サービス科学における大規模データと統計的モデリング

東北大学大学院　経済学研究科　石垣　司

2012年12月7日　日本機械学会計算力学部門「設計情報学研究会」in 強羅

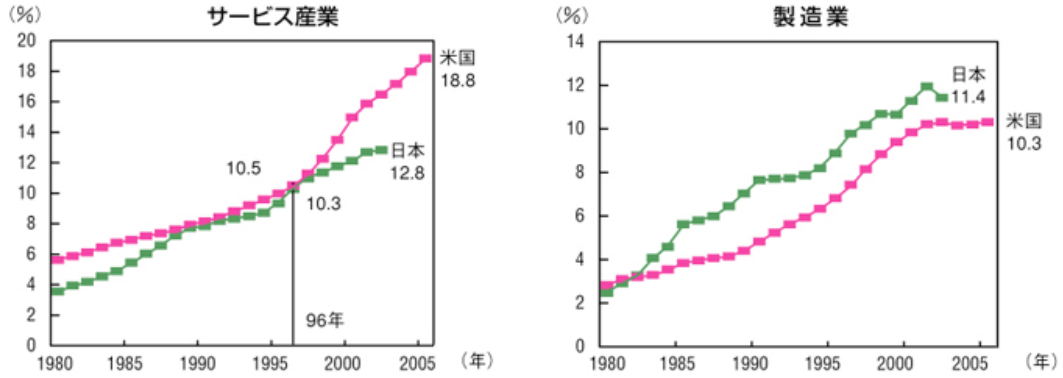
日本のサービス産業を取り巻く状況

- 日本のサービス産業
  - 実質GDPの7割弱
  - 雇用の3分の2
- しかしながら，サービス産業の生産性の成長率
  - 日本の製造業と比べ低い
  - 海外のサービス産業と比べて低い

労働生産性上昇率（1995～2003）

	米国	英国	ドイツ	日本
製造業	3.3%	2.0%	1.7%	4.1%
サービス業	2.3%	1.3%	0.9%	0.8%

OECD（経済協力開発機構）
compendium of productivity Indicator 2005 より



（備考）1．日本：IT資本ストック比率＝IT資本ストック（1995年価格実質ベース）／総資本ストック（1995年価格実質ベース）。
2．米国：IT資本ストック比率＝IT資本ストック（2000年価格実質ベース）／総資本ストック（2000年価格実質ベース）。
（資料）独立行政法人経済産業研究所「JIPデータベース2006」、米国商務省経済分析局Webサイトから作成。

日米におけるIT資本ストックの総資本ストックに占める比率

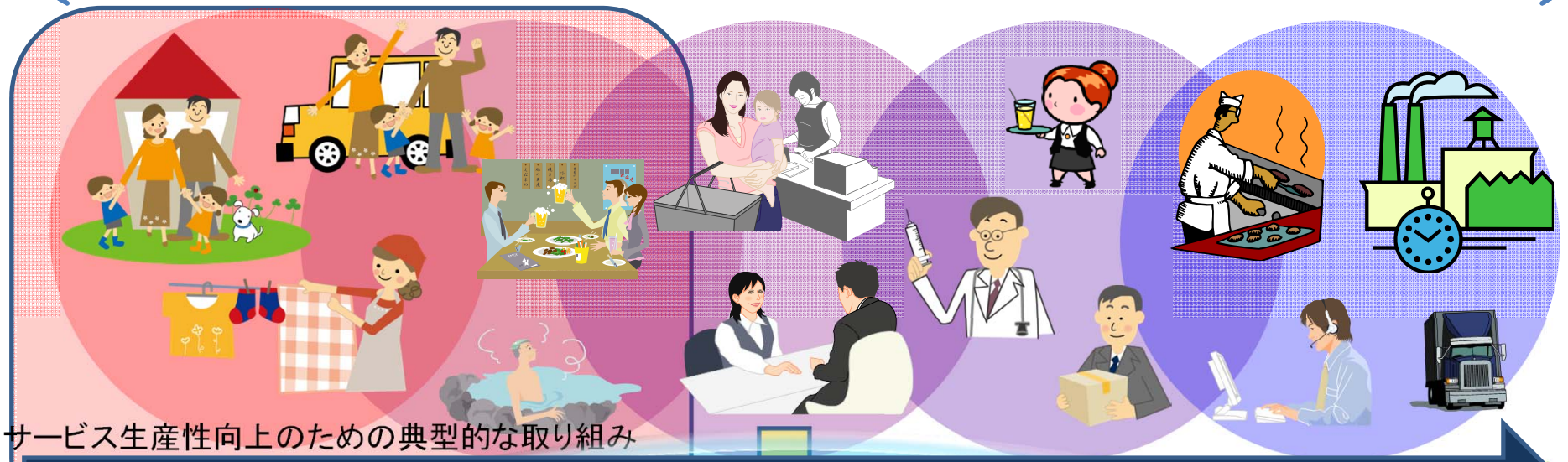
我が国サービス産業の労働生産性上昇率低迷の要因は、IT資本蓄積の不足とTFP上昇率の低迷にあることが分かった。
我が国サービス産業においては、TFP上昇とIT資本蓄積は金融仲介業、通信業及び卸売業で限定的に見られるにすぎず、
今後、幅広いサービス業種においてこれらに取り組むことによって労働生産性を高める必要がある。

サービス研究を推進する産学官の動向

- 1993年：IBM　サービス・サイエンス研究部門設立
- 2002年4月：東京大学人工物工学研究センター　サービス工学研究部門設立
- 2004年12月：米パルミザーノ・レポート（サービス・イノベーション）
- 2005年9月：中国「第11次5ヶ年計画」
- 2005年10月：北陸先端科学技術大学院大学MOT　サービスサイエンス
- 2006年7月：日本の財政・経済一体改革会議　「経済成長戦略」策定
- 2006年7月：韓国「社会サービス向上企画団」発足
- 2006年：経産省サービス工学検討チーム発足
- 2006年10月：東大サービス・イノベーション研究会
- 2007年5月：サービス生産性協議会発足
- 2007年4月：経産省「サービス産業生産性向上支援調査事業」
- 2007年9月：文科省「サービス・イノベーション人材育成推進プログラム」
- 2008年4月：産総研「サービス工学研究センター」設立
- 2010年4月：JST RISTEX「問題解決型サービス科学研究開発プログラム」
- 2010年4月：近畿大学次世代基盤技術研究所サービス工学研究センター設立
- 2012年1月：統計数理研究所サービス科学研究センター設立

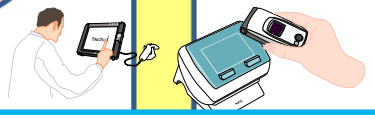
個人的な興味
大規模データに基づいた生活者起点のサービス科学

日常生活　利用行動　顧客接点　フロントヤード　バックヤード



サービス生産性向上のための典型的な取り組み

高効率化と高付加価値化の同時実現には生活者起点のサービス設計が必要



POS、ライフログデータ
アンケート・・・

観測

顧客接点で発生する大規模データ

適用

設計

分析

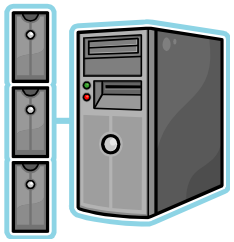
生活者理解・生活者行動モデル

大規模データと生活者の行動モデルに基づいた顧客接点の最適化

本日の話題

サービス業の
顧客接点

購買行動に関する
大規模データの蓄積



しかし現状は･･･、

データは未活用
社会的損失状態

大規模データの活用

企業生産性向上
顧客満足度向上

【本日の話題～小売サービスを例に】

大規模データ活用による「顧客理解の深化と個別最適化技術」

1. 百貨店の購買履歴データ
2. 流通量販店の購買履歴＋顧客アンケートデータの融合
3. 流通量販店の購買履歴＋確率効用モデル

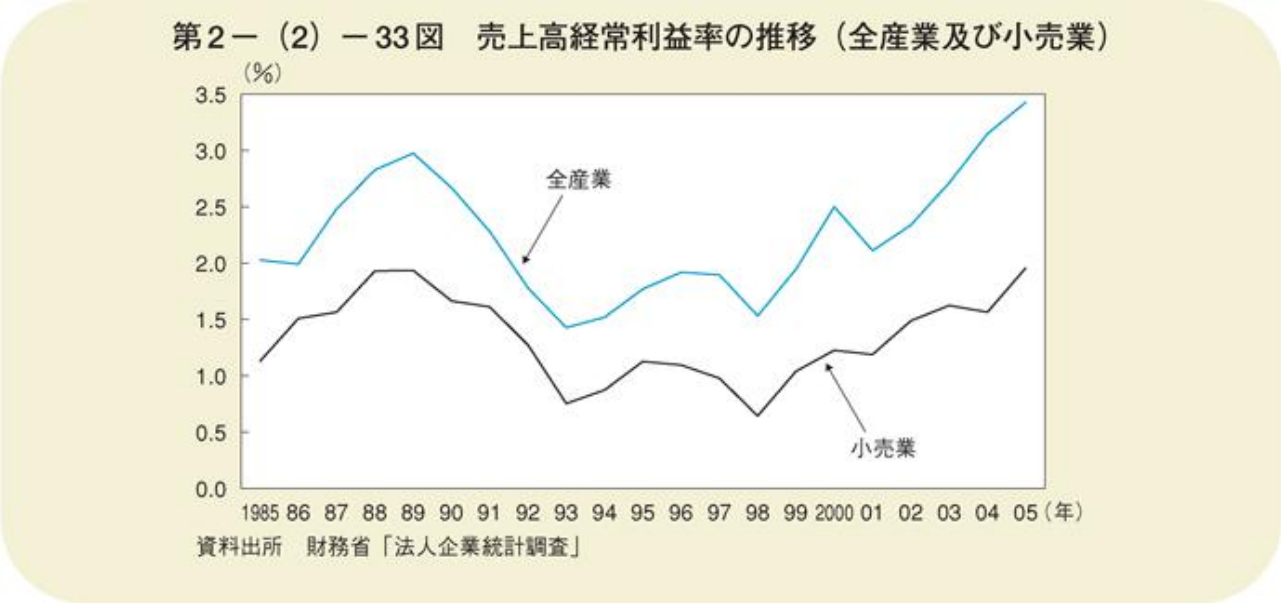
本日の話題

【小売サービスの現状】

- オーバーストア ⇒ 低価格化競争 ⇒ 利益率の低下
- 提供可能なサービス品質の低下へ ⇒ 生活者の享受価値減少へ
- 非低価格化戦略の重要性 ⇒ CRMや顧客満足度向上のための施策

