわかる. また表 4.2 より, 単調性＋凹性制約を課した場合は累積寄与率が増加しているとわかる.



第一主成分 (制約なし)　　第二主成分 (制約なし)



第一主成分 (単調性＋凹性)　　第二主成分 (単調性＋凹性)

図 4.6. 主成分分析の因子負荷量

表 4.2. 制約なし，単調性＋凹性の場合の第一主成分と第二主成分の寄与率

	寄与率		
	第一主成分	第二主成分	累積
制約なし	0.596	0.106	0.702
単調性＋凹性	0.845	0.116	0.961

得られた主成分について考察する．第一主成分は各累積来店回数でそれぞれ同程度の値をとり, 符号がすべて正であることから各累積来店回数の顧客の「平均的な定着度」の大きさと解釈できる．また第二主成分は累積来店回数が少ない場合には負の値をとり, 累積来店回数が多い場合には正の値をとることから, 累積来店回数が多い顧客と少ない顧客の「定着度の乖離」の大きさと解釈できる．第二主成分の値が他店舗と比較して相対的に大きい店舗は, 累積来店回数が少ない顧客の定着度が累積来店回数が多い顧客に比べて低いことを表し, 累積来店回数の少ない顧客について再来店率改善の余地があると考えることができる．一方で相対的に小さい店舗は累積来店回数の多い顧客について再来店率改善の余地があると考えることができる．

作成した因子負荷量を用いて, 2016 年 7 月から 2017 年 6 月の各月で各店舗の定着度に対

して主成分得点を算出し, 定着度推移の可視化を行った. 紙面の都合上, 図 4.7 では 2016 年 7 月を 1 ヶ月目として 1〜4, 4〜8, 8〜12 ヶ月目にあたる 2016 年 7 月〜10 月, 2016 年 10 月〜2017 年 2 月, 2017 年 2 月〜6 月についてのみ抽出し掲載する. 図中では年月を数字で表しており, 期間の始まりである 2016 年 7 月を 01 とし, 期間の終わりである 2017 年 6 月を 12 としている.

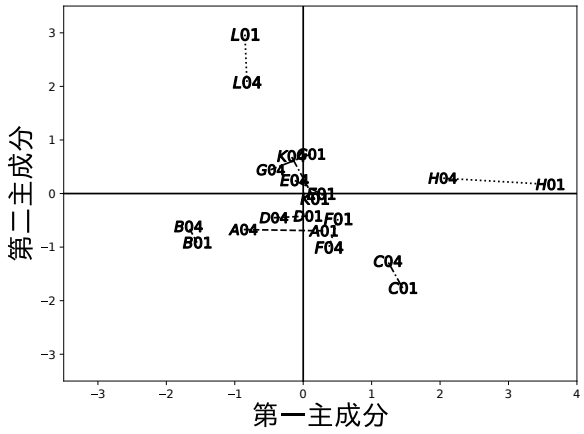
各店舗状況推移に対する考察

図 4.7 の定着度推移に基づいて、推移の特徴的な店舗に対して考察する.

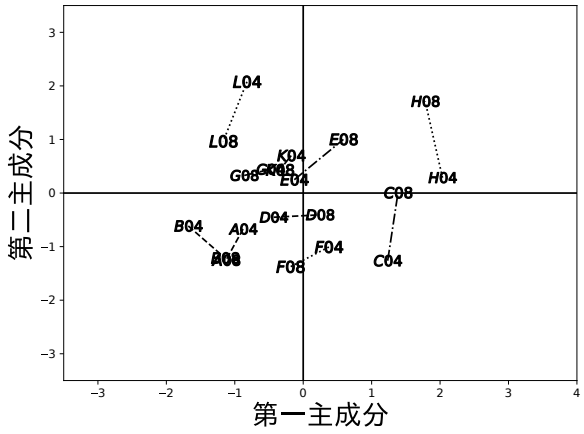
店舗 A 2016 年 7 月には第一主成分 (平均的な定着度) の値が他店舗と比較して高かったが, 2017 年 2 月まで減少し全店舗で下位 3 番目まで減少した. 平均的な定着度はその後増加するものの他店舗と比較すると依然として低い値であり, 12 ヶ月間を通して顧客全体の平均的な定着度は悪化したと解釈できる. また第二主成分 (定着度の乖離) については 12 ヶ月間を通して他店舗と比較して低い値をとっており, 相対的に累積来店回数の多い顧客の再来店率に課題があると解釈できる. 店舗 A が累積来店回数の多い顧客の再来店率の改善を行うためには, 12 ヶ月を通して他店舗と比較して平均的な定着度が高く, かつ定着度の乖離の値が 12 ヶ月間で負の値から正の値に推移し累積来店回数の多い顧客の再来店率に対する改善が行われたと思われる店舗 C のオペレーション改善が参考になると示唆される.

