ソシオネットワーク戦略ディスカッションペーパーシリーズ　ISSN 2434-9445
第 107 号　2023 年 3 月
RISS Discussion Paper Series
No.107　March, 2023

メディア選択データを利用した同時セグメンテーション
～無限関係モデルによる特徴的な視聴行動の抽出～

里村卓也

文部科学大臣認定　共同利用・共同研究拠点

関西大学ソシオネットワーク戦略研究機構

Research Institute for Socionetwork Strategies,
Kansai University
Joint Usage / Research Center, MEXT, Japan
Suita, Osaka, 564-8680, Japan
URL: https://www.kansai-u.ac.jp/riss/index.html
e-mail: riss@ml.kandai.jp
tel. 06-6368-1228
fax. 06-6330-3304

メディア選択データを利用した同時セグメンテーション
～無限関係モデルによる特徴的な視聴行動の抽出～[1]

Simultaneous Segmentation Based on Media Choice Data: Extraction of Distinctive Viewing Behaviors Using Infinite Relational Model

慶應義塾大学　里村卓也

概要

本研究では無限関係モデル(IRM)を利用したメディア選択行動の同時セグメンテーションを行った。TV 番組×視聴者からなる大規模な視聴の有無に関するデータに対して IRM による共クラスタリングを行うことで、TV 番組と視聴者の特徴のある関係を抽出して具体的な視聴者像を獲得することができる。さらに、IRM による共クラスタリングではノンパラメトリックベイズ手法を利用したセグメント数の自動決定も行うことができる。実証分析では、消費者パネルによるテレビ視聴データへ IRM を適用した。崩壊型ギブスサンプリングを用いてモデルの推定を行い、セグメント数とパラメータの推定値を得た。この推定値を用いた分析結果からは特定のニュース番組やドラマ番組などの番組セグメントが形成され、それらの番組群の視聴確率が高い視聴者のセグメントを抽出することができた。

1.はじめに

具体的な消費者像を得ることはマーケティング実務において重要である。このような具体的な消費者像を得ることは、近年ではマーケティング実務においてはペルソナの獲得として実践されている(Herskovitz and Crystal 2010, Aimé, Berger-Remy and Laporte 2022)。本研究ではそのような消費者の行う活動のなかでも TV 番組の視聴行動について、具体的な視聴行動を大規模なデータの中から抽出することを試みる。このような視聴者の具体像は、各番組のコアとなる視聴者を把握することが必要な番組やメディアの提供側にとっても役立つ情報である。そこで本研究では、TV 番組視聴データを利用して番組と視聴者の同時セグメンテーションを行い特徴的な視聴行動を抽出することを目的とする。特徴的な視聴行動の抽出を行うことで具体的な消費者像を得ることも可能となる。

具体的な消費者像を得るための方法としてために、本研究では Kemp, Tenenbaum,

[1] この研究は関西大学ソシオネットワーク戦略研究機構の共同利用・共同研究によって行った。

Griffiths, Yamada and Ueda (2006) による無限関係モデル(Infinite Relational Model: 以下IRM)を利用して、具体的な視聴行動を把握していく。IRM を利用することで、TV 番組の個人別視聴データ（行動データ）から TV 番組と視聴者の同時セグメンテーションを行って具体的な視聴者像を得る。

同時セグメンテーションでは変数と個体の分類を同時に行う。マーケティング・リサーチにおいて伝統的に利用されてきたクラスター分析は、複数の変数の値をもとに個体間の類似性を評価して個体を分類する、あるいは、複数の個体の値をもとに変数間の類似性を評価して変数を分類するものである。それに対して、同時セグメンテーションでは個体と変数とを同時に分類するものである。本研究においては TV 番組と視聴者とを同時に分類する。このようなセグメンテーション方法においては、番組と視聴者からなるブロックを形成する。このブロック化によりある番組グループの視聴確率が高い視聴者グループというブロックができあがり、このブロックを利用することで番組と視聴者の強い関係を見出すことが可能となる。このような同時セグメンテーションを行うために IRM を用いる。

本研究で利用する IRM を用いたセグメンテーションの特徴として１）生成モデルとしての妥当性、２）ハードクラスタリングによるセグメント間での番組や視聴者の重複の回避、３）セグメント数の自動決定、がある。

１番目の生成モデルとしての妥当性であるが、同時セグメンテーションを行う方法としては、機械学習の手法であるトピックモデル（Blei, Ng and Jordan 2003）を利用することも考えられる。トピックモデルを利用した研究として、里村(2020）では複数メディアの利用頻度データに対してトピックモデルを適用することで、トピックとして TV 番組やウェッブサイトを分類し、さらにトピックに対応させる形で消費者の分類を行っている。しかしながら日別番組別の視聴の有無を変数として考えると、そのような変数は視聴したか否かとい２値のカテゴリカル変数となる。もちろん、同じ番組名で毎日あるいは毎週繰り返しで方法されるニュースやドラマ、バラエティ番組などであれば、変数の値は頻度（視聴回数）となるので多項分布を用いたトピックモデルを利用することもできる。しかしながら、特番やスポーツのように基本的に１度しか放映されない番組は、頻度が０か１のみであるため、これらの番組も含めた視聴データを分析するためにはトピックモデルのような２回以上の頻度も考えたモデルを用いることは、データの生成過程から考えても適切ではない。これに対して IRM を用いたセグメンテーションでは、２値データをそのまま利用するため、生成モデルとして適切であるといえる。

２番目のセグメント間での番組や視聴者の重複の回避であるが、トピックモデルも含む潜在クラスモデルでは番組や個人は複数のセグメントに確率的に所属するソフトクラスタリング手法である。ソフトクラスタリング手法は番組視聴行動を分析するため、これまでの研究においても用いられてきた（Rust, Kamakura and Alpert 1992；Danaher and Mawhinney 2001)。しかしながら、ソフトクラスタリング手法では視聴者と番組の組み合わせにおいても、多くのそれらが混合した平均的な値でしか解釈することができず、具体的な解釈をする