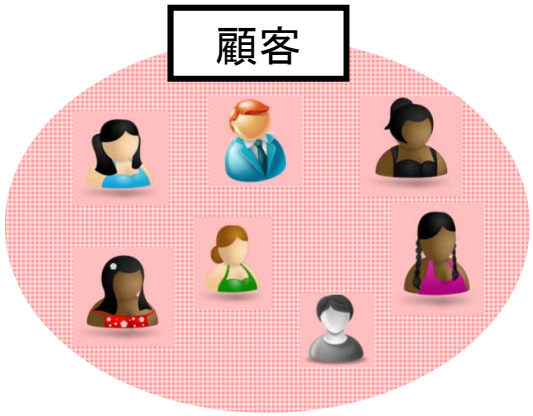
【小売サービスのビッグデータ（ID-POSデータ）】

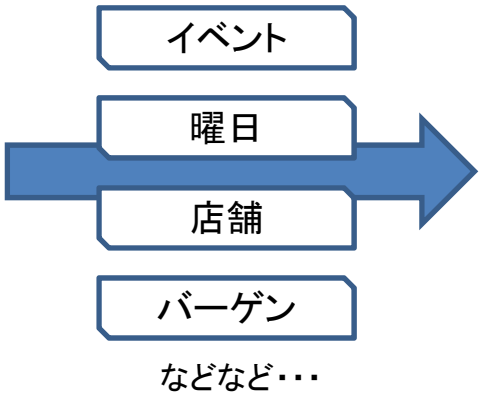
- 「いつ」「どこで」「だれが」「なにを」「なんこ」購買したか
- データ数が膨大（チェーン店では1年間で数十億件規模）



小売サービス～百貨店ID-POSを例に

【顧客の要求とサービスレベルのマッチングには】
・大量生産大量消費時代は終焉へ
・さらに十人十色から一人十色の時代へ





・来店行動
・選択行動
・購買行動

様々な顧客が、　様々な状況で、　様々な商品を購入。

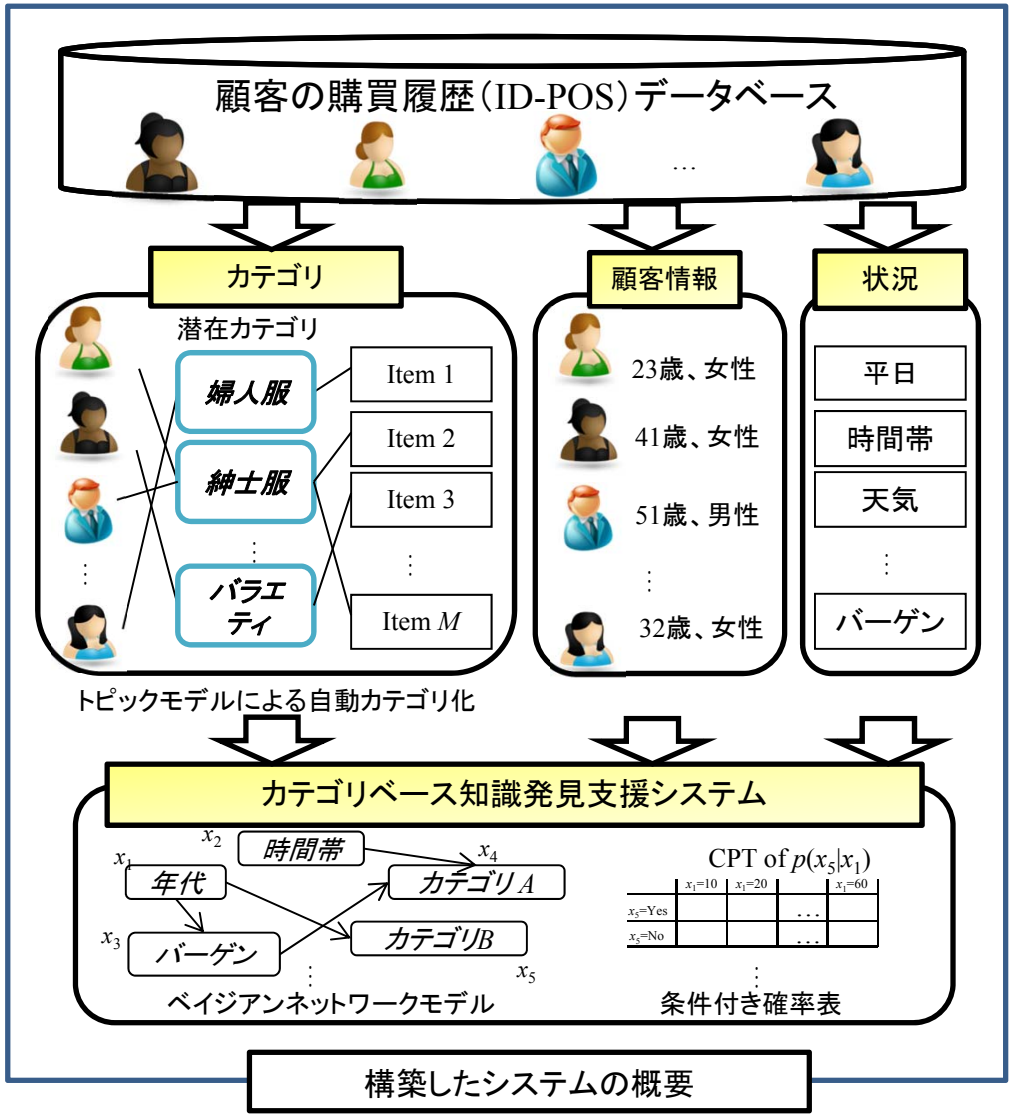
顧客の多様性を保持しながら購買行動の状況依存性を知りたい

小売サービス～百貨店ID-POSを例に
知識発見支援システム

【購買履歴データからの知識発見支援システム】
・ID-POSデータから、
－顧客・商品カテゴリを推定
－顧客情報のタグ付け
－購買状況のタグ付け

・ベイジアンネットワークによる
購買行動の計算モデル化

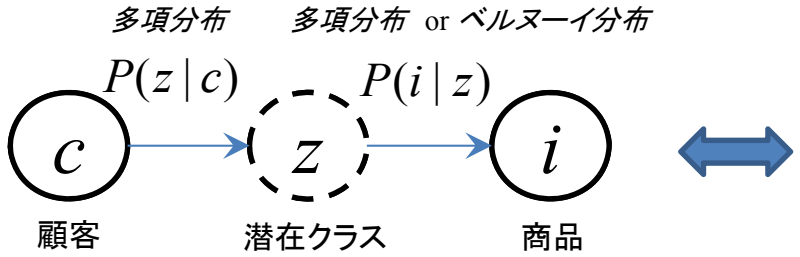
【使用データ】
・百貨店のID-POS
・3店舗の1年間
・約3,000,000件の履歴
・顧客ID:約20,000名分
・商品ID:126基本アイテム



トピックモデル（潜在クラスモデル）

【次元圧縮】
- 主成分分析、因子分析、独立成分分析
- データクラスタリング
- トピックモデル、潜在クラスモデル　⇒　離散確率値

【トピックモデル・潜在クラスモデル】
- 顧客セグメンテーションに利用、（1970年代のマーケティング研究）
- PLSI, LDAなどデータマイニング、機械学習の分野で研究が盛ん
- 圧縮基準　⇒　局所独立性



顧客cを潜在クラスへセグメンテーション

$$P(i|c) = \sum_z P(i|z)P(z|c)$$

“顧客数C、商品数I　>>　潜在クラス数K”

（cとiはzにより局所独立）

トピックモデル（潜在クラスモデル）

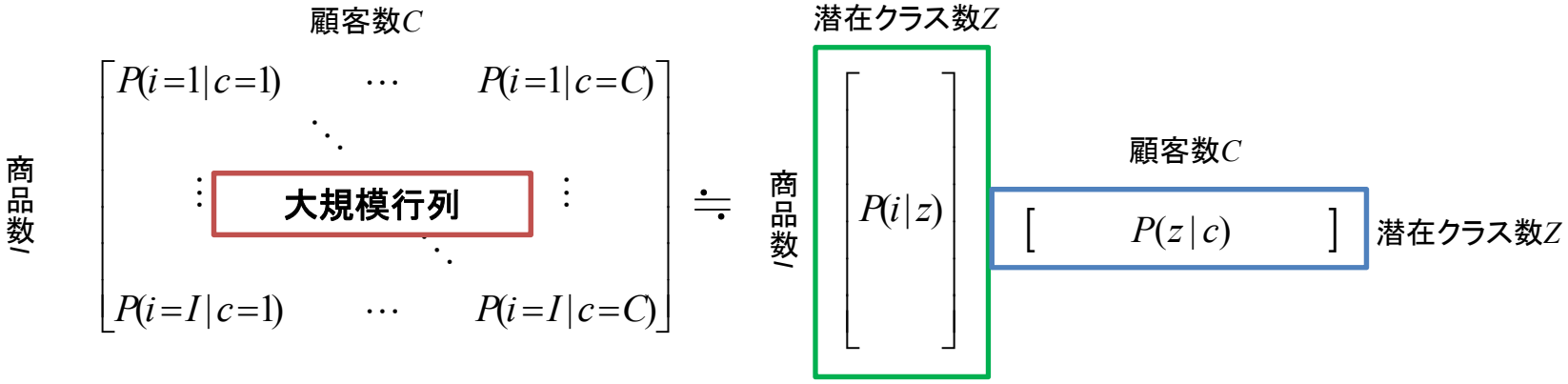
【マーケティングにおける潜在クラスの解釈】

- 顧客セグメント
  - 購買パターンでのセグメンテーション（同じような商品を買う人のセグメント）
  - 代表的な消費者として要約
- 商品カテゴリ
  - 上記の顧客セグメントが良く買う傾向にある商品の集合
  - 顧客購買パターンに基づいた商品カテゴリ（商品特性カテゴリではない）