店舗 L 第一主成分の値は他店舗と比較して小さい値であり, 期間全体を通じてほぼ同程度の値をとっている. 一方, 第二主成分の値は 2016 年 6 月から 2017 年 2 月にかけて低下していたものの, それ以降 2017 年 6 月までは増加に転じている. このことから店舗 L は他店舗と比較して直近では新規顧客の再来店率が改善の傾向にあると解釈できる. 店舗 L がさらに新規顧客の再来店率改善を行うためには, 12 ヶ月目において平均的な定着度が高い値をもち, 定着度の乖離の値が低い店舗 D, F などにおける顧客に対するオペレーションが参考になると示唆される.

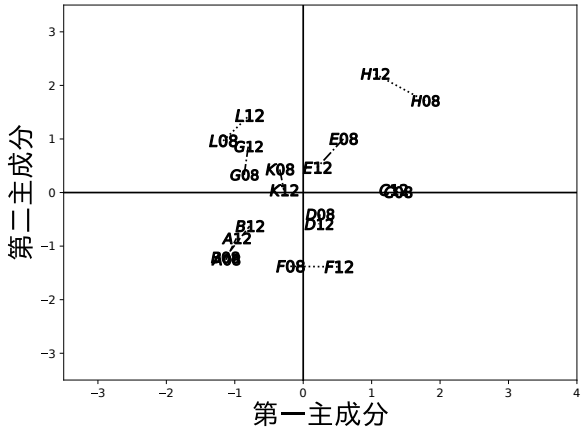
店舗状況を表すデータが今後更新された場合にも, 同じ因子負荷量を用いて同じ軸での解釈が可能になり, 状況の変化を連続的に把握することができる.



(a) 2016 年 7 月 : 01 → 2016 年 10 月 : 04 の推移



(b) 2016 年 10 月:04 → 2017 年 2 月:08 の推移



(c) 2017 年 2 月:08 → 2017 年 6 月 : 12 の推移

図 4.7. 各店舗の因子負荷量の推移

4.5 まとめ

本章では, 複数の店舗をもつヘアサロンを対象とし, 累積来店回数と再来店率に関する単調性と凹性を考慮した比例ハザードモデルにより再来店状況を分析した. ヘアサロンチェーンの実際のデータを用いた分析の結果, 累積来店回数と再来店率の関係を考慮し制約条件を課すことで比例ハザードモデルの予測性能が向上し, 特に来客数が少ない小規模店舗についてその効果が大きいという結果を得ることができた.

また, 各店舗の顧客再来店の状況推移を把握するため, 各月のデータを用いて比例ハザードモデルにより推定した定着度を主成分分析により可視化した. これにより各店舗における顧客の再来店状況の推移を捉えることが可能となり, それぞれの店舗における再来店率改善のための指針を得ることができた.

今後の課題としては, 累積来店回数や顧客属性だけでなく会計情報などを用いてより詳細に再来店状況を分析することが挙げられる. また本章では各店舗, 時期による再来店状況の異質性を考慮するため, 店舗別, 月別で層別してモデルを構築している. 既存研究ではこれらの異質性を考慮するために潜在クラスモデルや階層ベイズモデルなどのアプローチ [76, 79] が提案されており, これらの手法を組み合わせた分析も今後の課題である.

第 5 章

結論

5.1　主要な結果

本研究は，顧客の商品選択行動の解析により商品選択予測精度の向上と，顧客の商品選択の過程に関する知見の獲得を目的としたものであった．その目的のため，各章では，顧客の実際の商品選択行動を記録したデータを解析し，顧客の商品選択に影響を与える特徴量を探索し，その特徴量の観察から，顧客の商品選択に関する性質を仮定し，仮定を形状制約として課して選択予測モデルのパラメータを推定した．

顧客の商品選択に関する性質として，第 2 章では，2 次元確率表の単調性＋凸性＋凹性に加えて商品類型ごとの商品選択傾向の異質性を仮定した潜在クラス確率表モデルを提案した．第 3 章では，2 次元確率表と比較してより詳細な商品閲覧に関する情報を利用するため，商品閲覧数の時系列を表す閲覧数列に対して単調性を仮定した形状制約閲覧数列モデルを提案した．第 4 章では，顧客の継続選択の性質を仮定した形状制約比例ハザードモデルを提案した．

以下では，これらの章で示した本論文の主要な結果をまとめる．

第 2 章では，各顧客の過去の閲覧履歴の最新度と頻度の特徴量から商品選択確率を推定する潜在クラス型の形状制約モデルを提案した．提案手法では，商品をいくつかの潜在クラスに分類し，それぞれのクラスについて最新度と頻度に関する単調性，凸性，凹性の制約条件を満たすように 2 次元の確率表を推定した．また，潜在クラス型確率表を推定するための EM アルゴリズムについても提案した．さらに，実際のページ遷移データを用いて作成した潜在クラスモデルから，商品類型ごとの顧客の商品選択行動について分析した．

