

令和 6 年 6 月 24 日現在

機関番号：35302

研究種目：基盤研究(C)（一般）

研究期間：2021～2023

課題番号：21K11799

研究課題名（和文）複雑化データの情報抽出のための次元・変数縮約と同時推定及び計算環境の研究

研究課題名（英文）Dimension and variable selection, simultaneous estimation, and computational environment for information extraction from complex data

研究代表者

森 裕一（MORI, Yuichi）

岡山理科大学・経営学部・教授

研究者番号：80230085

交付決定額（研究期間全体）：（直接経費） 3,100,000円

研究成果の概要（和文）：本研究では、(i)情報損失を最小限に抑えた大規模さの軽減、(ii)縮約した情報や複雑性の関係の効率的な把握、(iii)尺度混在データやテキストデータの処理を可能とする手順の提供、(iv)計算の効率化を実現することを目的とし、尺度混在と高次元のデータを既存手法で扱える手法や手順を開発した。既存手法は、主成分分析、ファジィc平均法、テキストマイニングで、非計量主成分分析による数量化、数量化と次元縮約の同時推定、トピックモデルとヒートマップを利用し、複雑性と大規模性の軽減を実現した。また、効率的に結果を得るために、変数縮減と計算加速化を提案し、性能評価を行うことで、提案手法の有効性を確認した。

研究成果の学術的意義や社会的意義

数量化と次元縮約を同時に行うことで、個々の手法を別々に適用する方法では観察できなかった知見が得られること、このことにより、量的データのために開発されたすべての手法をカテゴリカルデータに適用できるようになることが本研究の学術的な特色である。また、アルゴリズム面で大規模性を克服しようとすることは、迅速に処理結果が得られる計算環境を提供するという点で独創性がある。これらの手法はマーケティングや社会心理学の個別課題を念頭に開発しているが、手法の良い点はお互いに利用し合うことで分野横断的な成果が得られる。さらに、提案手法がデータ分析の発見的情報把握の強力なツールとして活用されることも期待される。

研究成果の概要（英文）：In this study, we developed methods and procedures to deal with mixed measurement level and high-dimensional data in existing methods by (i) reduction of the scale size with minimum information loss, (ii) efficient analysis of the reduced information and complexity, (iii) propose of a procedure that enables processing of mixed measurement level data and text data, and (iv) efficient computation. The existing methods we used are principal component analysis, fuzzy c-means, and text mining. Quantification by non-metric principal component analysis, simultaneous estimation of quantification and dimension reduction, topic model and heat map are used to reduce the size and complexity. In order to obtain the results efficiently, variable reduction and computational acceleration are also proposed. The effectiveness of the proposed methods/procedures are confirmed by the performance evaluation.

研究分野：計算機統計学

キーワード：数量化 カテゴリカルデータ クラスタリング 変数選択 加速化 主成分分析

## 1. 研究開始当初の背景

意図せず集められたビッグデータには、冗長な情報やノイズなど、無駄な情報が混在する。計画的に収集した調査やテストのデータにしても、収集手段の高機能化により、可能な限り多くの情報を得ようとして、大規模化、複雑化する傾向にある。たとえば、マーケティング分野では、購入サイトの大規模なアクセスデータや顧客データを利用して、顧客を分類し、個別集団や集団間の特徴から価値意識を明らかにしたり、販売戦略を見極めたりすることが重要になっている。そこでは、SNS 上のコメントなどを調査項目に加味し、分析精度を高めたいという要求も出てきている。社会心理学の分野でも、人の態度や行動を分析する際、場面、グループ、観点などをさまざまに変化させた大量の調査項目や観察文書を同時に処理し、多角的に対象者の特徴をつかむことが必要になってきている。この両者で得られるデータは、その目的から、名義尺度や順序尺度の回答、自由記述などのテキストデータを大量に含むことになる。