ことが困難である。一方、本研究で利用する IRM はハードクラスタリング手法であるため、セグメント間での重複はない。そのためセグメントの特徴を把握するためには限定された番組や視聴者のみを利用できるため、より特徴を際立たせた分析が可能になると考えらえる。

３番目のセグメント数の決定であるが IRM では、セグメント数を自動的に決定することができる。潜在クラス分析やクラスター分析では手法を適用する前に、事前にセグメント数を固定しておく必要がある。そのため、異なるセグメント数で分析をしておき、最後に結果を比較してセグメント数を決定することになる。これに対して IRM ではノンパラメトリックベイズ手法を利用することで、データに応じてセグメント数を自動的に決定することができる。そのために自動的な分析が可能となる。

本研究では IRM の手法を利用して TV 番組視聴データから番組と視聴者の同時セグメンテーションを行う。この手法を用いることである番組グループの視聴確率が高い視聴者グループという、番組と視聴者の強い関係を見出すことが可能となる。

以降の本論文の構成は次のようになる。２.では同時セグメンテーションで用いる手法である共クラスタリングについての説明を行う。３．では IRM を利用したセグメンテーションを実行するアルゴリズムについて説明する。４．では実証分析とその結果について説明し、５．ではまとめについて述べる。

２．共クラスタリングについて

2.1 共クラスタリングの考え方

同時セグメンテーションで用いる手法は、共クラスタリング(co-clustering)やバイクラスタリング(bi-clustering)、ブロッククラスタリング(block-clustering)と呼ばれている。石井・上田(2014) によると共クラスタリングとは、異種オブジェクトを同時にクラスタリングする手法である。顧客と商品を共クラスタリングすることで、顧客群と商品群の関係が抽出される。この手法は最初に Hartigan(1972)によって提案されたが、Cheng and Church(2000)により遺伝子データの解析で注目された。そして近年は生物医学分野を中心に応用が進んでいる（Xie, Ma, Fennell, Ma and Zhao 2019)。共クラスタリングの手法はマーケット・セグメンテーション（Carmore, Kara and Maxwell 1999, Frances and Ghose 2016, Kaiser, Dolnicar, Lazarewski and Leischc 2017, 山田・佐藤 2020）でも利用がなされている。

図１に共クラスタリングの簡単な例を示す。この例では行と列で異なるオブジェクトであると考える。また、行と列の組み合わせにおいて、ある変数が０もしくは１の値をとるものとする。例えば、図１の左側のように、行をテレビ番組、列を視聴者、それらの組み合わせにより各視聴者による各番組視聴の有無（１は視聴、０はそれ以外）を表現するとする。このような番組と視聴者の組み合わせについて、行と列の入れ替えを行い、その結果、図１の

右側のような並び替えを行ったとする。このとき、図の右側のようにクラスタリング後には列で 2 グループ、行で 2 グループに分かれており、合計 4 ブロックで構成されるように並び替えが行われている。このときブロック内がなるべく同じ値になるように並び替えられていることがわかる。

>>図 1 挿入<<

さて、それぞれのブロックでの選択確率は図 2 のようになる。ただし$\theta(i,j)$は番組ブロックiにおける視聴者ブロックjでの視聴確率である。$\theta(i,j)$の値をみると 0 もしくは 1 に近くなっており、行と列の入れ替えによって、特徴のある組み合わせが得られたことがわかる。

>>図 2 挿入<<

2.2 IRM について

無限関係モデル(IRM)は共クラスタリングデータを確率的に生成するモデル（Kemp, Tenenbaum, Griffiths, Yamada and Ueda 2006）である。IRM を利用することで、共クラスタリングを行うことができる。IRM は所属クラスの事前分布がディリクレ過程のひとつである中華料理店過程(Chinese Restaurant Process: 以下 CRP)に従うと考えたノンパラメトリックベイズモデル（石井・上田 2014）となっている。そのため、本研究においても、番組と視聴者のセグメント数を自動的に決定することができるため、事前にセグメント数を設定する必要がない。

ノンパラメトリックベイズモデルは無限個のパラメータを持つことができるモデルである。IRM では番組や視聴者が各クラスへ割り振られる過程において CRP を利用することで、事前分布として無限次元でのセグメント数を想定することができる。CRP において新しい個体をクラスに割り当てる際には、既にあるクラスへ新しい個体を割り当てる、あるいは、新しくクラスを作ってそこに新しい個体を割り当てる、という操作を行う。このとき既存の個体数が多いクラスほど、新しい個体が割り当てられる確率が高くなる。一方で、既存の個体数がゼロの新しいクラスであっても、ある確率で新しい個体が割り当てられるようになっている。

3. IRM を利用した視聴番組選択行動のモデル化と推定

ここでは、石井・上田（2014）による IRM を用いた共クラスタリングに従って視聴番組選択行動のモデル化を行い、次にその推定方法について述べる。

3.1 IRM による視聴番組選択行動のモデル化

番組$k(=1,\cdots,K)$における視聴者$l(=1,\cdots,L)$の視聴の有無について考える。R_{kl}は視聴者lが番組kを視聴した場合には 1、それ以外では 0 をとる変数とする。またRをk行l列目の要素がR_{kl}である、大きさK行L列の行列とする。

さてここで、番組と視聴者について複数のクラスを想定する。番組のセグメント数をc^1、視聴者のセグメント数をc^2とする。s_k^1を番組kの所属クラス、s_l^2を視聴者lの所属クラスとする。s_k^1とs_l^2は潜在変数であり、直接観測はできないものとする。また$s^1=\{s_k^1,\ k=1,\cdots,K\}$、$s^2=\{s_l^2,\ l=1,\cdots,L\}$とする。番組$k$の所属クラスは$s_k^1=\{w_i^1,\ i=1,\cdots,c^1\}$の値をとり、視聴者$l$の所属クラスは$s_l^2=\{w_j^2,\ j=1,\cdots,c^2\}$の値をとるものとする。