【直感的なイメージ】

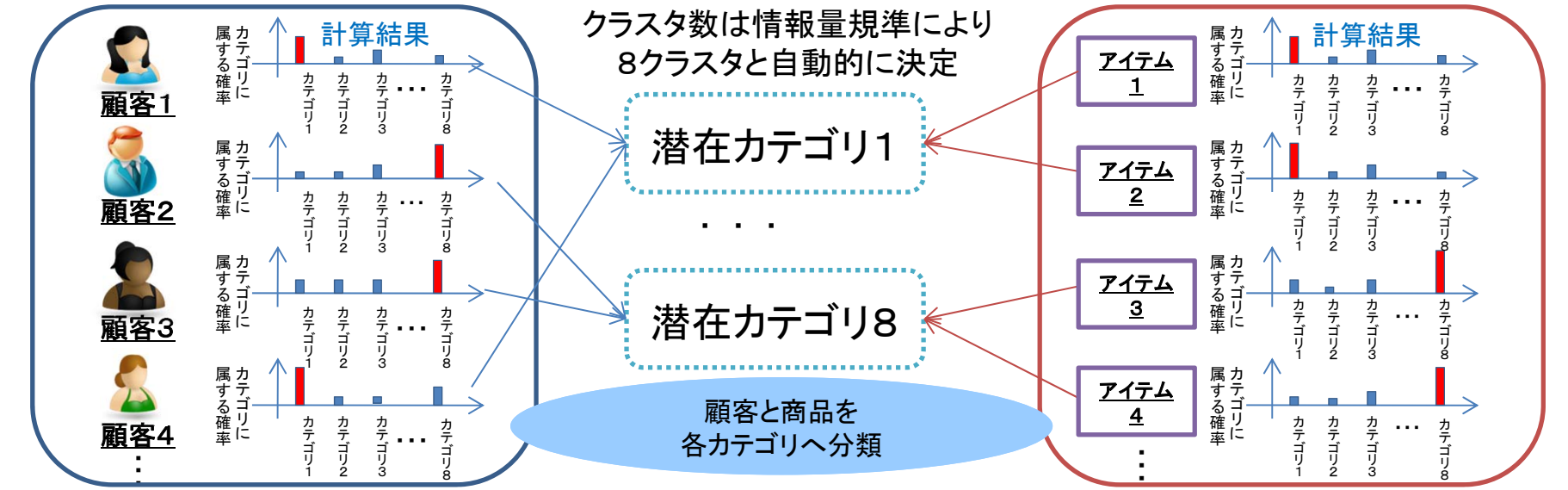
- 顧客・商品に関する多項分布パラメータの大規模行列
  ⇒ 商品・潜在クラス × 潜在クラス・顧客（2つの小行列に分解）

$$P(i|c) \cong \sum_z P(i|z)P(z|c)$$



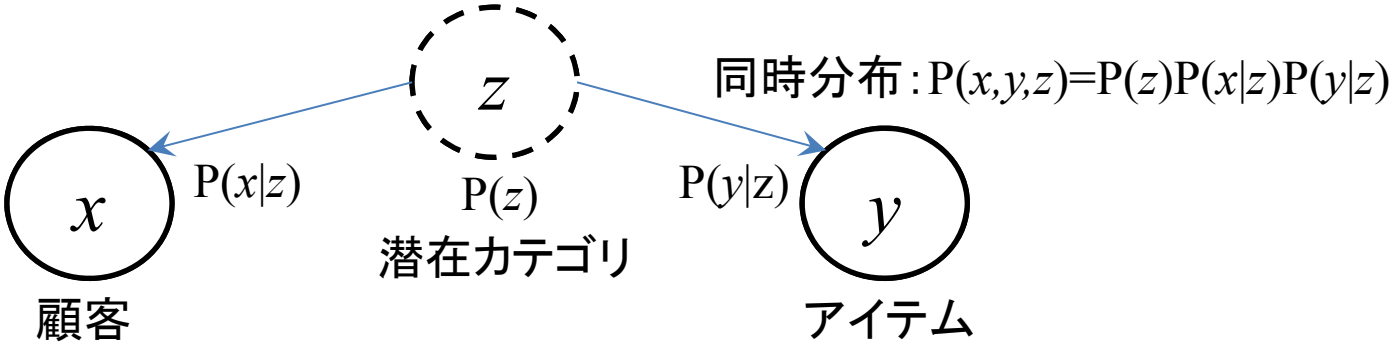
小売サービス～百貨店ID-POSを例に トピックモデル

・顧客ーアイテム双方向の局所独立性を仮定した潜在クラス分析
・情報量規準AICに基づきクラスタ数は8と決定



推定結果：顧客がある意味カテゴリに属する確率

推定結果：商品がある意味カテゴリに属する確率

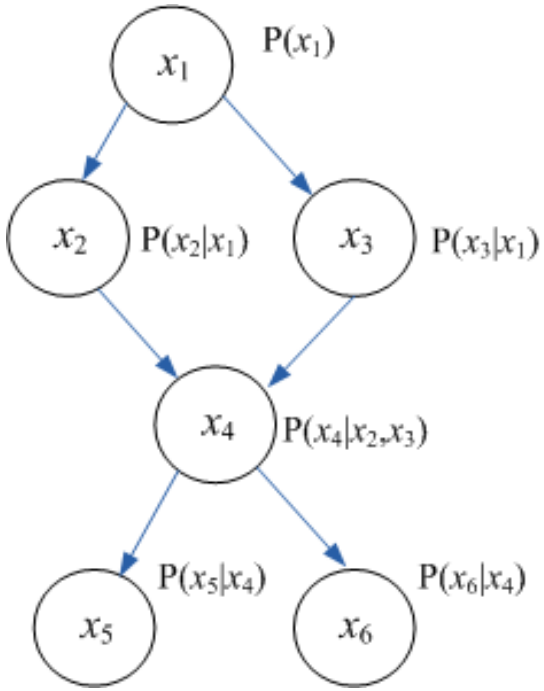


同時分布：$P(x,y,z)=P(z)P(x|z)P(y|z)$

小売サービス～百貨店ID-POSを例に ベイジアンネットワーク構築のための変数

・潜在カテゴリと状況変数の確率モデル化

ベイジアンネットワーク



上図で表現される確率モデルの同時分布

$$P(x_1,x_2,x_3,x_4,x_5,x_6)=P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2,x_3)P(x_5|x_4)P(x_6|x_4)$$

$P(a|b)$：事象bが生じる条件の下でのaが生じる条件付き確率

$P(x_a|x_b)$の条件付き確率表の例

x_a \ x_b	$x_b=b_1$	$x_b=b_2$	・・・	$x_b=b_M$
$x_a=a_1$	0.36	0.13	・・・	0.26
$x_a=a_2$	0.01	0.42	・・・	0.03
⋮	⋮	⋮	⋱	⋮
$x_a=a_N$	0.25	0.23	・・・	0.17

小売サービス～百貨店ID-POSを例に
ベイジアンネットワーク構築のための変数

使用データ：百貨店における1年間（3店舗）のID-POSデータ

顧客の属性

①年齢　：10代、20代、30代、40代、50代、60代以上の6分類
②性別　：男性、女性の2分類
③住所　：各都道府県の47分類
④購買個数　：合計購買個数に対し上位20%＝A、上位20～50%＝B、下位50%＝C
⑤購買金額　：合計購買金額に対し上位20%＝A、上位20～50%＝B、下位50%＝C
⑥来店頻度　：来店頻度（日毎）に対し上位20%＝A、上位20～50%＝B、下位50%＝C

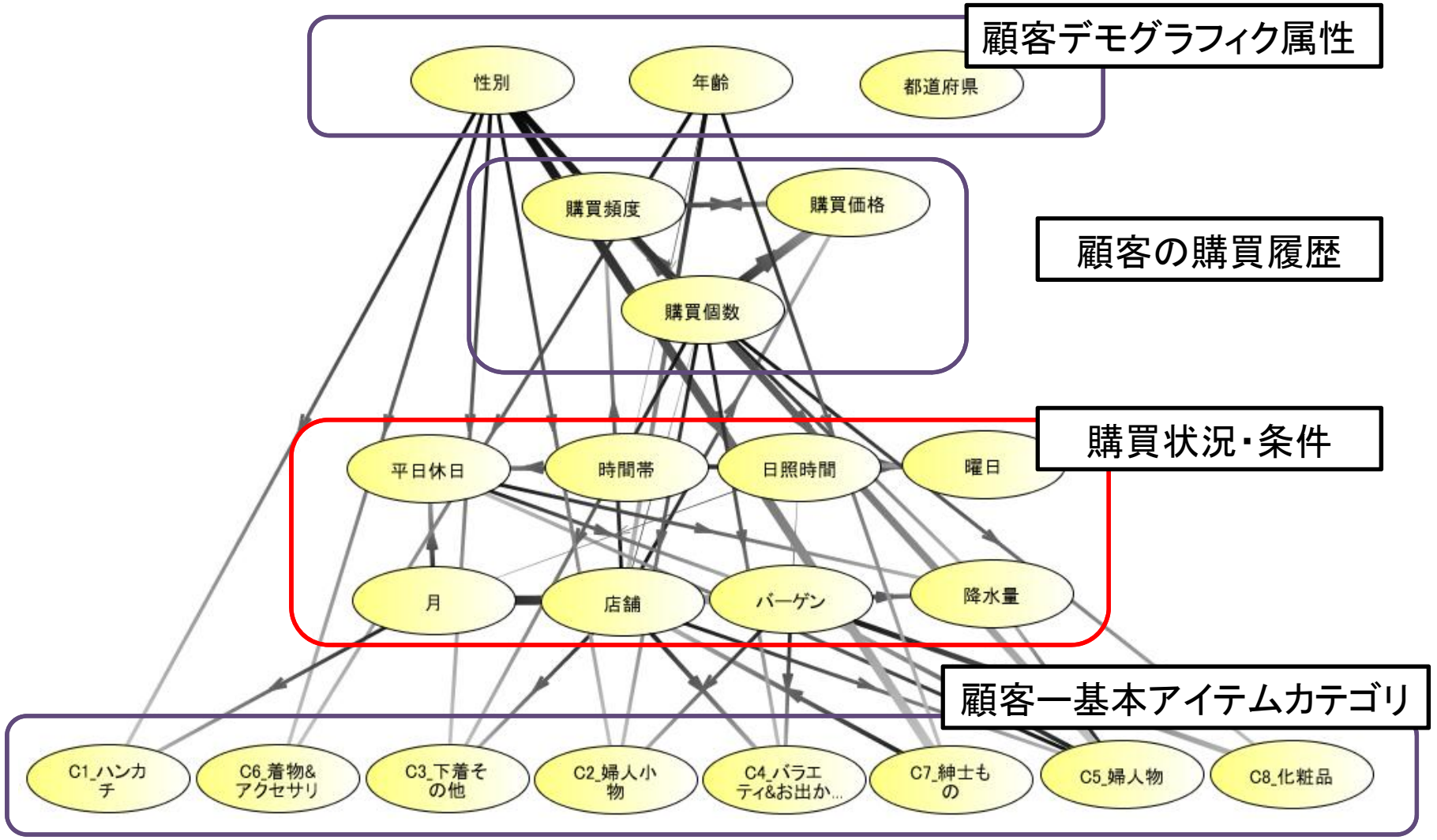
購買状況と条件（レシートデータ情報）

⑦月間 ：2008年4月～2009年3月までの1か月毎の12分類
⑧曜日 ：月曜から日曜までの7分類
⑨時間帯　：午前、昼、夕方、夜のどの時間帯で購買されたかの4分類
⑩平日or休日　：平日に購買されたか、それとも休日に購買されたかの2分類
⑪降水量 ：無し・極少量・少量・雨・大雨までの5分類
⑫日照時間 ：無・小・中・多までの4分類
⑬店　：どの店での購買されたかの3分類
⑭バーゲン：その商品はバーゲン品か否かの2分類