これらのデータに対して適用される既存の手法は、量的データのための分析が主で、質的データはそのままでは扱えず、大規模性や高次元性への対応も十分ではない。また、テキストに対しても他の特徴量と一緒に分析するなどの研究はまだ多くを見ない。計算においても、並列処理による計算時間短縮の策は講じられるが、計算そのものの効率化を図ったり、隠された特徴を試行錯誤的に把握したりする工夫は指向されていない。

このような背景から、量的データに対する単純な大規模データ処理だけでなく、質的データや非構造的なデータを前提に、冗長性、複雑性を考慮して、高次元の中で集団がもつ潜在的な特徴をいかに考察しやすい形で表現していくか、そのための試行錯誤的なデータ分析手法をどう提供していくかが求められているといえる。

## 2. 研究の目的

本研究では、

- (i) 情報損失を最小限に抑えた大規模さの軽減
- (ii) 縮約した情報や複雑性の関係の効率的な把握
- (iii) 尺度混在データの処理とテキストデータの処理を含む分析手順の提供  
(具体的には、数量化やデータ変換をともなった次元縮約と分類などとの同時推定)
- (iv) 計算の効率化

を順に実現していく形で、情報の損失を最大限に抑え、冗長性や複雑性を考慮した分類や潜在構造の推定により、隠れた構造や特徴を効果的に表現する手法と、その分析を高速に進める環境を開発・提供することである。これらにより、それまでの手法では観察できなかった知見が得られる可能性が大きく、何よりも大規模データ処理の実際的な要求に応えるものとなり、データ分析の発見的情報把握の強力なツールとしての活用が期待される。

## 3. 研究の方法

本研究の目的達成のために、次の機能をもつ情報抽出手法と計算環境の開発を行う。

- ・ 主成分分析の文脈で、次元縮約と変数選択により、重要な次元と変数を絞り込める。
- ・ 次元縮約とクラスタリングなどを組み合わせ、対象の分類や関係の把握ができる。
- ・ 尺度混在データと非構造化データを変換して特徴量として分析に取り込める。
- ・ 計算アルゴリズムの工夫による高速な計算ができる。

研究は、2の(i)~(iv)に対応して、次のように進めていく。

複雑性を処理する(i)は、既存手法の評価とその拡張の検討を行うことである。主成分分析を中心とした従来の次元縮約手法をベースとした手法を検討することを行う。縮約した情報と潜在構造の関係を把握する(ii)は、次元縮約とクラスタリングの同時推定を行う **Reduced k-平均法**の手順を応用して手法の開発を試みる。(iii)では、交互最小二乗法を利用した数量化の研究成果を任意の手法に拡張することを行う。また、本研究で扱う非構造化データをテキストデータの範囲に絞り、テキストマイニングの手法によりテキスト情報を複数の変数へ落とし込み、尺度混在データとして統一的に分析できるようにする。(iv)の加速化は、数値解析的なアプローチを用いて、すでに加速化に成功している主成分分析以外の多変量手法に適用して、その性能を確認する。いずれも人工データや実データで性能の評価を行っていく。

## 4. 研究成果

### 4.1 数量化と次元縮約を伴ったファジィ c-平均法

高次元のカテゴリカルデータを分類することを目的に、数量化と次元縮約を伴ったファジィ c-平均法を提案する。これは、クラスタ分析に存在する「手法の多くが数量データのみを分析の対象としていることから、カテゴリカルデータの分析を前提としていない問題」と「高次元データを扱う際に分類精度が落ちてしまったり、結果の解釈が難しくなったりする問題」の2つの問題を解決することに端を発し、両問題を同時に解決するクラスタリングの新しい推定方法を、クラスタリングの1つの手法であるファジィ c-平均法 (FCM, Fuzzy c-means Clustering) において実