$\theta_{kl}=\theta(s_k^1,s_l^2)$は番組$k$おける視聴者$l$の視聴確率であり、これは番組$k$の所属クラス$s_k^1$と視聴者$l$の所属クラス$s_l^2$によって決まるものとする。視聴確率は番組と視聴者の組み合わせで決まるが、同じクラスに所属する番組と視聴者の集合は、同じパラメータ（ブロックパラメータ）を持つものと仮定する。またΘはi行j列目の要素が$\theta(i,j)$である、大きさc^1行c^2列の行列とする。αを CRP のパラメータ、a,bをベータ分布のパラメータとする。

モデルでは、まず、所属クラスは CRP に従って決まるものとする。

$$s^1|\alpha\sim CRP(\alpha)$$
$$s^2|\alpha\sim CRP(\alpha)$$

次に、ブロックパラメータ$\theta(s_k^1,s_l^2)$はベータ分布に従うとする。

$$\theta(s_k^1,s_l^2)|a,b\sim Beta(a,b)$$

さらに、視聴の有無はベルヌーイ分布に従うとする。

$$R(k,l)|s_k^1,s_l^2,\Theta\sim Bernulli(\theta_{kl})$$

このときの観測データの尤度$P(R|s^1,s^1)$は次のとおりになる。

$$P(R|s^1,s^2)=\prod_{k=1}^{K}\prod_{l=1}^{L}\theta(s_k^1,s_l^2)^{R(k,l)}\{1-\theta(s_k^1,s_l^2)\}^{1-R(k,l)}$$

さらに、s^1とs^2の同時事後分布$P(s^1,s^2)$は以下のようになる。

$$P(s^1,s^2)\propto P(s^1)P(s^2)P(R|s^1,s^2)$$

また、以降での説明のためにνを以下のように定義しておく。

$$\nu=P(s^1)P(s^2)P(R|s^1,s^2)$$

3.2. IRM モデルの推定

IRM モデルの推定は、石井・上田（2014）と同様に崩壊型ギブスサンプリングでモデルを推定する。推定ステップは以下の通りである。

Step 1：初期設定

$s^1,\ s^2$の初期値を設定する。本研究では、まず CRP から乱数を発生させ、この乱数でのセグメント数$c^1,\ c^2$を求める。このセグメント数を所与として顧客、番組をそれぞれについて k-mean 法によりクラスタリングを行い、これを$s^1,\ s^2$の初期値とする。初期値での

νを計算しν_max とする。

Step 2：所属クラスの更新

まず$k = 1, \cdots, K$についてs_k^1を更新する。kを現在の所属クラスから除外する。このために空きクラスが発生すれば空きクラスを削除する。s_k^1の更新では、CRP に従いs_k^1を確率的に決定する。もしkの所属クラスが新規クラスの場合にはセグメント数を更新する。

次に$l = 1, \cdots, L$についてs_l^2を更新する。lを現在の所属クラスから除外する。このために空きクラスが発生すれば空きクラスを削除する。s_k^2の更新では、CRP に従いs_k^2を確率的に決定する。所属クラスが新規クラスの場合にはセグメント数を更新する。

Step 3：事後確率最大化

現時点でのs^1，s^2を用いてνを計算する。

νを計算し$\nu >$ ν_max ならs^1，s^2を更新する。それ以外であれば更新前のν_max とs^1，s^2を維持する。

Step 4：終了判定

ν_max が更新されない状態が十分に継続した場合、ギブスサンプリングを終了する。それ以外の場合には、Step2 に戻る。

４．個人別視聴データを用いた実証分析

4.1. データについて

個人別テレビ番組の資料データとして野村総合研究所　Insight Signal データ 2020 年 2-3 月調査を利用した。調査時期は 2020 年 1 月 25 日～2020 年 4 月 4 日である。調査期間中の日数は 71 日間である。このうち、週末（土日）を含まない日数は 50 日間であり、さらに週末（土日）と祝日を含まない日数は 47 日間である。

このデータは関東（１都６県）在住の 16～69 歳の男女約 3,000 名に PC、携帯にて調査を行ったものである。データには媒体接触状況、広告出稿状況、購買プロセス状況がある。分析対象のテレビ番組と視聴者としては、テレビ番組は視聴者数が 200 以上の 1061 番組（全番組数は 13,799）とした。また視聴者は、上記番組の視聴番組数が 300 以上 1000 未満の 1305 名（全視聴者数は 2,914 名）とした。分析対象者の年齢分布は図 3 のとおりである。

対象者を抽出する前のデータでは国勢調査での人口比率に応じて性別年齢区分別のサンプリングがされていたが、視聴番組数で対象者を選定した結果では、図 3 のように男性において 26 歳から 40 歳までのサンプル数が女性と比べて少なくなっている。

>>図 3 挿入<<

図 4 は個人別の番組視聴状況である。この黒い点がある部分は、当該行の視聴者が当該列の番組を視聴したことを表しており、白い部分は逆に視聴していないことを表している。個人により視聴番組の番組や総視聴数に違いがあることがわかる。

>> 図 4 挿入 <<

4.2 モデル推定の結果

崩壊型ギブスサンプリングを利用してモデルの推定を行った。図 5 はサンプリングの繰り返しごとの$\log(\nu)$の値である。本研究における実証分析では、事後確率が最大となったサンプリング結果を推定値として採用することとした。事後確率が最大となるときには$\log(\nu)$も最大となる。そこでサンプリングを 100 回繰り返し、$\log(\nu)$が最大となる番組セグメント数 26 と視聴者セグメント数 448 を得た。