顧客ー基本アイテムカテゴリ

⑮潜在カテゴリ：購買された商品が属するクラスタに対してラベルを付与（2分類×8クラスタ）

小売サービス～百貨店ID-POSを例に
知識発見支援システム



小売サービス～百貨店ID-POSを例に
知識発見支援システム

各種結果

結果の詳細は、
「石垣司、竹中毅、本村陽一、百貨店ID付きPOSデータからのカテゴリ別状況依存的変数間関係の自動抽出法、オペレーションズ・リサーチ, Vol. 56, No. 2, pp. 77-83, 2011」
をご参照ください。

本研究で使用したデータは、平成21年度データ解析コンペティションより提供を受けました。具体事例に関してはWeb上での公開許可を得ておりませんので、上記の既出論文の結果をご参照ください。

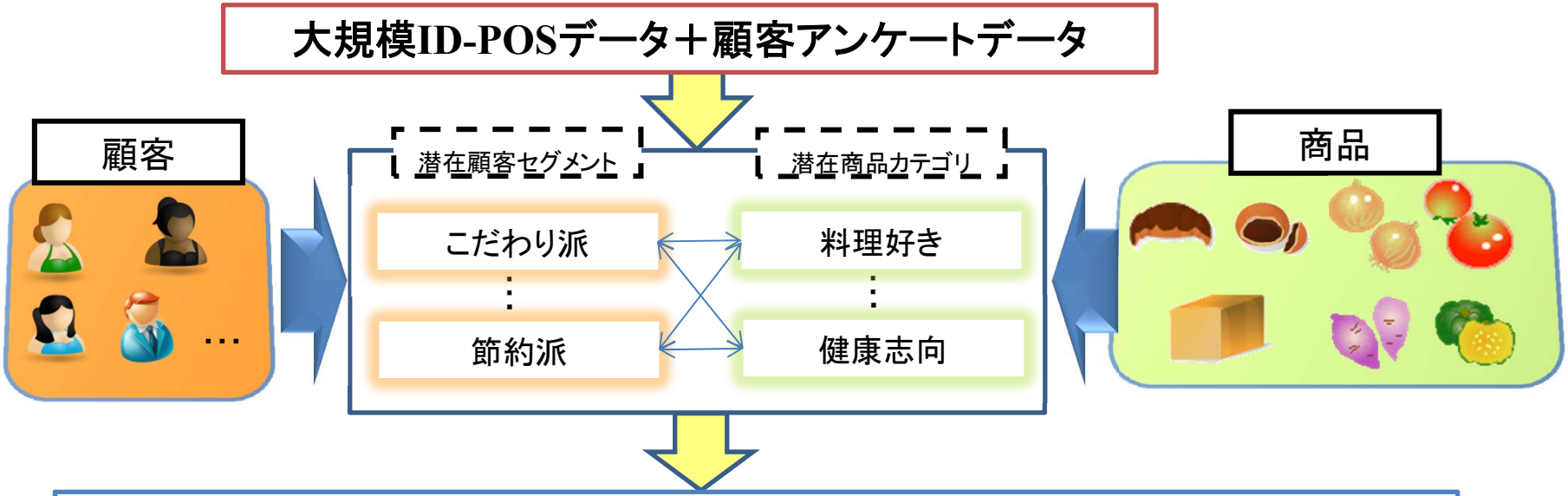
・時間帯に合わせて棚割や販促商品を変えることで、
年代別のセグメントに効果的な販促ができるのでは？

小売サービス～流通量販店における大規模データ融合

大規模購買関連データの融合

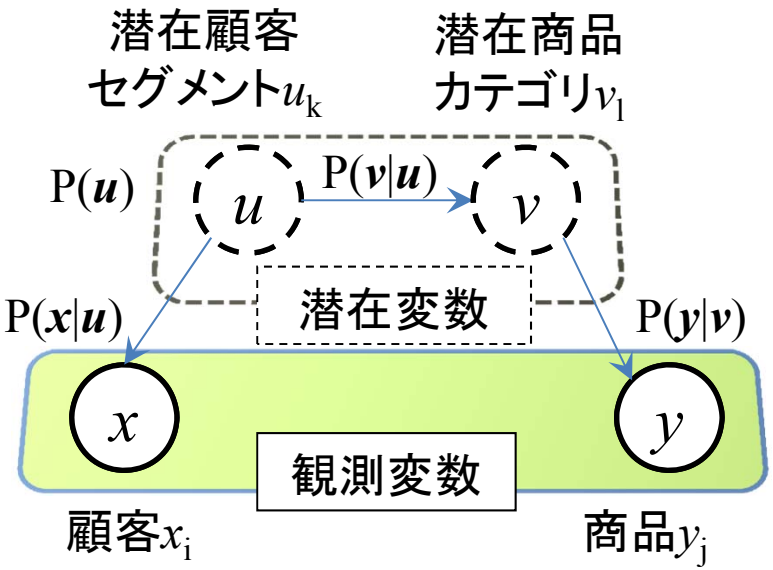
・小売サービスの現場では自動的かつエビデンスベースの
顧客や商品のセグメンテーション技術が求められている
・従来では年齢・性別属性等による顧客分類、流通業者の都合による商品分類が主
・生活者視点での顧客と商品のカテゴリ分類が必要

顧客と商品の生活者視点カテゴリを購買行動に関する大規模データからマイニング



小売サービス～流通量販店における大規模データ融合
提案モデルと使用データ

【提案モデル：多層潜在クラスモデル】



【使用するデータ】
神戸を中心に約150店舗を展開する
流通量販店コープこうべにおける
①ID-POSデータ
・期間：2008年10月1日～2009年9月30日
・トランザクション数：約6.7億件（669,511,467件）
②顧客アンケートデータ
・ライフスタイルに関する質問項目
・約4000人（3965人）の回答を得た

アンケートに回答した約4000人の
約420万件のレシートデータを利用

対数尤度：
$$L = \sum_{i}^{X}\sum_{j}^{Y} N_{ij} \log\left\{\sum_{k}^{U}\sum_{l}^{V} P(u_k)P(x_i \mid u_k)P(v_l \mid u_k)P(y_j \mid v_l)\right\}$$

【提案モデルの仮定】
1：各顧客はライフスタイルに基づいたセグメントに分類できる
2：各商品はカテゴリに分類できる
3：各顧客セグメントは特定の商品カテゴリを購買する傾向がある

小売サービス～流通量販店における大規模データ融合
顧客ライフスタイルアンケート

ライフスタイルアンケート結果の分析による6つの消費・生活因子の抽出

回転後の因子行列a

	因子 1	2	3	4	5	6
Q16バランスの良い食事	0.2336805	0.5484452	0.0408709	0.0097327	0.1557524	0.0594485
Q17毎日の生活が充実	0.0855746	0.6026855	0.1764881	-0.0392687	0.0331666	0.0867818
Q18料理好き	0.1962424	0.5076026	0.211716	0.0818154	0.0677555	-0.0316921
Q19お弁当	-0.0331312	0.1235969	0.130132	0.1064101	-0.0367461	0.1442449
Q20低カロリー	0.3140799	0.1955978	0.0373392	0.097738	0.0732867	0.0579539
Q21家計簿	0.0979278	-0.0170043	0.0381553	0.1084744	0.382077	0.0663099
Q22安ければ遠くても	0.0126347	-0.0280407	0.1280185	0.5886937	0.0560822	0.0859867
Q23チラシお得	0.1101945	0.0037253	0.0488645	0.7052306	0.0795292	0.0516476
Q24スーパー早くすませたい	0.0565589	-0.0340723	-0.0053603	0.0636841	0.0214505	0.4885447
Q25献立はスーパーで	0.0558405	-0.0295026	0.1124582	0.0424565	-0.2193077	0.1272927
Q26無駄遣い	0.0780097	-0.2121282	0.2667936	-0.0149716	-0.421225	0.0515053
Q27新商品	0.2424196	-0.0596534	0.4309869	0.0860724	-0.1797376	0.0462965
Q28コープでしか買わない商品	0.3818731	-0.0139799	0.0518595	0.0712535	0.0438735	0.0683186
Q29高くても健康	0.7100942	0.0974914	0.1466559	-0.1320333	0.0088925	0.0021707
Q30産地レシピ関心	0.5397713	0.1537658	0.230683	0.0721101	0.0494689	-0.0635669
Q31にぎやかな所が	0.0622048	0.0035026	0.5018811	0.0665499	-0.0247638	-0.0184657
Q32きちょうめん	0.2001072	0.0943064	0.103723	0.0465794	0.3645636	0.0498431
Q33気分が変わりやすい	0.0396625	-0.2796193	0.0775764	0.0435094	-0.0142338	0.0936903
Q34新しい体験	0.122036	0.1159401	0.555494	-0.0180691	0.0288857	0.0735427
Q35友人と買い物	0.0668866	0.1074661	0.3374588	0.0791719	0.0204875	0.0056362

因子抽出法: 主因子法
回転法: Kaiserの正規化を伴うﾊﾞﾘﾏｯｸｽ法

因子分析から，特長のある6つの因子が抽出できた．
(因子の妥当性を確認)

⇒これらの因子の組み合わせとして消費者の分類を行う

第１因子：こだわり消費派：高くても健康に良いものを選び，産地への関心，こだわりのブランドがある
第２因子：家庭生活充実派：料理が好きで食事も生活も充実している．気分も安定している
第３因子：アクティブ消費派：外向的で，新商品や話題の商品は試しに買ってみる．ただ無駄遣いは多い
第４因子：節約消費派：チラシを見てお得な商品を買う．安ければ少々遠い店にも行く．高い商品は買わない
第５因子：堅実生活派：几帳面で家計簿をつけ，無駄遣いはしない．毎日の献立はスーパーに行く前に決める
第６因子：パパっと消費派：スーパーでの買い物はできるだけ早くすませたい．お弁当を作ることがある