数値実験の結果，提案手法は検証で利用したデータに対して，潜在クラスロジスティック回帰モデルやサポートベクターマシン，人工ニューラルネットワークよりも予測精度が高かった．また，推定された潜在クラスの分析を，それぞれの商品類型に対する販売促進戦略を計画するためにも有用であることを示唆した．

第 3 章では，商品閲覧の時系列である閲覧数列に対して，頻度と最新度に関する形状制約つき最適化問題を解くことで商品選択確率を推定する手法を提案した．また閲覧数列について，半順序の簡約表現であるハッセ図の構造を利用して冗長な制約条件を削除することで最適化計算を効率化する方法について提案した．

数値実験の結果，検証で利用したデータに対して，提案手法は期間数や閲覧数上限が同じとき先行研究の形状制約モデルである 2 次元確率表モデルと比較して予測精度が高いことがわかった．そして，同一の閲覧数列の特徴量を用いて学習を行ったとき，機械学習の手法である L2 正則化ロジスティック回帰，人工ニューラルネットワーク，ランダムフォレストと比較して予測精度が高いことがわかった．また，閲覧数列モデルによる商品選択確率の補正は，機械学習の手法による予測値に対しても適用可能で，それによって予測精度が改善することを確認した．閲覧数列モデルの商品選択確率を推定するための最適化計算を行うとき，ハッセ図の構造を利用して冗長な制約条件を削減することで計算時間が高速化することがわかった．

第 4 章では, 複数の店舗をもつヘアサロンを対象とし, 累積来店回数と再来店率に関して単調性と凹性を仮定した形状制約比例ハザードモデルを提案した．また，推定された形状制約比例ハザードモデルのパラメータを主成分分析により縮約した指標の時系列変化を可視化することで各店舗の再来店状況を分析した．

数値実験の結果，検証で利用したデータに対して，累積来店回数と再来店率の関係を仮定し形状制約を課すことで比例ハザードモデルの予測性能が向上した．特に形状制約の効果は，来客数が少ない小規模店舗で大きかった．推定された形状制約比例ハザードモデルのパラメータを縮約した指標の可視化では，各店舗について平均的な再来店率，新規顧客と継続顧客で分けた再来店率の観点で，再来店状況の推移を捉えることが可能となり, それぞれの店舗における再来店率改善のための示唆を得た．

5.2　商品選択予測における形状制約モデルの選択の指針

本研究では，各章で異なる商品選択予測の課題に対して手法を提案した．そこで，本節では各章で提案した形状制約モデルについて整理し，各章の結果をもとにどのような状況でそれぞれの形状制約モデルを適用すべきかについて指針を示す．

表 5.1 は，本論文で扱った形状制約モデルについて，形状制約を課す方法，他の特徴量の扱い方の観点で整理した表である．モデルは，既存手法の 2 次元確率表モデル [32]，第 3 章で説明した閲覧数列モデル，第 4 章で説明した形状制約比例ハザードモデルを含む他の統計モデル，機械学習モデルと大別する．

まず 2 次元確率表モデルでは，商品閲覧の最新度と頻度の 2 次元の特徴量に対して単調性＋凸性＋凹性の形状制約を仮定して，形状制約つきの最適化問題を解くことでモデルのパラメータを推定した．商品類型などの形状制約を仮定しない他の特徴量に関しては，第 2 章で述べたように，その異質性を潜在クラスとして考慮することで予測精度が向上することを示した．

続いて閲覧数列モデルでは，多次元である閲覧数列に対して，閲覧数列の最新度と頻度に基づく単調性の形状制約を仮定して形状制約つきの最適化問題を解くことでモデルのパラメータを推定した．閲覧数列モデルについても，閲覧数列以外の形状制約を仮定しない特徴量について考慮する場合には 2 次元確率表モデルと同様に潜在クラスモデルと組み合わせることが考えられる．

2 次元確率表モデルや閲覧数列モデル以外の統計モデルや機械学習モデルにおいて特徴量の

形状制約を考慮する方法については，2 つの方法を述べた．1 つ目の方法は，第 4 章で説明した形状制約比例ハザードモデルのように，モデルのパラメータに形状制約を課して推定する方法である．このようなモデルのパラメータ推定において形状制約を課す方法は，比例ハザードモデル以外のパラメトリックモデルについても適用可能と考えられ，その際にも 1.3 節で説明した手順で形状制約モデルの有効性が検証可能である．2 つ目の方法は，第 3 章で説明したように，モデルの予測値を入力として，形状制約つき最適化問題を解き予測値を補正する方法である．この方法も，第 3 章の閲覧数列モデルでの予測値の補正に限らず，既存手法の 2 次元確率表モデルを用いても可能である．