>>図 5 挿入<<

図 6 は番組と個人の共クラスタリング結果を表している。図 4 と同じデータを共クラスタリングの結果をもとに並べ替えたものが図 6 の(a)である。図ではセグメントサイズが大きい順に、番組は左から、視聴者は下から並べてある。各セグメントの境界は黒点もしくは白点の密度の違いにより確認できる。各セグメントの幅はセグメントに含まれる番組あるいは視聴者の数を表している。この図からは、番組セグメントと視聴者セグメントの組み合わせによって視聴の有無が明確に分かれた結果を得られたことがわかる。図 6 の(b)は、各ブロックのパラメータの値をヒートマップで表示したものである。ブロックパラメータは 0 から 1 の間の値をとるが、色が濃いほど、パラメータの値が 1 に近いことを表現している。また、セグメントの境界は番組・視聴者とも 5 アイテム以上を表示している。ヒートマップからも特定の視聴者セグメントにおいて特定の番組セグメントの視聴確率が高いことが分かる。

>>図 6 挿入<<

表 1 は番組セグメント別の番組数と番組内容である。番組セグメントは番組数が多い順に番号が付与されている。番組内容では、例えば A1 と A2 のようにアルファベットが同じで続く数字が異なるものは、同じ番組名で異なる時間帯に分かれているもの、あるいは同じ番組名で同じ時間帯であっても、日によって異なる番組セグメントに所属しているものを表現している。

表 1 を見ると、上位 3 位までの番組セグメントでは番組数が 100 を超えていることがわかる。番組セグメント 1 はバラエティやスポーツ、民放の夜のニュースなどが含まれ番組数は 200 を超える。番組セグメント 2 は NHK のニュースとバラエティ 2 番組という、NHK

の番組のみから構成されている。番組セグメント 3 は休日のバラエティ、情報、アニメなど、休日に放映される番組からなる。このように上位 3 つのセグメントには多くの番組が含まれているが、セグメント 2 とセグメント 3 では含まれる番組において特徴がみられる。

次の 4 位から 10 位までの番組セグメントでは番組数が 47 以上である。なお 47 という値は、4.1 でも述べたように、対象期間のうち週末（土日）と祝日を含まない日数である。これらの番組セグメントの中にはドラマやニュース・情報のように同じ番組名で週日に放映される番組でセグメントを形成しているものもある。IRM を利用して番組と視聴者の特徴のある組み合わせを抽出することで、番組名のような事前情報をもちいることなく、同じ番組が一つのセグメントに含まれるようなセグメンテーションを行うことができていることが確認できた。番組名をもとにした集計を行わずとも視聴行動の特徴から同じ番組が同じセグメントに分類できていることが確認できた。

>>表 1 挿入<<

図 7 は番組・視聴者セグメント毎の視聴確率（ブロックパラメータの値）の分布である。図では番組セグメント別に視聴者セグメント別の視聴確率の分布を箱ひげ図で表現している。これを見ると、番組セグメントによって視聴者セグメント別の視聴確率にばらつきがあることがわかる。視聴者セグメント別の視聴確率について、各番組セグメントでの中央値は箱の中の横線で表現されているが、それらの多くはゼロに近い値をとっている。ただし、いくつかの視聴者セグメントでは視聴確率が高い値をとっているものもある。例えば番組セグメント 8 や 12 においては、中央値はゼロに近いが、いくつかの視聴者セグメントにおいて視聴確率が高く、それらの中には 0.8 以上の値をとっているものも見られる。

>>図 7 挿入<<

そこで視聴者セグメントによる視聴確率のばらつき度合いを把握するために、番組セグメント別に視聴確率の平均値と中央値を表したのが図 8 である。これを見ると、視聴確率の平均値と中央値が近いものもあれば、乖離しているものもある。視聴者セグメントによる視聴確率のばらつきを評価するために、ばらつき度として視聴確率の中央値÷平均値の値を用いる。この値が大きいほど、視聴者セグメント間での視聴確率のばらつきは小さくなる。視聴確率の平均値が中央値の 2 倍以下の場合にはばらつき度の値は 0.5 以上となる。そこで、ばらつき度の値が 0.5 を超えているか否かで視聴確率のばらつき度を判定することとする。中央値÷平均値が 0.5 以上である、すなわち視聴者セグメント間での視聴確率のばらつきが小さい番組セグメントは 1, 3, 4, 9, 11, 14, 17, 20, 21 であり、これらはバラエティやドラマの一部、休日ニュース、音楽・映画などが含まれる。一方、中央値÷平均値の値が 0.5 未満よりも小さい、すなわち視聴者セグメント間での視聴確率のばらつきが大きい番組セグメントは 2, 5, 6, 7, 8, 10, 12, 13, 15, 16, 18, 19, 22, 23, 24, 25, 26 であり、これらは NHK 番

組全般、民放ニュース、ドラマの一部などが含まれるである。後者のようにばらつき度が大きいセグメントは、コアな視聴者がいる一方で、その他多くの視聴者にとっては視聴されない番組セグメントであるといえる。

>>図 8 挿入<<

図 9 は視聴者セグメント毎の視聴者数である。視聴者セグメントの番号は、人数が多い順に番号を付与してある。視聴者セグメント 1 は 23 名、視聴者セグメント 2 から 5 は 13 名となっており、上位１ 4 位までは視聴者数が 10 名以上である。一方で、視聴者数が少ないセグメントも多く、視聴者数が 1 名のセグメントは 181、視聴者数が 2 名のセグメントは 86、視聴者数が 3 名のセグメントは 58、視聴者数が 4 名のセグメントは 38 ある。これら 4 名以下のセグメントは視聴者セグメント数の 81.0%を、全体人数の 52.0%を占める。