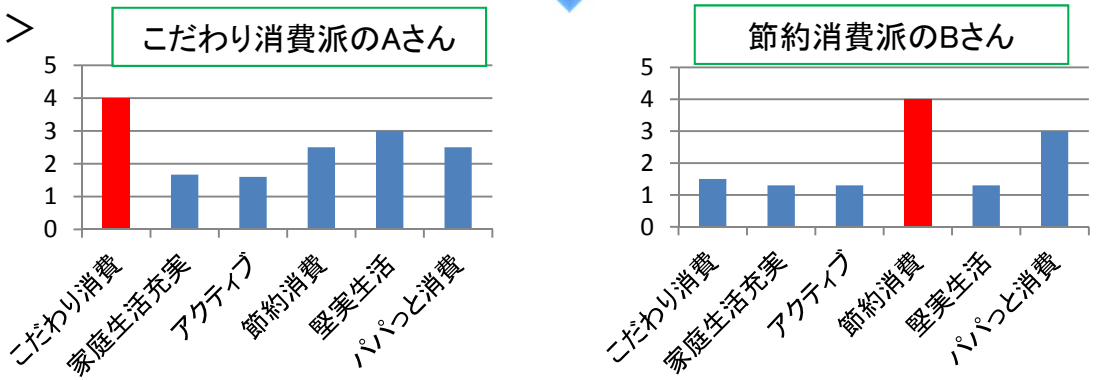
小売サービス～流通量販店における大規模データ融合
顧客ライフスタイルアンケート

アンケートの結果から各ライフスタイルに属する得点が計算可能

第１因子：	**こだわり消費派**	（1670人、43%）
第２因子：	**家庭生活充実派**	（1385人、34%）
第３因子：	**アクティブ消費派**	（384人、10%）
第４因子：	**節約消費派**	（707人、18%）
第５因子：	**堅実生活派**	（364人、9%）
第６因子：	**パパっと消費派**	（869人、22%）

各顧客に対してライフスタイルに関するプロファイリングが可能に

＜例＞

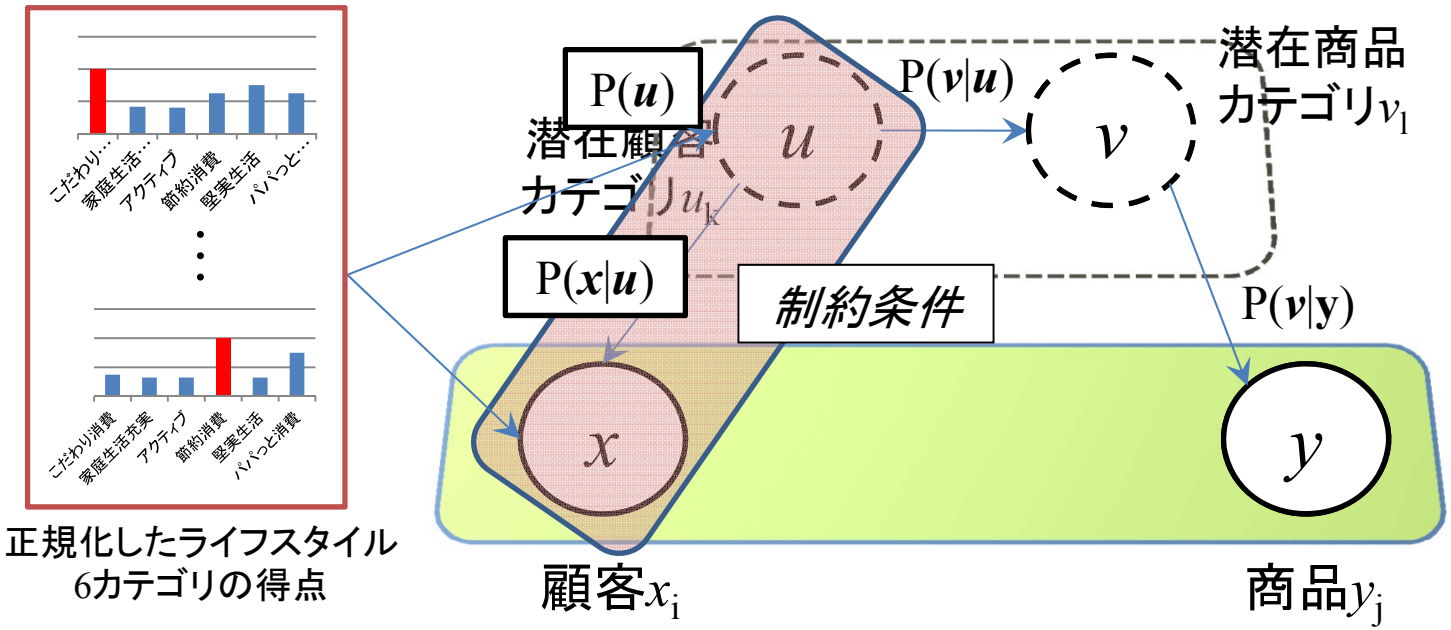


各因子に属するアンケート項目の4段階回答を合計し、項目数で正規化した得点
最大4点、最小1点

小売サービス～流通量販店における大規模データ融合

【制約条件の導入】

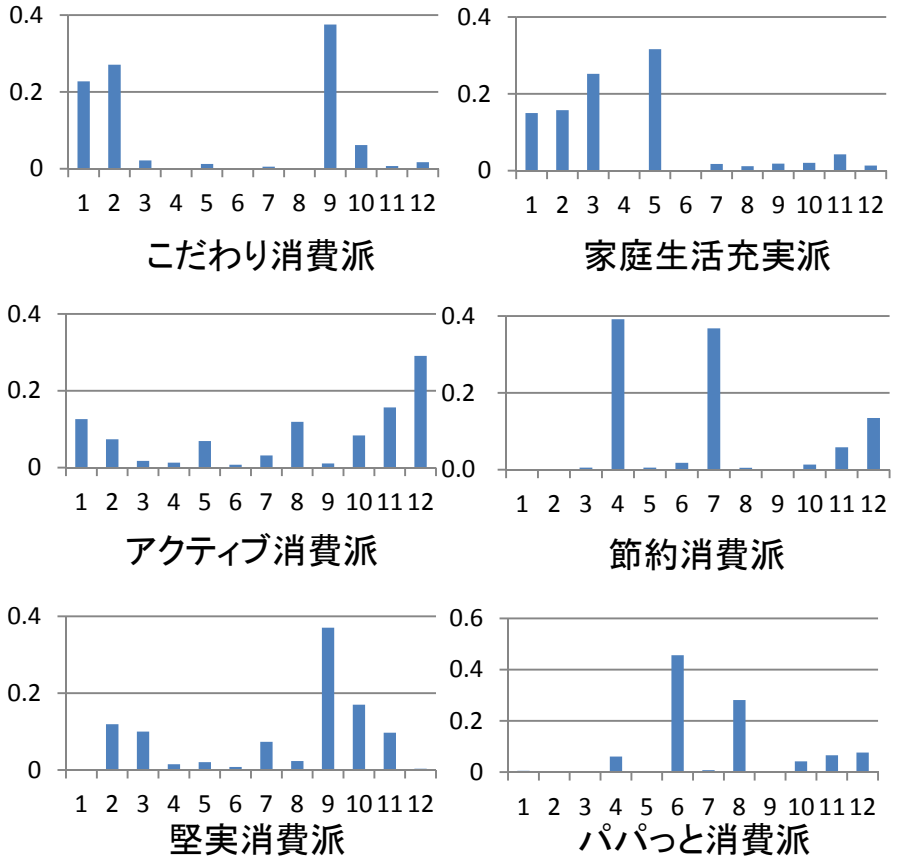
・正規化したアンケート得点を制約条件としてパラメータを固定
・潜在顧客セグメント数U → ライフスタイルカテゴリ数6
・潜在商品カテゴリ数V → 情報量規準に基づき12カテゴリ



顧客ライフスタイルを多層潜在クラスモデルへ反映し、最尤推定

小売サービス～流通量販店における大規模データ融合 自動分類の結果

【パラメータ$p(v|u)$（潜在顧客セグメント→潜在商品カテゴリ） の推定結果】



12商品カテゴリの特徴的な商品

No.	特徴商品	No.	特徴商品
1	高品質PB	7	セール頻出品
2	生野菜・生鮮	8	多種混合
3	日配品	9	高価格帯野菜
4	低価格帯商品	10	小サイズ野菜・日配
5	鮮魚・肉類	11	飲料
6	肉・パン・飲料	12	惣菜・飲料

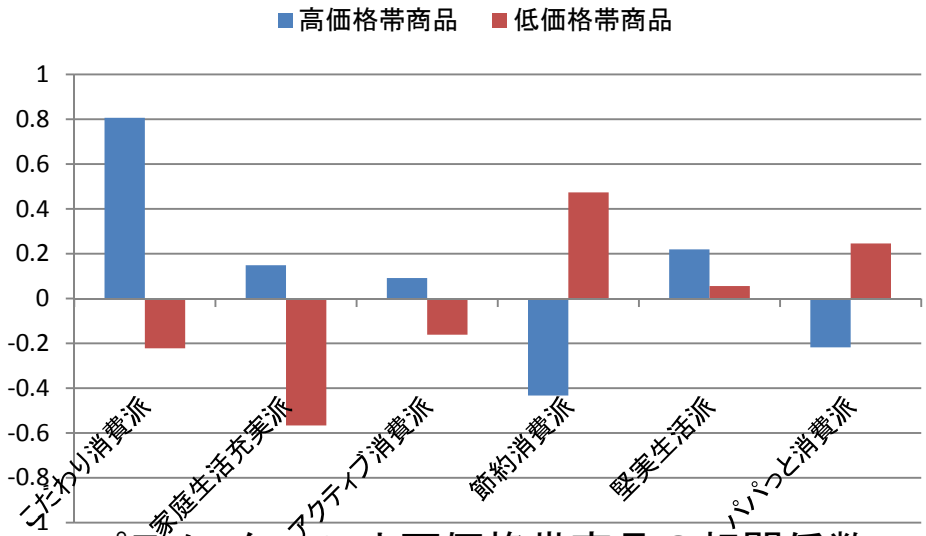
【分類結果の一例】
- 1000商品中、見切り品が2商品あり、その両方がカテゴリ4へ
- 19種ある10個詰めたまごのうち平均単価が高い5商品がカテゴリ1,2,9へ
- また、最も安い商品はカテゴリ7へ
- カテゴリ1,2,3,5には調理済みの総菜などがほとんど分類されていない