形状制約を課さない他の特徴量については，モデルのパラメータ推定時に考慮する場合には，形状比例ハザードモデルでのパラメータ推定と同様にそれぞれ特徴量として追加をすればよい．

表 5.1. 形状制約モデルの分類

モデル	形状制約を課す方法	他の特徴量の扱い方
2 次元確率表モデル [32]	商品閲覧の最新度と頻度に基づく単調性＋凸性＋凹性	潜在クラスとして考慮：**第 2 章**
閲覧数列モデル	閲覧数列の最新度と頻度に基づく単調性：**第 3 章**	潜在クラスとして考慮
他の統計モデル，機械学習モデル（形状制約比例ハザードモデル）	モデルのパラメータ推定時に考慮（定着度の単調性＋凹性：**第 4 章**），予測値の後処理として補正（閲覧数列の最新度と頻度に基づく単調性：**第 3 章**）	特徴量として追加

続いて，図 5.1 は各章の結果をもとにどのような状況でそれぞれの形状制約モデルを適用すべきかについて表したフローチャートである．これは，1.3 節で説明した実践的な問題解決のために，どのように形状制約モデルを選択するかについて指針を示したものである．

商品選択予測モデルの作成では，まず商品選択予測の対象データを解析し，どのような特徴量が商品選択と関連があるかを観察する．対象とするデータによってどのような特徴量が商品選択と関連があるかは異なるが，まず商品選択において重要な特徴量であることが実証されている [18, 19, 34, 60, 61] 商品閲覧の最新度と頻度の特徴量に着目して，単調性，凸性，凹性などの性質がみられるか確認するとよい．これらの特徴量に性質が見られない場合には，形状制約を仮定せずに統計モデル，機械学習モデルを利用すればよい．

次に，最新度と頻度以外に有効な特徴量が存在するかを確認する．最新度と頻度以外に特に有効な特徴量がない場合には，商品や顧客の異質性を考慮すべきか検討し，考慮すべき場合には，潜在クラスを考慮した 2 次元確率表モデル，閲覧数列モデルを用いる．2 次元確率表モデルを用いるか，閲覧数列モデルを用いるかについては，第 3 章の数値実験の結果から，標本サ

イズの大きさをもとに判断するのがよい．標本サイズが大きい場合には表現可能なパターン数がが多い閲覧数列モデルを，標本サイズが小さい場合には表現可能なパターン数が少ない 2 次元確率表モデルを用いるのがよい．

最新度や頻度以外に有効な特徴量が存在する場合には，他の特徴量を明示的に扱うことができるパラメトリック（線形）モデルやノンパラメトリック（非線形）モデルを用いる．標本サイズが大きい場合には，表現力の高いノンパラメトリック（非線形）モデルを用いて予測値の後処理として形状を考慮した補正するのがよい．一方で，標本サイズが小さい場合には，パラメトリック（線形）モデルのパラメータ推定にて形状制約を課すのがよい．