>>図 9 挿入<<

図 10 は視聴者セグメント別の番組セグメント視聴確率である。この図では視聴者セグメント 1 と 2 について表示してある。視聴者セグメント 1 は女性比率が 26.1%、平均年齢は 55.1 歳の 23 名からなる。この視聴者セグメントでは番組セグメントでは 2（NHK ニュース、NHK バラエティ A、NHK バラエティ B)、18（NHK ドラマ B)、19（民放休日ニュース A）の視聴確率が高くなっている。このように視聴番組と性別年齢の構成に特徴があるセグメントとなっている。視聴者セグメント 2 は女性比率が 46.2%、平均年齢は 53.1 歳の 13 名からなる。この視聴者セグメントでは番組セグメントでは 6（民放夜ニュース B)、11（休日ニュース)、15（ドラマ群 A）の視聴確率が高くなっている。 2 つの視聴者セグメントをとってみても、全く異なる番組セグメントを視聴していることが分かる。IRM を用いた同時セグメンテーションによって視聴番組と性別年齢の構成に特徴があるセグメントをブロックとして抽出することができていることが確認できた。

>>図 10 挿入<<

以上のように、IRM を用いた同時セグメンテーションにより、番組視聴の具体的特徴が容易に把握できるセグメンテーションを行うことができた。番組セグメントを理解するうえでは、潜在クラスモデルを用いたセグメンテーションのようなクラス間での番組の重複がないために、セグメントの理解が容易となっている。また視聴者セグメントにおいても、特徴のある消費者の姿を具体的に得ることができる。

６．おわりに

本研究では無限関係モデル(IRM)を利用したメディア接触行動の同時セグメンテーションを行った。TV 番組×視聴者からなる大規模な視聴の有無に関するデータに対して IRM による共クラスタリングを行うことで、TV 番組と視聴者の特徴のある関係を抽出することを試みた。このような共クラスタリングから具体的な視聴者像を獲得することができる。さらに、IRM による共クラスタリングではノンパラメトリックベイズ手法を利用することでセグメント数を自動決定するもできる。

実証分析では、消費者パネルによるテレビ視聴データへ IRM を適用した。崩壊型ギブスサンプリングを用いてモデルの推定を行い、事後確率が最大となったサンプリング結果を採用し、セグメント数とパラメータの推定値を得た。この分析結果からは特定のニュース番組やドラマ番組などの番組セグメントが形成され、それらの番組群の視聴確率が高い視聴者のセグメントを抽出することができた。このように番組と消費者をブロックとして抽出する手法は視聴者像を具体的に把握するために有効であることが確認できた。

今後の課題としては、次の点があげられよう。まずは、視聴者セグメントにおいて、少数の視聴者からなるセグメントが多数存在していることである。本研究で用いた手法はハードクラスタリングであり番組や視聴者が複数のセグメントへ所属することを許していないため、他の視聴者と類似した視聴行動をとらない視聴者はたとえ１名であっても新しいセグメントを作ってしまう。本研究の目的のように、番組と視聴者の組み合わせで特徴があるセグメントを抽出するためにはそのような特性は問題とはならないが、視聴行動の予測やプロモーションを行う場合には、少数の視聴者からなるセグメントが多数存在する場合には活用の上で課題があるといえる。また、本研究では TV 番組の視聴の有無というデータを用いたが、商品選択等の異なるデータでの応用や、頻度や金額のようなカテゴリカル２値データ以外への対応も今後の課題といえよう。

参考文献

Aimé, I., F. Berger-Remy and M.E. Laporte (2022) “The brand, the persona and the algorithm: How datafication is reconfiguring marketing work,” Journal of Business Research, 145, 814–827.

Blei, D.M., A.Y. Ng and M.I. Jordan (2003) “Latent Dirichlet Allocation,” Journal of Machine Learning Research, 3, pp.993–1022.

Carmone, F. J., A. Kara, and S. Maxwell (1999) “HINoV: A New Model to Improve Market Segment Definition by Identifying Noisy Variables.” Journal of Marketing Research, 36(4), 501–509.

Cheng, Y., & G.M. Church (2000) “Biclustering of Expression Data,” Proceedings. International Conference on Intelligent Systems for Molecular Biology, 8, 93-103.

Danaher, P. J., and D.F. Mawhinney (2001) “Optimizing television program schedules using choice modeling,” Journal of Marketing Research, 38(3), 298–312.

France , S. L., and G. Sanjoy (2016) “An Analysis and Visualization Methodology for Identifying and Testing MarketStructure,” Marketing Science, 35(1), 182-197.

Hartigan, J. A. (1972) “Direct Clustering of a Data Matrix,” Journal of the American Statistical Association, 67(337), 123-129.

Herskovitz, S. and M. Crystal (2010) “The essential brand persona: storytelling and branding, Journal of Business Strategy, Vol. 31 No. 3, pp. 21-28.

Kaiser S., S. Dolnicar, K. Lazarevski and F. Leisch (2017) “Overcoming data dimensionality problems in market segmentation,” in Applied Biclustering Methods for Big and High-Dimensional Data Using R (A. Kasim, Z. Shkedy, S. Kaiser, S. Hochreiter and W. Talloen ed.), Chapman & Hall.

Kemp C., J.B. Tenenbaum, T.L. Griffiths, T. Yamada and N. Ueda (2006) “Learning systems of concepts with an infinite relational model,” Proceedings of the 21st national conference on artificial intelligence, 381-388.

Rust, R. T., W.A. Kamakura and M.I. Alpert (1992) “Viewer preference segmentation and viewing choice models for network television,” Journal of Advertising, 21(1), 1–18.

Xie, J., A. Ma , A. Fennell, Q. Ma, J. Zhao (2019) “It is time to apply biclustering: a comprehensive review of biclustering applications in biological and biomedical data,” Briefings in Bioinformatics, 20(4), 1450–1465.

石井健一郎・上田修功(2014)「続・わかりやすいパターン認識　－教師なし学習入門ー」オーム社

里村卓也(2020)「消費者の複合的メディア消費行動の統合的分析モデル」オペレーションズ・リサーチ, 65-2, 76 – 84.

