各データセットにおいて、PCAを適用し次元圧縮を行い2次元グラフに可視化した図およびC=3でRCCMM法を適用、ラフメンバシップ値で色分けした図を以下に示す。

結果から分析できる点：

1. Needs、二値化を行ったmovieLensは視覚的にはクラスター構造が確認できない
2. 二値化を行わないmovieLensはデータの散らばりが比較的大きいように見える

TODO: データの違いにおいても、考察を進める

* Needs-scan/panel

グラフ, 散布図

自動的に生成された説明

図1：PCAにより2次元圧縮しプロットしたneeds-scan/panel

【主成分の考察】

* 言えること

1. PC1はほとんどすべての製品の固有ベクトルが負なので，製品を所有すればするほどそのユーザーのPC1は小さくなる．（中・小電気冷蔵庫(206)：0.1231333 だけ例外）
2. グラフでいうと，多くの種類の製品を所有している（かつ中・小電気冷蔵庫を持っていない）ユーザーは左側，そうでないユーザーは右側に位置するといえる．
3. PC2は**ピアノ(340)，大型電気冷蔵庫(858)，電気乾燥機(242)**の固有ベクトルが大きくマイナス，**ワープロ(506)，VD(325)，自動二輪車(294)，中・小電気冷蔵庫(206)，オーブン(347)**が大きくプラス．あんまりはっきり言えるわけではないけど，所有世帯数が多い製品が負の影響を与えて，200とか300の特殊っぽい製品がプラスの影響を与えがち？
4. 平均的なユーザーが下側に行って，特殊めなユーザーが上に行ってる感じ

* 固有ベクトル

自動車(825)：-0.15049394，0.002329907940268947

ピアノ(340)：-0.36946269，**-0.1494865427313055**

VTR(933)：-0.11202948，0.0363792177384905

ルームエアコン(911)：-0.15799141，-0.06848782673657076

パソコン(588)：-0.49071821，-0.1325133187797386

ワープロ(506)：-0.25390201，**0.6376444273473912**

CD(844)：-0.28065201，0.00027368836929280793

VD(325)：-0.24121207，**0.34757293022230074**

自動二輪車(294)：-0.11300577，**0.1029684196181354**

自転車(893)：-0.10562772，-0.04107484336224411

大型電気冷蔵庫(858)：-0.17867588，**-0.3968588251278407**

中・小電気冷蔵庫(206)：0.1231333，**0.46004134640129246**

電子レンジ(962)：-0.00148637，-0.04698998838969434

オーブン(347)：-0.31861813，**0.11578749893968888**

コーヒーメーカー(617)：-0.36215479，-0.029411956654792164

電気洗濯機(986)：0.0070824，0.003750639410060031

衣料乾燥機(226)：-0.18521648，0.0037931785660436316

電気乾燥機(242)：-0.15189497，**-0.169541071062793**

* PC1，PC2の主成分得点のtop,bottom10の実際のデータベクトル

テーブル

低い精度で自動的に生成された説明

背景パターン, テーブル

自動的に生成された説明

グラフ, 散布図

自動的に生成された説明

図２：C=3でRSCCMM法を適用し、ラフメンバシップ値で色分けした結果

auc: 0.8515453684630208、delta: -6.0

（delta: [-5, -5.5, -6, -6.5, -7, -7.5, -8, -8.5] 各10回試行の中からaucが最大であった結果を採用）

* 二値化したmovieLens

グラフ, 散布図

自動的に生成された説明

図3: PCAにより2次元圧縮しプロットした二値化したmovieLens

【主成分考察】

1. PC1は主成分得点のトップ10とボトム10のそれぞれの要素を比較すると，

トップ10のユーザーの平均評価値は

**[0.31989708 0.38936535 0.40566038 0.39708405 0.39965695 0.42024014 0.41766724 0.38850772 0.42281304 0.43910806]**

ボトム10の平均評価値は

**[0.62521441 0.62607204 0.65265866 0.6380789 0.58404803 0.61835334 0.59605489 0.62521441 0.61921098 0.59948542]**

よって，ボトムの方が平均評価値が高い

1. 映画を平均的にいい評価をつける人が左側に，そうでない人が右側に行く
2. PC2は主成分得点のトップ10とボトム10のそれぞれの要素を比較すると，平均評価値はPC1ほど差はないのに対して，未評価値の数がボトム10の方がだいぶ多い
3. 多くの映画を評価したユーザーが上側に行き，評価した映画の数が少ないユーザーが下側に行く

グラフ, 散布図

自動的に生成された説明

図4: C=3でRSCCMM法を適用し、ラフメンバシップ値で色分けした結果

Auc: 0.7124551028552769, delta: -6.9

（delta: [-6.5, -6.7, -6.9, … -8.9]で 各10回試行の中からaucが最大であった結果を採用）

* 未評価値を各ユーザーの平均評価値で補完したmovieLens(※)

グラフ, 散布図

自動的に生成された説明

図5: PCAにより2次元圧縮しプロットした(※)

【主成分考察】

1. PC1はbinary と同じく，平均評価値が高いほど左に行く
2. PC2はbinaryと同じく，評価した映画数が多いほど上に行く

グラフ, 散布図

自動的に生成された説明

図4: C=3でRSCCMM法を適用し、ラフメンバシップ値で色分けした結果

auc: 0.7091151328700238, delta: -7.430

（delta: [-7.428, -7.429, -7.430, …, -7.432]で 各10回試行の中からaucが最大であった結果を採用）

検証方法について，確認ありがとうございます．

各データで主成分得点のトップ10，ボトム10や各固有ベクトルを調べて検証してみたところ，以下のようになりました．

エビデンスとなる実験結果の数値も含めて修士論文にまとめ，こちらに提出するつもりです．

ご確認よろしくお願いします．

Needs

Pc1: 中・小電気冷蔵庫(206)を除いて，ほぼすべての製品の固有ベクトルが負の値なので，多くの種類の製品を所有している（かつ中・小電気冷蔵庫を持っていない）ユーザーは左側，そうでないユーザーは右側に位置するといえる．

Pc2: 所有世帯数が多い製品が負の影響を与えて，少ない製品が正の影響を与えている．（逆の動きをする製品などもある）

Pc2からはっきりしたことは言えないが，平均的な特徴を持ったユーザーが下側，特殊な嗜好パターンをっ持つユーザーが上側に位置する傾向がある（？）

movieLens(二値化した and 二値化せず未評価値をユーザーの平均評価値で補完した)

Pc1: トップ10とボトム10の平均評価値に大きく差があり，平均的に良い評価をつけるユーザーがグラフの左側に，低い評価をつけるユーザーが右側に位置する．

Pc2: トップ10とボトム10の未評価値の数に大きく差があり，評価した映画が多いほどグラフの上側に，少ないほどグラフの下側に位置する．