現しようとするものである。本手法を **categorical reduced fuzzy c-means (catRFCM)** とする。

数量化と次元縮約の同時推定の手順は以下のとおりである。

カテゴリカル変数からなる  $n \times p$  のデータ  $\mathbf{X}$  が与えられたとする。数量化と次元縮約の両方を取り入れた FCM の目的関数を考えるにあたり、FCM に、本多ほか (2006) の **categorical fuzzy c-means (catFCM)** と西田 (2010) の **reduced fuzzy c-means (RFCM)** の目的関数を組み合わせる。すなわち、FCM の目的関数をベースとして、まず、catFCM から  $\mathbf{X}$  の各データ  $\mathbf{x}_i$  を数量化した  $\mathbf{g}_{ij}^T \mathbf{q}_j$  に置き換える。ここで、 $\mathbf{g}_j$  はカテゴリカルデータ  $\mathbf{X}$  から生成された第  $j$  変数のダミー変数行列、 $\mathbf{q}_j$  は第  $j$  変数のカテゴリ得点である ( $j = 1, \dots, p$ )。次に、RFCM を利用して、中心座標を主成分座標上の  $\mathbf{A}\mathbf{f}_k$  に置き換える。ここで、 $\mathbf{A}$  は  $p \times r$  の主成分負荷行列、 $\mathbf{f}_k$  はクラスタ中心における  $c \times r$  の主成分得点行列  $\mathbf{F}$  の  $k$  番目の行である ( $c$  はカテゴリの数、 $r$  は主成分数、 $1 \leq r < p$  である)。以上より、提案手法である catRFCM の目的関数を以下のように定める。

$$\text{catRFCM}(\mathbf{U}, \mathbf{H}, \mathbf{A} | \mathbf{X}) = \sum_{i=1}^n \sum_{k=1}^c u_{ik}^m \|\mathbf{g}_{ij}^T \mathbf{q}_j - \mathbf{A}\mathbf{f}_k\|^2$$

catRFCM は、この目的関数を最小化する。つまり、提案手法は、数量化された個体と主成分座標上のクラスタ中心との距離の二乗和が最小となるようにクラスタリングを行うものである。

catRFCM のアルゴリズムは以下のとおりである。

[Step1] ファジィ化パラメータ  $m$ , クラスタ数  $c$ , 主成分数  $r$  の値を決め、制約を満たしたうえで、メンバシップ行列  $\mathbf{U}$ , クラスタ中心行列  $\mathbf{H}$ , 主成分負荷行列  $\mathbf{A}$  を初期化する。

[Step2] 第  $j$  変数のカテゴリ得点  $\mathbf{q}_j$  を求め、以下、 $\mathbf{g}_{ij}^T \mathbf{q}_j$  を  $\mathbf{x}_{ij}$  とおく。

[Step3]  $\mathbf{H}$  を計算し、 $\mathbf{F} = \mathbf{H}\mathbf{A}$  で主成分得点行列  $\mathbf{F}$  を更新する。

[Step4]  $\mathbf{A}$  を更新する。

[Step5]  $\mathbf{U}$  を更新する。

[Step6] すべてのパラメータ ( $\mathbf{U}$ ,  $\mathbf{H}$ ,  $\mathbf{A}$ ) が収束していれば終了、そうでなければ、Step2 へ戻り計算を続ける。

$n=100$ ,  $p=12$ ,  $c=4$  として、指定したクラスタ数をもつ 100 セットの人工データを Fuzziness Performance Index (FPI), Normalized Classification Entropy (NCE), Misclassification Probability (MCP) の 3 つの指標で評価した結果が図 4.1 である。また、ファッションに関するアンケート調査 (大藪 他, 2018; Ohyabu et al., 2019) に提案手法を適用して得られた分類結果が図 4.2 である。これらの結果から、クラスタをより明確に分離すること、タンデムクラスタリングと比較して分類精度が高いこと、実際の場面に解釈可能な分析結果を導出することが確認できた。