山田浩喜・佐藤忠彦(2020)「ポアソン潜在ブロックモデルによるドラッグストアにおける購買データの解析」行動計量学, 47-2, 161-172.

表 1　番組セグメント別の番組数と番組内容

番組セグメント	番組数	番組内容
1	208	バラエティ,アイドル, 民放夜ニュースA, スポーツ
2	138	NHKニュース, NHKバラエティA, NHKバラエティB
3	125	休日バラエティ, 休日情報, 休日アニメ
4	62	平日19-21時
5	61	NHKドラマA
6	50	民放夜ニュースB
7	50	民放朝ニュースA 1
8	48	民放朝ニュースB 1
9	47	プライムタイムドラマ群
10	47	民放情報A1
11	38	休日ニュース群
12	36	民放朝ニュースA 2
13	21	民放夕ニュースA
14	19	バラエティ群A
15	17	ドラマ群A
16	16	ドラマ群B
17	11	音楽特番, 映画
18	10	NHKドラマB
19	10	民放休日ニュースA
20	10	バラエティ番組A
21	9	ドラマA
22	8	ドラマ深夜A
23	8	休日ニュースA
24	5	民放朝ニュースB2,民放朝ニュースB3
25	3	民放情報A2
26	2	NHKバラエティA[再]

図１　共クラスタリングの例

(a)共クラスタリング前

	1	2	3	4	5	6
1	1	1	0	0	0	1
2	0	0	1	1	1	0
3	0	0	1	1	1	0
4	0	0	1	0	1	1
5	1	1	0	0	1	0
6	1	1	0	0	0	1

(b) 共クラスタリング後

	3	4	5	1	2	6
2	1	1	1	0	0	0
3	1	1	1	0	0	0
4	1	0	1	0	0	1
1	0	0	0	1	1	1
5	0	0	1	1	1	0
6	0	0	0	1	1	1

図２　図１での共クラスタリング後の各ブロックでの選択確率

	3	4	5	1	2	6
2						
3	$\theta(1,1)=\frac{8}{9}$			$\theta(1,2)=\frac{1}{9}$		
4						
1						
5	$\theta(2,1)=\frac{1}{9}$			$\theta(2,2)=\frac{8}{9}$		
6						