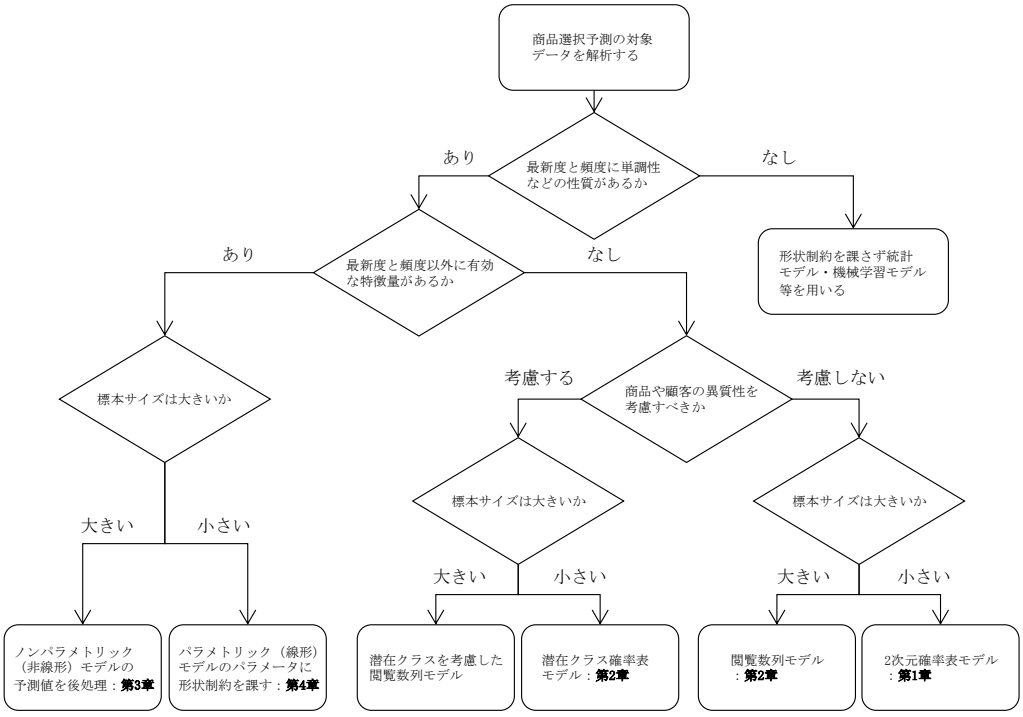


図 5.1. 商品選択予測における形状制約モデルの選択のフローチャート

5.3 今後の展望

本研究では，顧客の商品選択行動の解析により商品選択予測精度の向上と，顧客の商品選択の過程に関する知見の獲得を目的として，いくつかのモデルと特徴量について形状制約を仮定し実データを用いて検証した．

各章で述べた以外にも，顧客の商品選択に対して影響を与える特徴量や形状制約を課すこと

のできるモデルは多く考えられ，商品選択予測の精度向上のための特徴量の探索と性質のモデル化を試みることが本研究の今後の展望である．

具体的には，比例ハザードモデル以外のパラメトリックモデルに対して特徴量の性質を考慮した形状制約を課してパラメータを推定することや，第 3 章の数値実験で述べた以外の統計モデル，機械学習モデルの後処理として閲覧数列モデルや 2 次元確率表モデルにより補正することの有効性を検証することが挙げられる．

また，第 3 章の閲覧数列モデルや，第 4 章の形状制約比例ハザードモデルにおいて商品や顧客の異質性を考慮するために，第 2 章と同様に潜在クラスモデルと組み合わせることの有効性を検証することも課題である．第 2 章の数値実験では，標本サイズが大きい場合に潜在クラスを考慮することで予測精度が向上しており，十分にデータが収集されている状況では閲覧数列モデルや形状制約比例ハザードモデルにおいても商品や顧客の異質性の考慮によって予測精度が向上することが期待できる．

閲覧数列の新たな性質を探索することも課題である．閲覧数列に期待される性質として，閲覧数列の頻度に基づく単調性と，最新度に基づく二つの単調性について検証を行ったが，それ以外にも有効な性質を発見することで，更なる予測精度向上の余地がある．

最後に，第 3 章で説明した形状正則化モデルを閲覧数列モデルだけでなく 2 次元確率表モデルなどの他の形状制約モデルに拡張することも課題である．

謝辞

このように本研究を行うことができたのは，様々な形でご指導，ご支援，励ましをいただいた多くの方々のおかげです．

特に，指導教員である高野祐一先生には，私が大学学部生の頃から大学院を卒業し，現在に至るまで，研究内容にとどまらず，あらゆる面にわたってご指導いただきました．高野先生の熱心な教育によって，データ解析者としての素養を得たと感じております．

次に，共同研究者である岩永二郎さん，鮭川矩義先生，小林健さん，吉住宗朔さんからは研究活動やデータ解析コンペティションでの分析を通して議論をさせていただいた中で，多くのことを学ばせていただきました．

また，お忙しい中時間を割き，本論文の審査を担当していただき，有益な助言をいただいた安東弘泰先生，岡田幸彦先生，繁野麻衣子先生，今倉暁先生にも感謝の意を表します．

そして，本学への進学機会と経済的支援を与えて下さった株式会社リクルート，株式会社リクルートライフスタイルの関係者の方々に対しても大いに感謝をしております．

第 4 章の数値実験で利用したデータを提供いただいた，経営科学系研究部会連合協議会の方々にも感謝の意を表します．

最後に，家族の支援や理解がなければ，この論文が完成しなかったことをここに記し，重ねて感謝いたします．

参考文献

[1] Scott T. Acton and Alan C. Bovik. Nonlinear image estimation using piecewise and local image models. *IEEE Transactions on Image Processing*, Vol. 7, No. 7, pp. 979–991, 1998.

[2] Yacine Aıt-Sahalia and Jefferson Duarte. Nonparametric option pricing under shape restrictions. *Journal of Econometrics*, Vol. 116, No. 1-2, pp. 9–47, 2003.

[3] Richard E. Barlow, David J. Bartholomew, James M. Bremner, and H. Daniel Brunk. Statistical inference under order restrictions: The theory and application of isotonic regression. Technical report, Wiley New York, 1972.

[4] Amit Bhatnagar and Sanjoy Ghose. A latent class segmentation analysis of e-shoppers. *Journal of Business Research*, Vol. 57, No. 7, pp. 758–767, 2004.