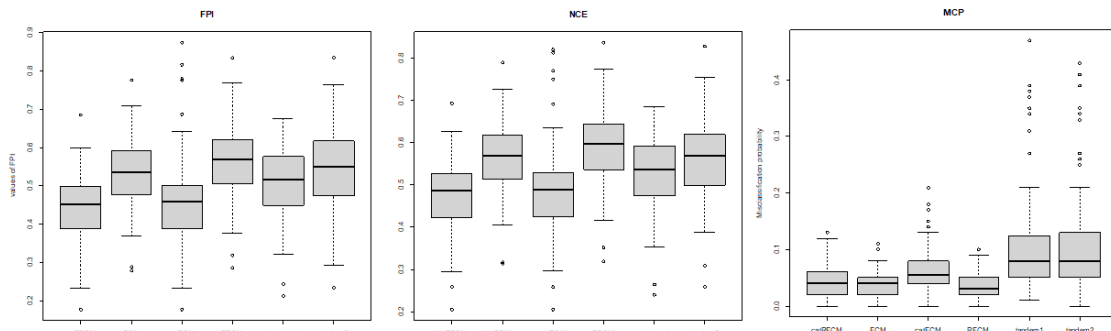


図 4.1 手法ごとの 100 回のシミュレーションによる FPI, NCE, MCP の箱ひげ図 (各図の左から, catRFCM, FCM, catFCM, RFCM, tandem1, tandem2)

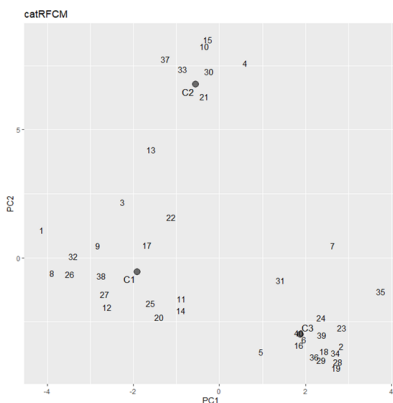


図 4.2 クラスタリング結果

Ohyabu, A., Kuroda, M., Seino, S., & Zhang, J. (2019). Exploring interplay among consumer engagements with multiple objects. The 10th Years Naples Forum on Service-Service Dominant Logic, [http://www.naplesforumonservice.it/uploads/files/2018/BookOfAbstract/BoA\\_Ohyabu-Kuroda-Seino-Zhang.pdf](http://www.naplesforumonservice.it/uploads/files/2018/BookOfAbstract/BoA_Ohyabu-Kuroda-Seino-Zhang.pdf)

大藪 亮, 黒田正博, 山口隆久 (2018). ファッション消費経験における消費者エンゲージメント. 日本マーケティング学会カンファレンス・プロシーディングス, vol.7, 238-240.

西田豊 (2010). 曖昧さを表現する概念形成モデル. 知能と情報 (日本知能情報ファジィ学会誌), 22(4), 434-442.

本多克宏・上杉亮・市橋秀友 (2006). 混合データベースの FCM クラスタリング. 知能と情報 (日本知能情報ファジィ学会誌), 18(4), 598-608.

#### 4.2 非計量主成分分析を用いた項目選択

複雑な構造をもつデータでの項目選択について、「ファッション」、「ブランド」、「ショップスタッフ」という3つの消費者エンゲージメントを調査した消費者エンゲージメントのデータ(大藪 他, 2018; Ohyabu et al., 2019) のデータを用い, 全 85 項目から項目選択をする場合(ケース i) と, 3つのエンゲージメントの群が独立の情報を収集していると考えて, 群ごとに項目選択をする場合(ケース ii) の2つのケースから, 非計量主成分分析における項目選択を検討する。

ここでは, カテゴリカルデータの項目選択の方法として, Mori et al. (2017) で理論が整理された非計量主成分分析(NLPCA, Nonlinear Principal Component Analysis)と拡張主成分分析(M.PCA, Modified Principal Component Analysis)を同時に行うNL.M.PCA(NL.M.PCA, Nonlinear Modified Principal Component Analysis)を利用する。すなわち, 本研究は, NL.M.PCAの実データへの応用事例がまだないことに対して, 実データを用いた項目選択を実行することにより, 選択のあり方を検討するとともに, データの構造に応じたNL.M.PCAの利用方法の可能性を探るものである。