図３　分析対象者の年齢分布

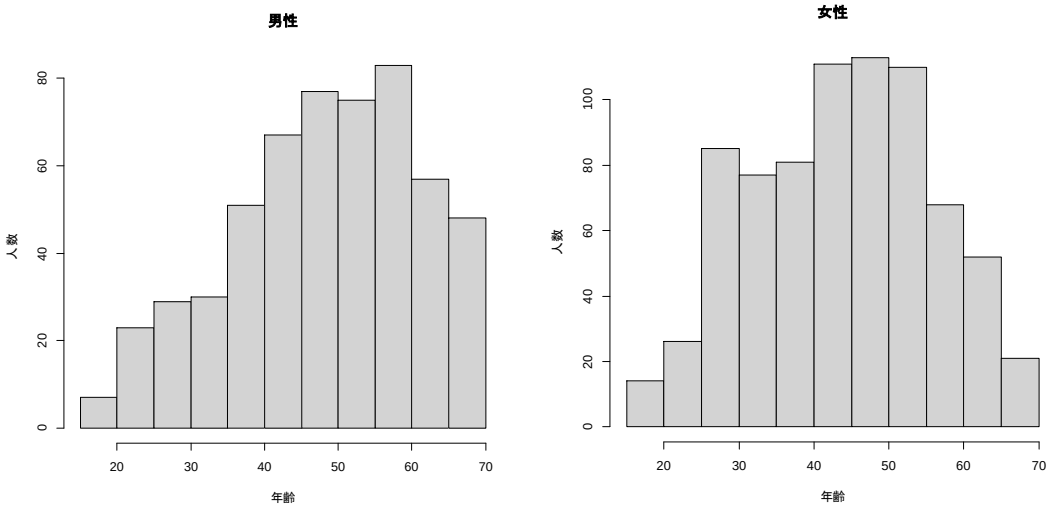


図４　個人別の番組視聴状況

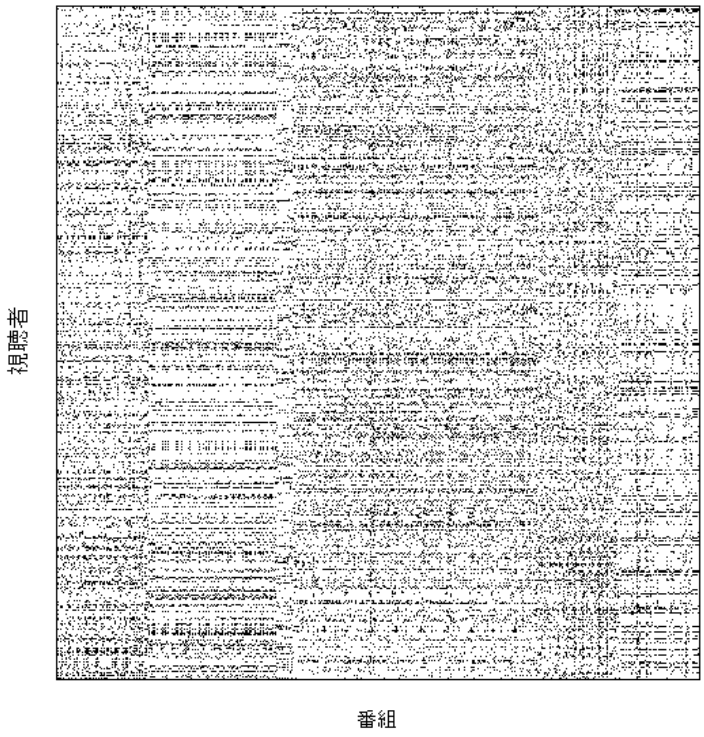


図５ サンプリングの繰り返し数とlog (ν)の変化

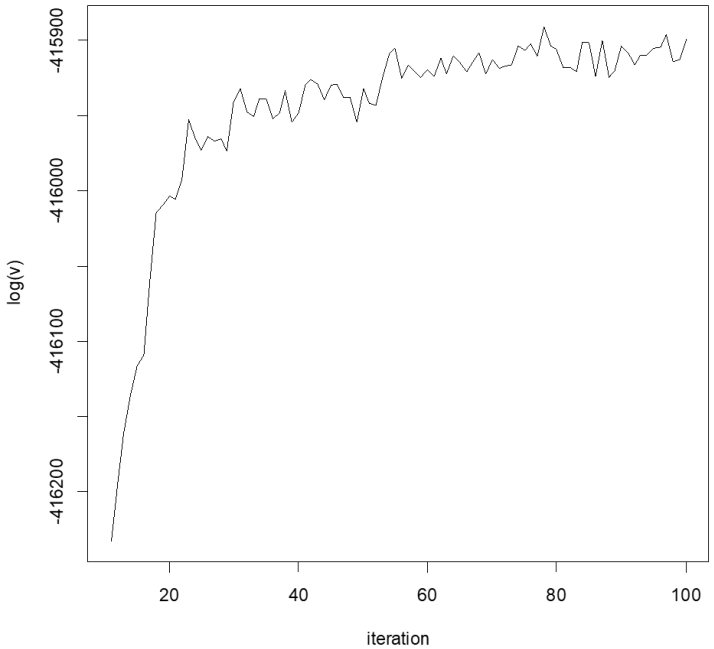
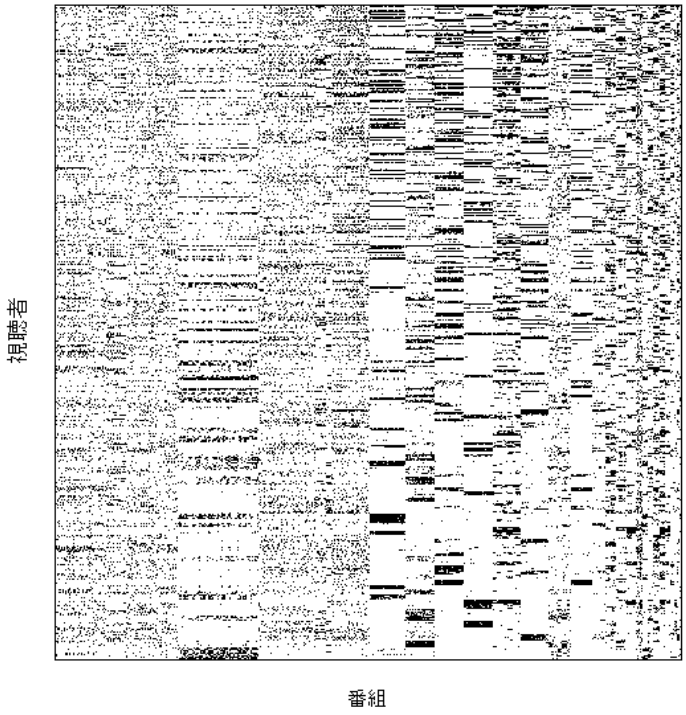
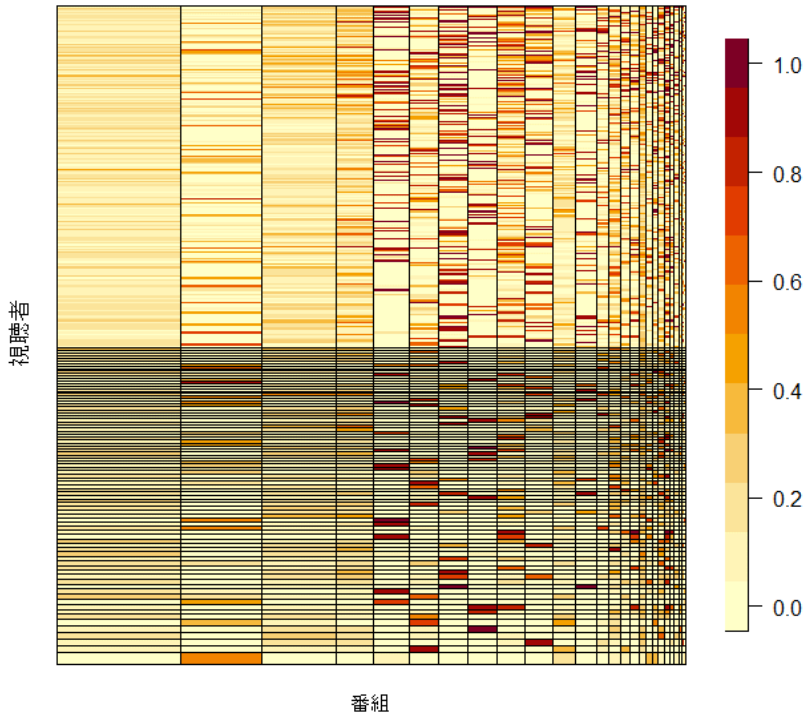