小売サービス～流通量販店における大規模データ融合 自動分類の結果

【商品の価格帯でみる分類の妥当性】

・牛乳、たまご、PB、見切り品から
　高価格帯商品（71商品）、低価格帯商品（20商品）を選別

・その商品カテゴリ毎の所属確率を求め、推定した$p(v|u)$との相関をみる

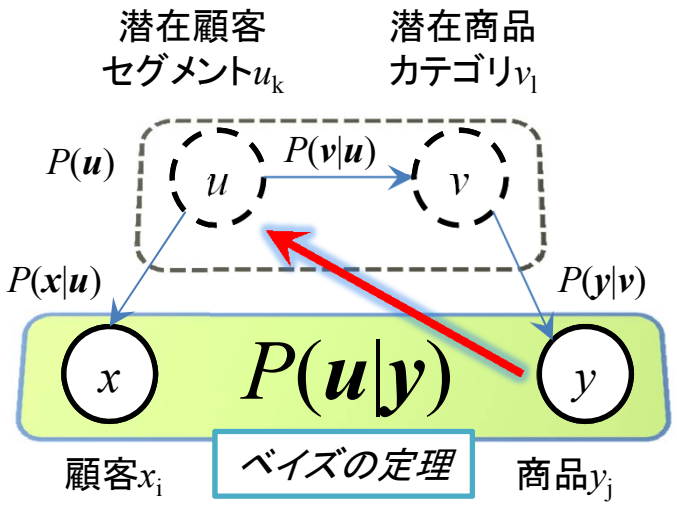


パラメータ$p(v|u)$と両価格帯商品の相関係数

第１因子：**こだわり消費派**：高くても健康に良いものを選び，産地への関心，こだわりのブランドがある
第２因子：**家庭生活充実派**：料理が好きで食事も生活も充実している．気分も安定している
第３因子：**アクティブ消費派**：外向的で，新商品や話題の商品は試しに買ってみる．ただ無駄遣いは多い
第４因子：**節約消費派**：チラシを見てお得な商品を買う．安ければ少々遠い店にも行く．高い商品は買わない
第５因子：**堅実生活派**：几帳面で家計簿をつけ，無駄遣いはしない．毎日の献立はスーパーに行く前に決める
第６因子：**パパっと消費派**：スーパーでの買い物はできるだけ早くすませたい．お弁当を作ることがある

小売サービス～流通量販店における大規模データ融合
応用1：商品DNAの自動付与

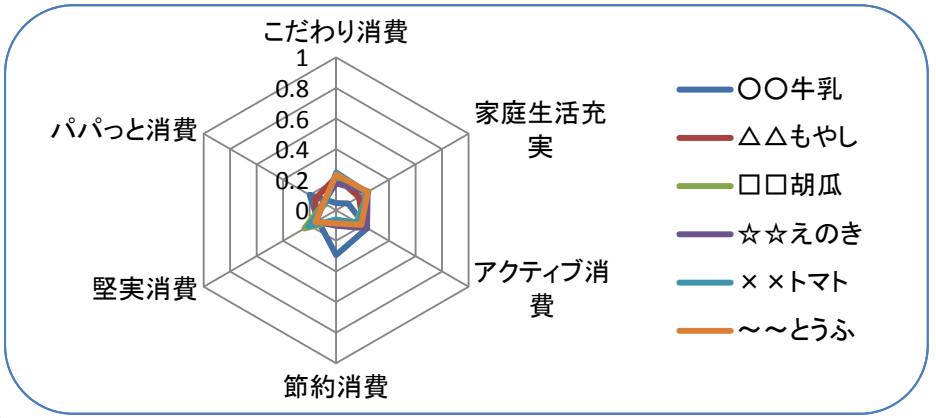
【商品毎のライフスタイルプロファイル（商品DNA）の付与】
・パラメータの推定結果から個々の商品と潜在顧客セグメント（ライフスタイルカテゴリ）の関係を計算可能



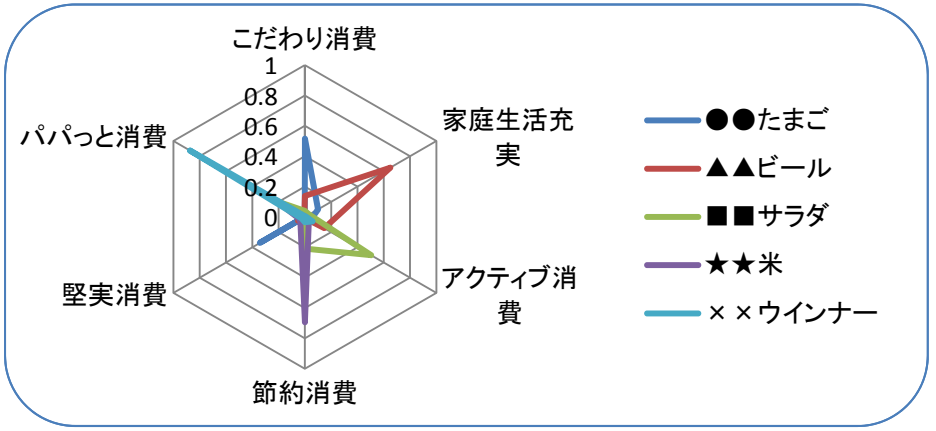
$P(\text{ライフスタイル}|\text{商品})$

各商品が与えられた時の
ライフスタイルの確率が計算可能に

～各商品のライフスタイルプロファイル～



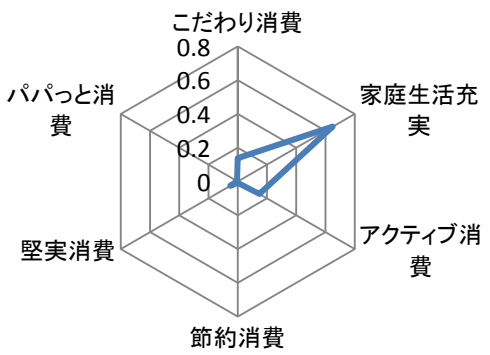
売上個数上位に位置する6商品



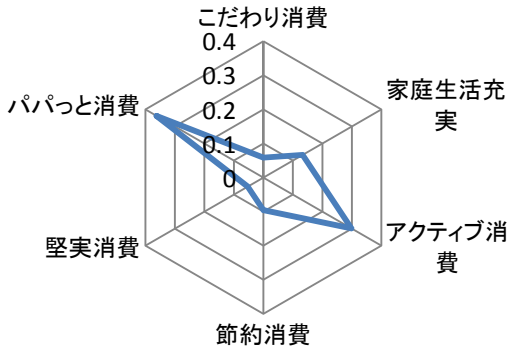
一つの軸に大きな値を持つ特徴的な5商品

小売サービス～流通量販店における大規模データ融合
応用1：商品DNAの自動付与

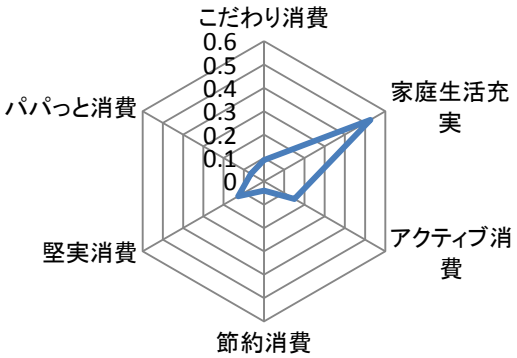
【商品毎のライフスタイルプロファイル：アルコール飲料の一例】



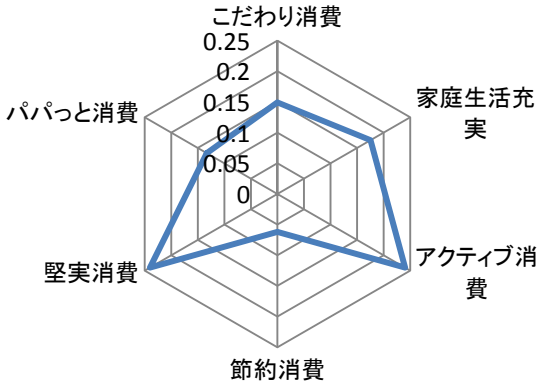
X社：ビール、350ml×6缶セット



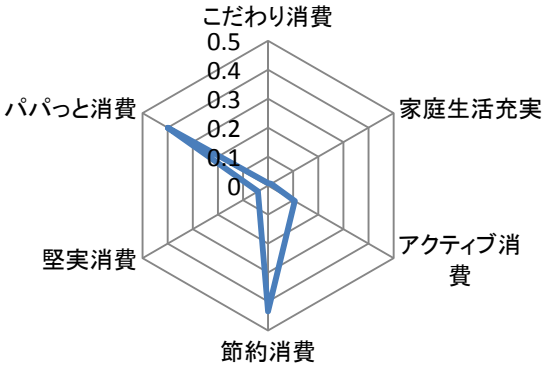
X社：第3のビール、350ml×6缶セット



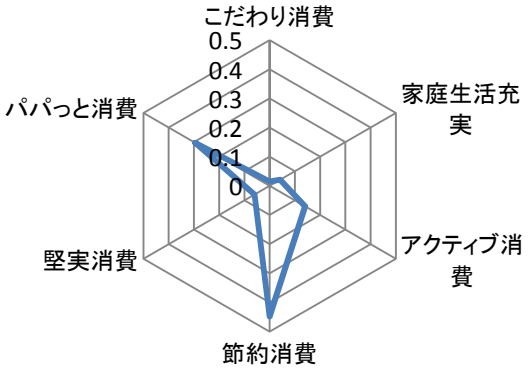
Y社：第3のビール、350ml×6缶セット



X社：ビール、500ml×1缶



X社：第3のビール、500ml×1缶



Y社：第3のビール、500ml×1缶

同系統の商品でも商品特性により得点傾向には差がある

少しアカデミックな話題へ

【大規模データと消費者行動理論に基づくモデル】

・マーケティング、消費者行動研究は膨大な実績

・しかし、理論と実践の遊離が問題に

・実活用可能な消費者行動モデルが重要に

- **マーケティング研究における確率的選択モデル（演繹型モデル）**
  ⇒　大規模データへの対応が困難
- **大規模データ処理技術による帰納的次元圧縮**
  ⇒　個別顧客の特性をモデル化が困難

ベイズモデリングによる両者の統合

- 大規模な顧客・商品数での効用が計算可能
- マーケティング変数を介した購買シミュレーションが可能

効用原理に基づく確率的選択モデル

【マーケティングにおける顧客行動モデル】

- 離散選択モデル（D. McFadden: 2000年ノーベル経済学賞）
- 効用の概念　⇒　効用最適化原理
- “効用＝確定的部分＋確率的部分”　⇒　確率的選択

$$\text{効用}：U_i = \underset{\text{確定的部分}}{V_i} + \underset{\text{確率的部分}}{\varepsilon_i}$$