[5] Emad Gohari Boroujerdi, Soroush Mehri, Saeed Sadeghi Garmaroudi, Mohammad Pezeshki, Farid Rashidi Mehrabadi, SeyyedSalim Malakouti, and Shahram Khadivi. A study on prediction of user's tendency toward purchases in websites based on behavior models. In *2014 6th Conference on Information and Knowledge Technology (IKT)*, pp. 61–66, 2014.

[6] Norman Breslow. Covariance analysis of censored survival data. *Biometrics*, pp. 89–99, 1974.

[7] Hugh D. Brunk. Maximum likelihood estimates of monotone parameters. *The Annals of Mathematical Statistics*, Vol. 26, No. 4, pp. 607–616, 1955.

[8] Hugh D. Brunk. On the estimation of parameters restricted by inequalities. *The Annals of Mathematical Statistics*, Vol. 29, No. 2, pp. 437–454, 1958.

[9] Randolph E. Bucklin, James M. Lattin, Asim Ansari, Sunil Gupta, David Bell, Eloise Coupey, John D.C. Little, Carl Mela, Alan Montgomery, and Joel Steckel. Choice and the internet: From clickstream to research stream. *Marketing Letters*, Vol. 13, No. 3, pp. 245–258, 2002.

[10] Randolph E. Bucklin and Catarina Sismeiro. Click here for internet insight: Advances in clickstream data analysis in marketing. *Journal of Interactive Marketing*, Vol. 23, No. 1, pp. 35–48, 2009.

[11] Johnson R. Burke. Examining the validity structure of qualitative research. *Educa-*

tion, Vol. 118, No. 2, pp. 282–292, 1997.

[12] Zhen-Yu Chen and Zhi-Ping Fan. Distributed customer behavior prediction using multiplex data: A collaborative MK-SVM approach. *Knowledge-Based Systems*, Vol. 35, pp. 111–119, 2012.

[13] Guang Cheng. Semiparametric additive isotonic regression. *Journal of Statistical Planning and Inference*, Vol. 139, No. 6, pp. 1980–1991, 2009.

[14] David R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 34, No. 2, pp. 187–202, 1972.

[15] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 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.

[16] Lifeng Deng, Jieli Ding, Yanyan Liu, and Chengdong Wei. Regression analysis for the proportional hazards model with parameter constraints under case-cohort design. *Computational Statistics & Data Analysis*, Vol. 117, pp. 194–206, 2018.

[17] Brian Everitt. *Introduction to Optimization Methods and their Application in Statistics*. Springer science & business media, 2012.

[18] Peter S. Fader, Bruce G.S. Hardie, and Ka Lok Lee. "counting your customers" the easy way: An alternative to the pareto/nbd model. *Marketing Science*, Vol. 24, No. 2, pp. 275–284, 2005.

[19] Peter S. Fader, Bruce G.S. Hardie, and Ka Lok Lee. RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research*, Vol. 42, No. 4, pp. 415–430, 2005.

[20] Alberto Fernández, Sara del Río, Nitesh V. Chawla, and Francisco Herrera. An insight into imbalanced big data classification: Outcomes and challenges. *Complex & Intelligent Systems*, Vol. 3, No. 2, pp. 105–120, 2017.

[21] Anton K. Formann. Linear logistic latent class analysis. *Biometrical Journal*, Vol. 24, No. 2, pp. 171–190, 1982.

[22] Donald A.S. Fraser and Hélene Massam. A mixed primal-dual bases algorithm for regression under inequality constraints. application to concave regression. *Scandinavian Journal of Statistics*, pp. 65–74, 1989.

[23] Ronald A. Gallant and Gene H. Golub. Imposing curvature restrictions on flexible functional forms. *Journal of Econometrics*, Vol. 26, No. 3, pp. 295–321, 1984.

[24] Ulf Grenander. On the theory of mortality measurement: Part ii. *Scandinavian Actuarial Journal*, Vol. 1956, No. 2, pp. 125–153, 1956.

[25] Shelby J. Haberman. *Analysis of Qualitative Data. Volume.2, New Developments*. Academic Press, 1979.

[26] Jacques A. Hagenaars and Allan L. McCutcheon. *Applied Latent Class Analysis*. Cambridge University Press, 2002.

[27] Clifford Hildreth. Point estimates of ordinates of concave functions. *Journal of the American Statistical Association*, Vol. 49, No. 267, pp. 598–619, 1954.

[28] C.C. Holmes and N.A. Heard. Generalized monotonic regression using random change points. *Statistics in Medicine*, Vol. 22, No. 4, pp. 623–638, 2003.

[29] Tingliang Huang and Jan A. Van Mieghem. The promise of strategic customer behavior: On the value of click tracking. *Production and Operations Management*, Vol. 22, No. 3, pp. 489–502, 2013.

[30] Tingliang Huang and Jan A. Van Mieghem. Clickstream data and inventory management: Model and empirical analysis. *Production and Operations Management*, Vol. 23, No. 3, pp. 333–347, 2014.