図６　番組と個人の共クラスタリング結果

(a) セグメント別で並べ替えられた番組と視聴者における番組視聴状況



(b) 各ブロックのパラメータの値



※セグメントの境界は番組・視聴者とも５アイテム以上を表示

図７　番組・視聴者セグメント毎の視聴確率（ブロックパラメータの値）の分布

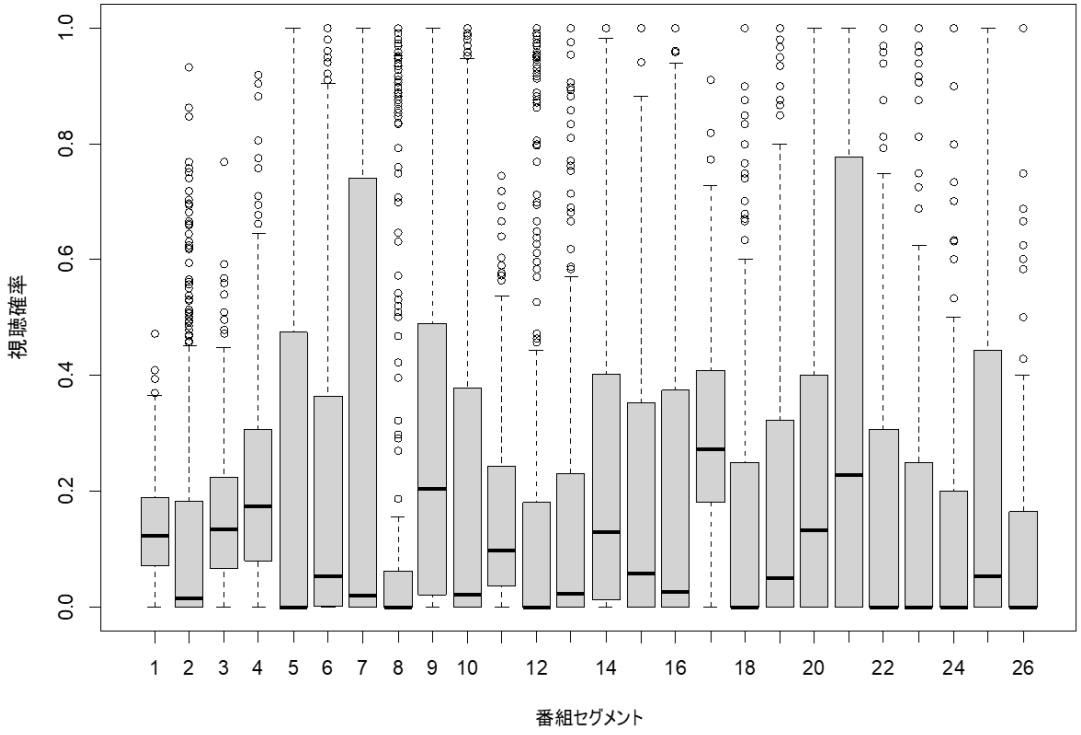


図８　視聴確率の平均値と中央値

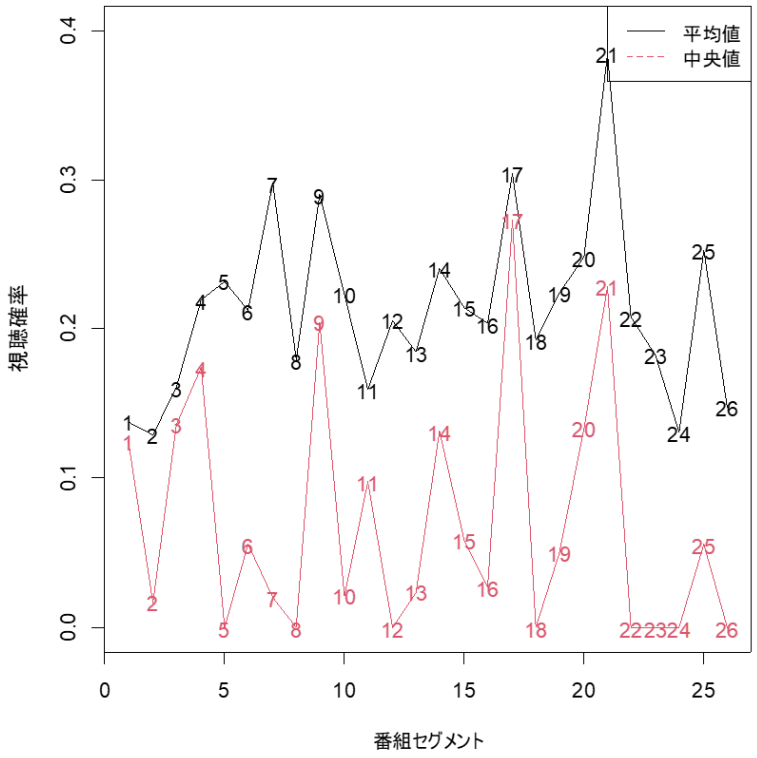


図９　視聴者セグメント毎の視聴者数

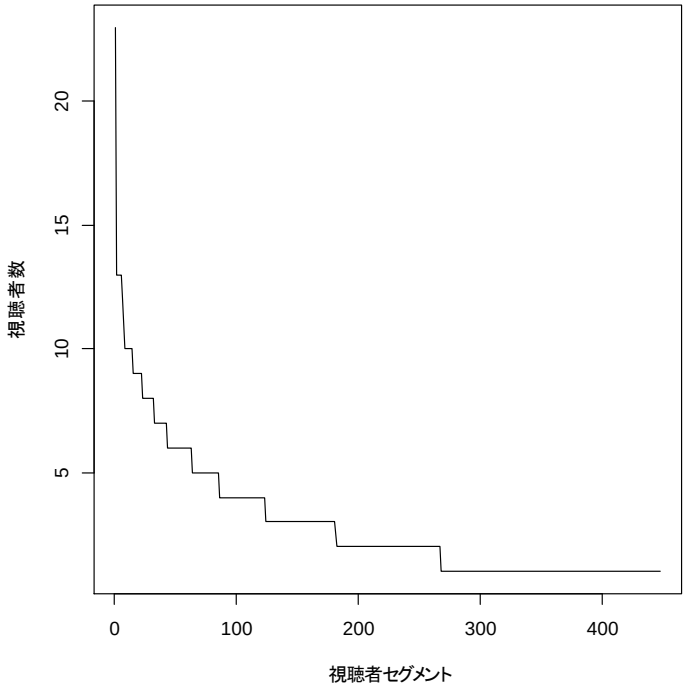


図 10　視聴者セグメント別の番組セグメント視聴確率の例