項目を特定する方法として, 調査者が任意の個数を指定するパターン a と規準値を用いるパターン b, これを全 85 項目から選択するケース i と, 群ごとに項目選択をするケース ii に当てはめて項目選択を行う。その結果, 各項目が選択過程において何回含まれていたかを示したものが表 4.1 である。太字の数字は全体の7割以上の回数で選ばれたものを表している。

これより, ケース i, ケース ii とともに, ショップスタッフに対する項目が早い段階から落とされ, ファッションとブランドについての項目が多く残っているが, 3つの群に分けて項目を選択した方がそれぞれの内容を代表する項目をバランスよく選んでいることがわかる。ここに残った項目が全体の情報をよく保持している質問であり, 今後の調査でも必ず用いるべきと判断される。一方, 選択された回数の少ない項目は, 他の質問で代表されうるので, 質問として必要であったかどうかを考える材料にもなる。以上より, NL.M.PCA がカテゴリカルデータに対して, 全体を再現する項目の精選が行えており, 逆に, 選択結果を質問自体の評価にもつなげることができるものであることがわかった。

表 4.1 項目が選ばれた回数

| 回答者の基本属性 |      | パターン      | V1        | V2        | V3        | V4        |           |           |           |           |           |           |           |           |  |  |
|----------|------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|--|--|
|          | i    |           | 32        | <b>73</b> | 39        | 28        |           |           |           |           |           |           |           |           |  |  |
|          | ii   |           | <b>12</b> | 7         | 7         | 9         |           |           |           |           |           |           |           |           |  |  |
| ファッション   | 関与   | パターン      | V5        | V6        | V7        | V8        |           |           |           |           |           |           |           |           |  |  |
|          |      | i         | <b>66</b> | 55        | 50        | 22        |           |           |           |           |           |           |           |           |  |  |
|          | ii   | 6         | 8         | 4         | 6         |           |           |           |           |           |           |           |           |           |  |  |
|          | 顧客   | パターン      | V17       | V18       | V19       | V20       | V21       | V22       | V23       | V24       | V25       | V26       | V27       |           |  |  |
|          |      | i         | 36        | <b>59</b> | 10        | 31        | 6         | <b>78</b> | <b>62</b> | 18        | 13        | <b>71</b> | 47        |           |  |  |
|          | ii   | <b>25</b> | 19        | 18        | 22        | 11        | <b>31</b> | 23        | 4         | 9         | <b>27</b> | <b>31</b> |           |           |  |  |
|          | 価値   | パターン      | V52       | V53       | V54       | V55       | V56       | V57       | V58       | V59       | V60       | V61       | V62       |           |  |  |
|          |      | i         | <b>81</b> | 16        | 37        | 38        | <b>57</b> | 44        | 24        | <b>81</b> | <b>69</b> | 53        | 20        |           |  |  |
|          | ii   | <b>29</b> | <b>30</b> | 20        | 22        | 13        | <b>30</b> | <b>26</b> | 7         | 14        | 15        | 9         |           |           |  |  |
|          | ブランド | 関与        | パターン      | V9        | V10       | V11       | V12       |           |           |           |           |           |           |           |  |  |
| i        |      |           | 45        | 11        | 9         | <b>72</b> |           |           |           |           |           |           |           |           |  |  |
| ii       |      | 6         | 10        | 12        | 6         |           |           |           |           |           |           |           |           |           |  |  |
| 顧客       |      | パターン      | V28       | V29       | V30       | V31       | V32       | V33       | V34       | V35       | V36       | V37       | V38       | V39       |  |  |
|          |      | i         | 23        | 52        | <b>76</b> | <b>63</b> | 29        | 12        | 42        | <b>58</b> | 26        | <b>81</b> | 35        | <b>81</b> |  |  |
| ii       |      | 7         | 20        | <b>31</b> | 22        | 5         | 19        | 17        | <b>29</b> | 10        | 24        | 24        | 13        |           |  |  |
| 価値       |      | パターン      | V63       | V64       | V65       | V66       | V67       | V68       | V69       | V70       | V71       | V72       | V73       | V74       |  |  |
|          |      | i         | 14        | <b>70</b> | 54        | 21        | 41        | <b>61</b> | <b>74</b> | 30        | 48        | 43        | <b>80</b> | <b>67</b> |  |  |
| ii       |      | 10        | 23        | <b>25</b> | 17        | 6         | 13        | 5         | 23        | <b>29</b> | 15        | 21        | 20        |           |  |  |
| ショップスタッフ |      | 関与        | パターン      | V13       | V14       | V15       | V16       |           |           |           |           |           |           |           |  |  |
|          | i    |           | 34        | 40        | 7         | <b>60</b> |           |           |           |           |           |           |           |           |  |  |
|          | ii   | 5         | 11        | 9         | 8         |           |           |           |           |           |           |           |           |           |  |  |
|          | 顧客   | パターン      | V40       | V41       | V42       | V43       | V44       | V45       | V46       | V47       | V48       | V49       | V50       | V51       |  |  |
|          |      | i         | 46        | 1         | <b>81</b> | 4         | 25        | <b>75</b> | 7         | 51        | 2         | <b>68</b> | 19        | <b>65</b> |  |  |
|          | ii   | 16        | 12        | 19        | 12        | 13        | 19        | 6         | 19        | 12        | 19        | 8         | <b>28</b> |           |  |  |
|          | 価値   | パターン      | V75       | V76       | V77       | V78       | V79       | V80       | V81       | V82       | V83       | V84       | V85       |           |  |  |
|          |      | i         | 17        | <b>79</b> | 8         | 56        | 33        | <b>77</b> | 16        | 3         | <b>64</b> | 27        | 49        |           |  |  |
|          | ii   | 18        | 2         | <b>28</b> | 3         | 20        | <b>29</b> | 15        | 2         | 22        | 15        | 8         |           |           |  |  |