[31] Arthur Middleton Hughes. *Strategic Database Marketing: The masterplan for starting and managing a profitable, customer-based marketing program*, Vol. 12. McGraw-Hill New York, 2000.

[32] Jiro Iwanaga, Naoki Nishimura, Noriyoshi Sukegawa, and Yuichi Takano. Estimating product-choice probabilities from recency and frequency of page views. *Knowledge-Based Systems*, Vol. 99, pp. 157–167, 2016.

[33] Tomoharu Iwata, Saito Kazumi, and Yamada Takeshi. Recommendation method for improving customer lifetime value. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 20, No. 9, pp. 1254–1263, 2008.

[34] Kinshuk Jerath, Peter S. Fader, and Bruce G.S. Hardie. New perspectives on customer "death" using a generalization of the Pareto/NBD model. *Marketing Science*, Vol. 30, No. 5, pp. 866–880, 2011.

[35] Wagner A. Kamakura and Gary J. Russell. A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research*, Vol. 26, No. 4, pp. 379–390, 1989.

[36] Komal Kapoor, Mingxuan Sun, Jaideep Srivastava, and Tao Ye. A hazard based approach to user return time prediction. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pp. 1719–1728, 2014.

[37] Paul Klemperer. Markets with consumer switching costs. *The Quarterly Journal of Economics*, Vol. 102, No. 2, pp. 375–394, 1987.

[38] Bartosz Krawczyk. Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, Vol. 5, No. 4, pp. 221–232, 2016.

[39] Paul Felix Lazarsfeld and Neil W. Henry. *Latent Structure Analysis*. Houghton Mifflin Co., 1968.

[40] Zikuan Liu, Jalal Almhana, Vartan Choulakian, and Robert McGorman. Online EM algorithm for mixture with application to internet traffic modeling. *Computational statistics & data analysis*, Vol. 50, No. 4, pp. 1052–1071, 2006.

[41] Malte Ludewig and Dietmar Jannach. Evaluation of session-based recommendation algorithms. *User Modeling and User-Adapted Interaction*, Vol. 28, No. 4-5, pp. 331–390, 2018.

[42] Lee Lynd. A perspective on engineering and its relationship to systems analysis and science [portable document format]. Retrieved from `http://www.dartmouth.edu/~sullivan/22files/Engineering_vs._science.pdf`, 2004.

[43] Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. *Natural Language Engineering*, Vol. 16, No. 1, pp. 100–103, 2010.

[44] Nathan Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, Vol. 50, pp. 163–170, 1966.

[45] Geoffrey McLachlan and Thriyambakam Krishnan. *Order Restricted Statistical Inference*, Vol. 1. John Wiley & Sons, 1988.

[46] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions*. John Wiley & Sons, 2007.

[47] Mary C. Meyer. Semi-parametric additive constrained regression. *Journal of Nonparametric Statistics*, Vol. 25, No. 3, pp. 715–730, 2013.

[48] Wendy W. Moe and Peter S. Fader. Dynamic conversion behavior at e-commerce sites. *Management Science*, Vol. 50, No. 3, pp. 326–335, 2004.

[49] Alan L. Montgomery. Applying quantitative marketing techniques to the internet. *Interfaces*, Vol. 31, No. 2, pp. 90–108, 2001.

[50] Alan L. Montgomery, Shibo Li, Kannan Srinivasan, and John C. Liechty. Modeling online browsing and path analysis using clickstream data. *Marketing science*, Vol. 23, No. 4, pp. 579–595, 2004.

[51] Eric W.T. Ngai, Li Xiu, and Dorothy C.K. Chau. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, Vol. 36, No. 2, pp. 2592–2602, 2009.

[52] Naoki Nishimura, Noriyoshi Sukegawa, Yuichi Takano, and Jiro Iwanaga. A latent-class model for estimating product-choice probabilities from clickstream data. *Information Sciences*, Vol. 429, pp. 406–420, 2018.

[53] Naoki Nishimura, Noriyoshi Sukegawa, Yuichi Takano, and Jiro Iwanaga. Estimating product-choice probabilities from sequence of page views. In *2019 International Symposium on Nonlinear Theory and Its Applications*, pp. 25–28, 2019.

[54] Jorge Nocedal and Stephen J. Wright. *Sequential Quadratic Programming*. Springer, 2006.

[55] Rainer Olbrich and Christian Holsing. Modeling consumer purchasing behavior in social shopping communities with clickstream data. *International Journal of Electronic Commerce*, Vol. 16, No. 2, pp. 15–40, 2011.

[56] Patryk Orzechowski, William La Cava, and Jason H. Moore. Where are we now?: A large benchmark study of recent symbolic regression methods. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 1183–1190, 2018.