（確定的部分）　マーケティングによって説明可能な部分

$$V_i = \beta_0 + \beta_1 \text{価格} + \beta_2 \text{販促} + \cdots$$

（確率的部分）　心理的な揺れなど、マーケティングでは説明できない部分

効用が最大となる商品を選択

問題設定とゴール

問題設定

顧客　$c = \{1, \cdots, C\}$

商品　$i = \{1, \cdots, I\}$

時刻　$t = \{1, \cdots, T\}$

データ　$y_{cit} = \{0,1\}$

時刻 t における商品 i のマーケティング変数（M要素）：

$\mathbf{x}_{it} = [x_{it1}, \cdots, x_{itm}, \cdots x_{itM}]^T$、

顧客 c の商品 i に対する
マーケティング変数に対する反応係数ベクトル：

$\boldsymbol{\alpha}_{ci} = [\alpha_{ci1}, \cdots, \alpha_{cim}, \cdots, \alpha_{ciM}]^T$

時刻 t における顧客cの商品 i に対する効用（潜在変数）を考える。

効用　$u_{cit} = f(\mathbf{x}_{it}) + \varepsilon_{cit} = \mathbf{x}_{it}^T \boldsymbol{\alpha}_{ci} + \varepsilon_{cit}$

ゴール

店舗オペレーションに必要な顧客・商品数に関する大規模データの下で、

$\boldsymbol{\alpha}_{ci}$ を推定したい

問題設定とゴール

【購買行動の定式化と問題点】

- 効用 u_{cit} のモデル　⇒　誤差εが正規分布（2項プロビットモデル）
- 時刻 t における顧客cが商品i を購買するベルヌーイ確率:（＝買うか買わないか、1 or 0 ）

2項プロビットモデル

$$p(y_{cit}=1)=p(u_{cit}>0)=p(\mathbf{x}_{it}^{T}\boldsymbol{\alpha}_{ci}+\varepsilon_{cit}>0)=F(\mathbf{x}_{it}^{T}\boldsymbol{\alpha}_{ci})$$

F : 標準正規分布の累積分布関数

$\boldsymbol{\alpha}_{ci}$の推定でゴール達成。　⇒　しかし・・・、

反応係数$\boldsymbol{\alpha}_{ci}$の数＝顧客数C×商品数I

データがスパース（全体では大規模データでも、個人毎、商品毎のデータは無 or 小）

⇒

潜在クラスによる次元圧縮

提案モデル（潜在クラス型階層ベイズ2項プロビットモデル）

【プロビットモデルと潜在クラスモデルの融合】

1. 購買確率（効用>0）の分解

$$p(u_{cit} > 0) \cong \sum_{z=1}^{Z} p(u_{cit} > 0 \mid z) C_{cz} \qquad C_{cz} \equiv p(z \mid c)$$

2. 潜在クラスがもつ効用

潜在クラス z の商品 i への反応係数ベクトル：$\boldsymbol{\beta}_{zi} = [\beta_{zi1}, \cdots, \beta_{zim}, \cdots, \beta_{ziM}]^T$

$$p(u_{cit} > 0 \mid z) = p(\mathbf{x}_{it}^T \boldsymbol{\beta}_{zi} + \varepsilon_{cit} > 0 \mid z) = F(\mathbf{x}_{it}^T \boldsymbol{\beta}_{zi} \mid z)$$

3. 潜在クラス型2項プロビットモデル

$$p(u_{cit} > 0) \cong \sum_{z=1}^{Z} F(\mathbf{x}_{it}^T \boldsymbol{\beta}_{zi} \mid z) C_{cz}$$

尤度：$$l = \prod_{c=1}^{C} \prod_{i \in i_c} \prod_{t \in t_c} \sum_{z=1}^{Z} \left[\left\{ F(\mathbf{x}_{it}^T \boldsymbol{\beta}_{zi} \mid z_c) C_{cz} \right\}^{y_{cit}} \left\{ 1 - F(\mathbf{x}_{it}^T \boldsymbol{\beta}_{zi} \mid z_c) C_{cz} \right\}^{1-y_{cit}} \right]$$

提案モデル（潜在クラス型階層ベイズ2項プロビットモデル）

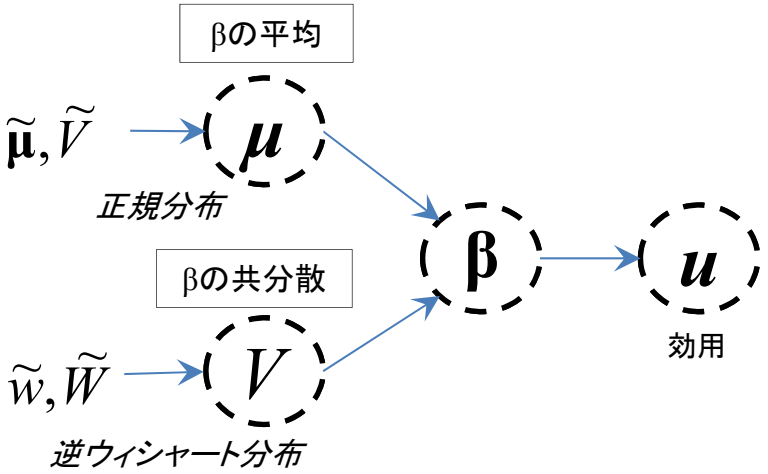
【階層ベイズモデル化】

4. 潜在クラスの異質性モデル

– βの階層ベイズモデル化

$$
\begin{aligned}
u_{ict} \mid z &\sim Normal(\mathbf{x}_{it}^{T}\boldsymbol{\beta}_{iz}, 1) \\
\boldsymbol{\beta}_{iz} &\sim Normal(\boldsymbol{\mu}_i, V_i) \\
\boldsymbol{\mu}_i &\sim Normal(\tilde{\boldsymbol{\mu}}, \tilde{V}) \\
V_i &\sim InverseWishart(\tilde{w}, \tilde{W})
\end{aligned}
$$

$\tilde{\boldsymbol{\mu}}, \tilde{V}, \tilde{w}, \tilde{W}$: Hyper parameters



階層ベイズ2項プロビットモデル

5. 潜在クラスのベイズモデル化

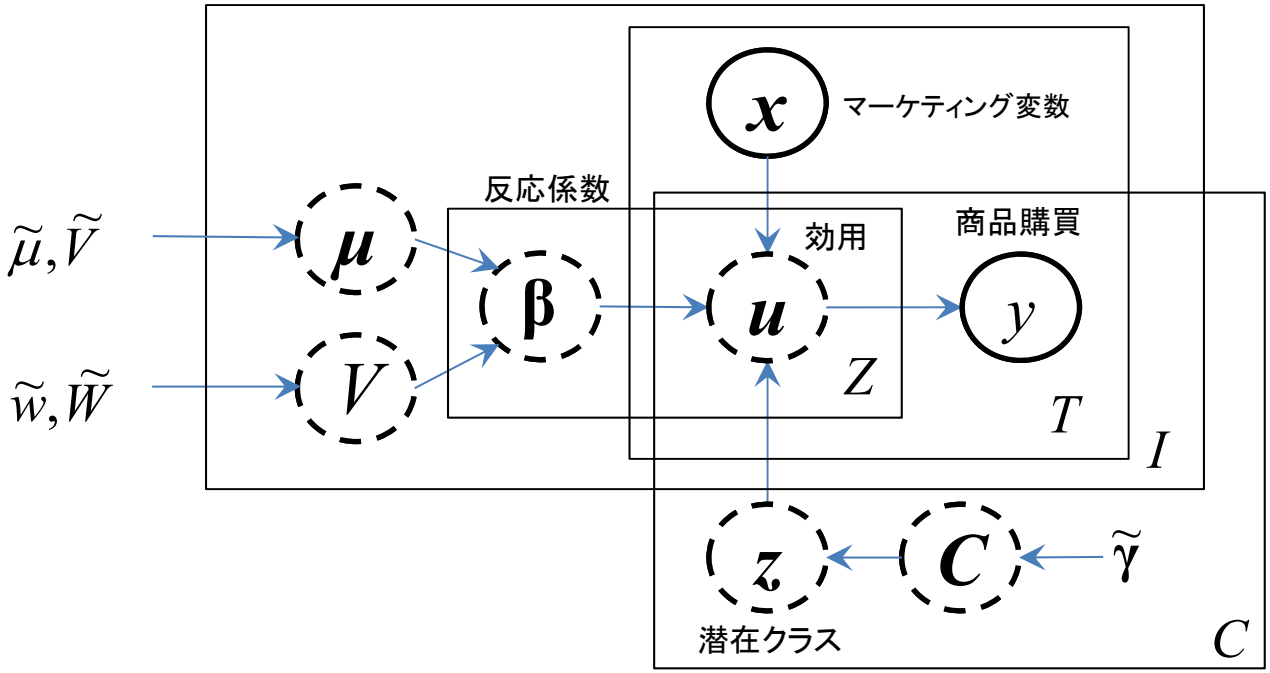
$$
\begin{aligned}
C_{cz} &\sim Dirichlet(\tilde{\gamma}) \\
z &\sim Multinomial(C_{cz})
\end{aligned}
$$

$\tilde{\gamma}$: Hyper parameter

Diriclhet – 多項分布モデル

提案モデル（潜在クラス型階層ベイズ2項プロビットモデル）

【潜在クラス型階層ベイズ2項プロビットモデル】



- 階層ベイズ2項プロビットモデルと潜在クラスモデルの統合モデル
- 演繹モデルと帰納モデルの融合

MCMCによるパラメータ推定

【MCMCによるサンプリング】

提案モデルの完全同時分布

$$P(\{\mathbf{C}_c\},\{z_c\},\{\boldsymbol{\beta}_{zi}\},\{u_{cit}\},\{\boldsymbol{\mu}_i\},\{\mathbf{V}_i\},\mathbf{X},\mathbf{Y};\tilde{\boldsymbol{\gamma}},\tilde{\boldsymbol{\mu}},\tilde{V},\tilde{w},\tilde{W},)$$

$$=\prod_{c=1}^{C} Diriclet(\mathbf{C}_c;\tilde{\boldsymbol{\gamma}})Multinomial(z_c\,|\,\mathbf{C}_c)$$

$$\cdot\prod_{i\in i_c}^{I} Normal(\boldsymbol{\mu}_i\,|\,V_i;\tilde{\mu},\tilde{V})InversWishart(V_i;\tilde{w},\tilde{W})$$

$$\cdot\prod_{z=1}^{Z} Normal(\boldsymbol{\beta}_{zi}\,|\,\boldsymbol{\mu}_i,V_i)\prod_{t\in t_c}^{T} p\left(u_{cit}>0\,|\,z,\mathbf{x}_{it}\right)^{y_{ict}}\,p\left(u_{cit}\le 0\,|\,z,\mathbf{x}_{it}\right)^{1-y_{ict}}$$