Mori, Y., Kuroda, M., Makino, N. (2017). Variable selection in nonlinear principal component analysis. Nonlinear Principal Component Analysis and Its Applications (JSS Research Series in Statistics), 31-45, Springer.

#### 4.3 テキストデータによる感情分析の性能向上への試み

非構造化データに対する特徴量の抽出については, テキストデータの感情分析を取り上げ, 汎用の極性辞書ではなく, 対象に適した辞書を構築することで, ネガティブ・ポジティブの判断をより正確に行うことを試みた。そこでは, トピックモデルとヒートマップを用いた。

既存の汎用辞書として, 単語感情極性対応表を用いる。対象レビューは, 「ラゲリより愛をこめて」, 「永遠の0」, 「海賊と呼ばれた男」のレビューである。その際, レビューにつけられているスコア 4, 5 を positive, 1, 2 を negative とし, positive に分類されたレビューから 400 件, negative から 220 件をランダムに抽出し, 分析に用いた。

新たな辞書の構築には, まず, 対象の分析データに LDA (Latent Dirichlet Allocation) を適用し, トピックの出現確率を外部変数を基に比較し, positive が大きければ positive トピックとする。negative が大きければ negative トピックとし, そのトピックにおける出現確率の値の係数に - を付与する。その後, ヒートマップで negative と positive での出現確率の差が 0.01 未満であるトピックは, 語の頻度を確認し, positive ビューでの出現が多ければ, positive 語, negative レビューでの出現が多ければ negative 語とした。最後に, トピックの出現確率とそのトピックを構成する語のトピック内での出現確率を掛け合わせ, その語の感情極性値として, 辞書を完成させた。