[57] Panos M. Pardalos and Guoliang Xue. Algorithms for a class of isotonic regression problems. *Algorithmica*, Vol. 23, No. 3, pp. 211–222, 1999.

[58] Jiangtao Qiu, Zhangxi Lin, and Yinghong Li. Predicting customer purchase behavior in the e-commerce context. *Electronic Commerce Research*, Vol. 15, No. 4, pp. 427–452, 2015.

[59] Sergio Ramírez-Gallego, Bartosz Krawczyk, Salvador García, Michał Woźniak, and Francisco Herrera. A survey on data preprocessing for data stream mining: Current status and future directions. *Neurocomputing*, Vol. 239, pp. 39–57, 2017.

[60] Werner J. Reinartz and Vijay Kumar. On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. *Journal of Marketing*, Vol. 64, No. 4, pp. 17–35, 2000.

[61] Werner J. Reinartz and Vita Kumar. The impact of customer relationship characteristics on profitable lifetime duration. *Journal of Marketing*, Vol. 67, No. 1, pp. 77–99, 2003.

[62] Shota Sato and Asahi Yumi. The model of purchasing and visiting behavior of customers in an e-commerce site for consumers. In *International Proceedings of Economics Development & Research*, Vol. 52, pp. 72–76, 2014.

[63] Catarina Sismeiro and Randolph E. Bucklin. Modeling purchase behavior at an e-commerce web site: A task-completion approach. *Journal of Marketing Research*, Vol. 41, No. 3, pp. 306–323, 2004.

[64] Robert E. Stake. Situational context as influence on evaluation design and use. *Studies in Educational Evaluation*, Vol. 16, No. 2, pp. 231–246, 1990.

[65] Bartolomeo Stellato, Goran Banjac, Paul Goulart, Alberto Bemporad, and Stephen Boyd. OSQP: An operator splitting solver for quadratic programs. In *2018 UKACC 12th International Conference on Control (CONTROL)*, pp. 339–339, 2018.

[66] Dek Terrell. Incorporating monotonicity and concavity conditions in flexible functional forms. *Journal of Applied Econometrics*, Vol. 11, No. 2, pp. 179–194, 1996.

[67] Ryan J. Tibshirani, Holger Hoefling, and Robert Tibshirani. Nearly-isotonic regression. *Technometrics*, Vol. 53, No. 1, pp. 54–61, 2011.

[68] Efraim Turban, Jon Outland, David King, Jae Kyu Lee, Ting-Peng Liang, and Deborrah C. Turban. *Electronic Commerce 2018: A managerial and social networks perspective*. Springer, 2017.

[69] Dirk Van den Poel and Wouter Buckinx. Predicting online-purchasing behaviour. *European journal of Operational Research*, Vol. 166, No. 2, pp. 557–575, 2005.

[70] Dirk Van den Poel and Bart Lariviere. Customer attrition analysis for financial

services using proportional hazard models. *European Journal of Operational Research*, Vol. 157, No. 1, pp. 196–217, 2004.

[71] Jiangdian Wang and Sujit K. Ghosh. Shape restricted nonparametric regression with bernstein polynomials. *Computational Statistics & Data Analysis*, Vol. 56, No. 9, pp. 2729–2741, 2012.

[72] Michel Wedel and Wayne S. DeSarbo. A review of recent developments in latent class regression models. *Advanced Methods of Marketing Research, R. Bagozzi (Ed.), Blackwell Pub*, pp. 352–388, 1994.

[73] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in Statistics*, pp. 196–202. Springer, 1992.

[74] Yongzheng Zhang and Marco Pennacchiotti. Predicting purchase behaviors from social media. In *Proceedings of the 22nd International Conference on World Wide Web*, pp. 1521–1532, 2013.

[75] 秋山純, 松本健, 西村直樹. インセンティブ付与対象決定システム及びプログラム. 日本国特許庁 (JP), 特開 2018-13827, 2018.

[76] 猪狩良介, 星野崇宏. Online–offline チャネルにおける消費者の購買間隔と購買金額の同時モデリング. オペレーションズ・リサーチ: 経営の科学, Vol. 61, No. 9, pp. 589–599, 2016.

[77] 小西葉子. 存続時間分析による美容院顧客の来店確率予測 (特集 予測と発見). 統計数理, Vol. 54, No. 2, pp. 445–459, 2006.

[78] 西村直樹, 小林健, 吉住宗朔. 制約つき比例ハザードモデルを用いたヘアサロンの再来店状況分析. オペレーションズ・リサーチ: 経営の科学, Vol. 64, No. 2, pp. 65–72, 2019.

[79] 山口景子. 頻度の時間変化を考慮した階層ベイズモデルによるウェブサイト訪問行動の分析. マーケティング・サイエンス, Vol. 22, No. 1, pp. 13–29, 2014.