Gibbs sampler

Draw 1 : $\mathbf{C}_c$（ディリクレ分布）

Draw 2 : z_c（多項分布）

Draw 3 : u_{cit}（切断正規分布）

Draw 4 : $\boldsymbol{\beta}_{zi}$（多変量正規分布）

Draw 5 : $\boldsymbol{\mu}_i$（多変量正規分布）

Draw 6 : V_i（逆ウィシャート分布）

個人の係数 α_{ic} の推定

個人に対する
プロビットモデルより：

$$p(y_{cit}=1)=F(\mathbf{x}_{it}^{T}\boldsymbol{\alpha}_{ci})$$

潜在クラスモデルより：

$$p(y_{cit}=1)\cong\sum_{z}^{Z}p(u_{ict}>0\,|\,z)C_{cz}$$

潜在クラスに対する
プロビットモデルより：

$$p(u_{ict}>0\,|\,z)=F(\mathbf{x}_{it}^{T}\boldsymbol{\beta}_{zi})$$

↓

上の3式より：

$$\mathbf{x}_{it}^{T}\boldsymbol{\alpha}_{ci}\cong F^{-1}\left\{\sum_{z}^{Z}C_{cz}F(\mathbf{x}_{it}^{T}\boldsymbol{\beta}_{zi})\right\}\equiv\psi_{cit}$$

↓

線形回帰モデルへ帰着

$$\psi_{cit}=\mathbf{x}_{it}^{T}\boldsymbol{\alpha}_{ci}+\varepsilon_{cit}$$

↓

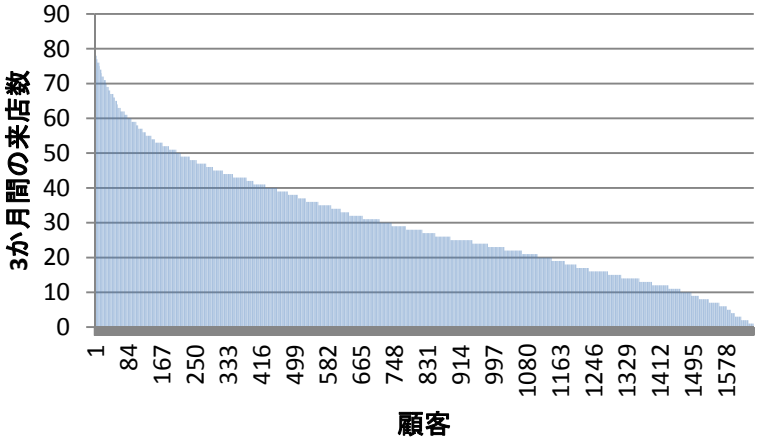
反応係数$\boldsymbol{\beta}_{zi}$から個人の$\boldsymbol{\alpha}_{ci}$の分布を算出

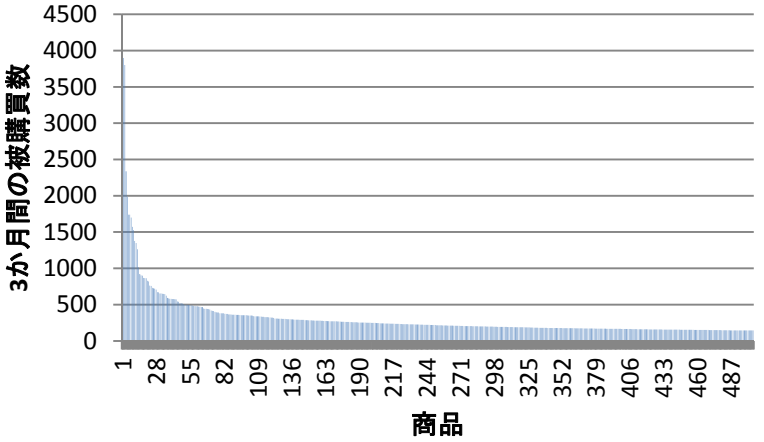
データとパラメータの推定

【推定に利用したデータ】

- 流通量販店1店舗
- 期間：2000年4月1日～6月30日までの3か月間
- 顧客数：1,647名パネルデータ
- 商品数505種類
  - I. 期間内にエンド陳列あり
  - II. 期間内にチラシ掲載あり
  - III. 被購買数が144以上
- 取引トランザクション数：159,494件
- マーケティング変数：
  1. 商品固有価値
  2. 価格（Price）（価格最大掛け率）
  3. エンド陳列の有無（Display）
  4. チラシ掲載の有無（Leaflet）

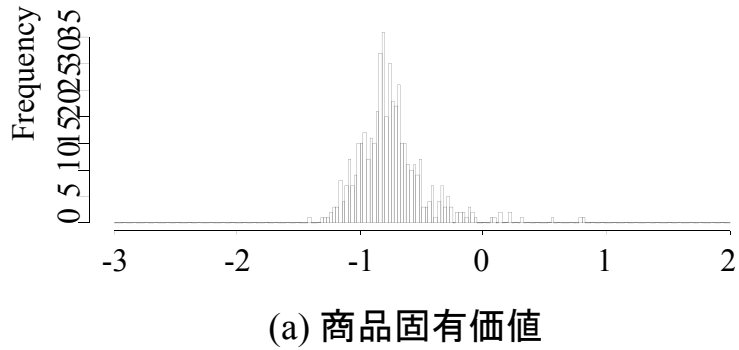
$$\mathbf{x}_{it} = [x_{it1}, \cdots, x_{itM}]^T = [1 \quad P_{it} \quad D_{it} \quad L_{it}]$$



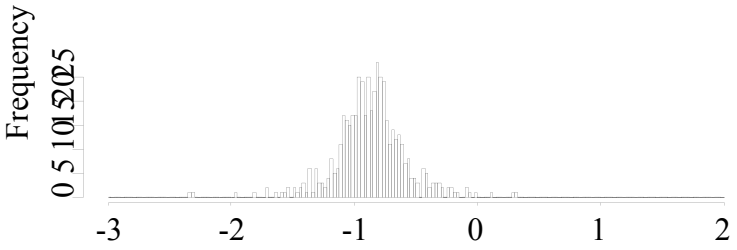


個別顧客・個別商品の反応係数（α_{ic}の事後平均）

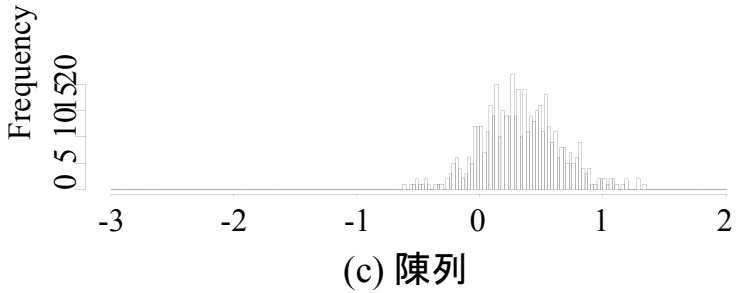
【商品に関するヒストグラム】



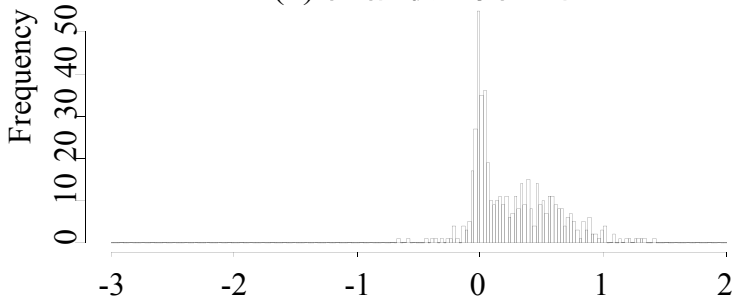
(a) 商品固有価値



(b) 価格最大掛け率



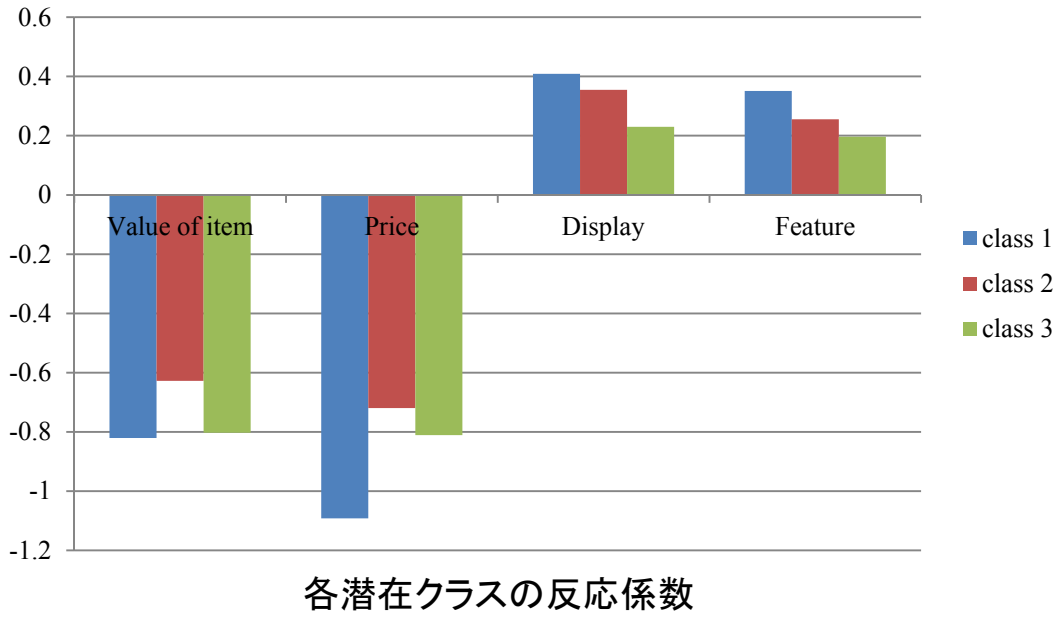
(c) 陳列



(d) チラシ

- 商品固有価値、価格は、ほとんどの商品が負の係数
- 陳列は、多くの商品が正の係数
- チラシは多くの商品が正の係数。ゼロに近い係数をもつ商品多数

潜在クラスの内容



各潜在クラスの反応係数

Latent class No.	1	2	3
The number of Customers	882	525	240

- クラス１：値引き、陳列、チラシに大きな反応をもつ
- クラス２：陳列、チラシのプロモーションには平均的な反応。値引きには反応しない。
- クラス３：値引きには反応するが、陳列、チラシのプロモーションには反応しない。

おわりに

【提案モデルのまとめ】

- サービスの研究の紹介
- 小売サービスでの研究事例紹介
  - ビッグデータ対応型の顧客理解深化
  - ビッグデータ対応型の個別最適化技術

【今後の課題】

- ビッグデータ活用の方法論
- 「設計」に関する議論