ラベル付けしたレビューを正解データとし、既存辞書と提案辞書の判定精度を比較した。その結果、positive レビューでは、従来の辞書では適切な判定ができなかったが、提案辞書により精度の向上を図ることができたこと、negative レビューでは、positive 判定された 37 件のうち 4 分の 1 は文中に肯定的な表現があり、positive 判定はおかしくないものであることがわかった。positive、negative の設定をスコアを基に行っていることに改良は必要と考えられるが、トピックモデルとヒートマップによって辞書生成が可能であり、かつ、当該分野に特化した辞書の有用性を確認することができた。

表 5.1 感情分析の結果

| 手法 | 表 4.2    |          |      |          |          |      |
|----|----------|----------|------|----------|----------|------|
|    | positive |          |      | negative |          |      |
|    | positive | negative | 判定不可 | positive | negative | 判定不可 |
| 既存 | 0        | 388      | 15   | 0        | 215      | 5    |
| 提案 | 131      | 267      | 2    | 37       | 182      | 1    |

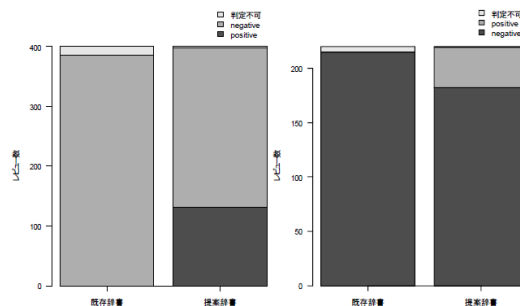


図 4.3 ネガポジ別感情分析の結果

#### 4.4 ファジィ c-平均法の加速化

ファジィ c-平均法の加速化については、これまで、ve アルゴリズム (Wynn, 1962) を用いた提案を行い、限られた範囲での性能の評価を行ってきた (Yoshioka et al., 2018 など) が、本研究では、そのシミュレーションの範囲を広げ、より詳細な性能の評価を行った。

アルゴリズムは、次のとおりである。

[Step1] チューニングパラメータ  $m$  と、クラスター数  $K$  の初期値を与える。

[Step2]  $A$  の初期値  $A^{(0)}$  を定める。定めた初期値を代入した  $A$  を  $\hat{A}^{(0)} = (\hat{c}_1^{(0)}, \dots, \hat{c}_k^{(0)})$  とする。

[Step3]  $\hat{A}^{(t)}$  を固定し、 $\hat{U}^{(t)} = \arg \min_{U \in M_f} F(U, \hat{A}^{(t)})$  を求める。

[Step4]  $\hat{U}$  を固定し、 $\hat{A}^{(t+1)} = \arg \min_A F(\hat{U}^{(t)}, A)$  を求める。

[Step5] ve アルゴリズムにより  $\{\hat{A}^{(t)}, \hat{A}^{(t+1)}, \hat{A}^{(t+2)}\}$  を用いて、 $\text{vec } \hat{A}^{(t)} = A^{(t+1)} + \left[ [\Delta A^{(t+1)}]^{-1} - [\Delta A^{(t)}]^{-1} \right]^{-1}$  を計算し、ve 加速列  $\{\text{vec } \hat{A}_\varepsilon^{(t)}\}_{t \geq 0}$  を生成する。ここで、 $\text{vec } \hat{A}$  は、 $\hat{A}$  をベクトル化したものである。

[Step6]  $\|\text{vec } \hat{A}_\varepsilon^{(t)} - \text{vec } \hat{A}_\varepsilon^{(t-1)}\|^2 < \delta$  より収束判定をする。終了していなければ、Step3 へ戻り計算を続ける。

収束判定を  $10^{-9}$  とし、シミュレーション回数を 100 回、人工データの発生には R の MixSim を使い、反復回数と CPU 時間 (秒) で評価を行った。比較は、加速しないオリジナルの計算 (FCM と略記)、メンバーシップ  $U$  を加速する計算 (ve-FCM1)、中心  $H$  を加速する計算 (ve-FCM2) の 3 つである。データのサイズを変化させた場合、クラスタの重なりやファジネスを変化させた場合について 100 回の挙動を見た (一部を図 4.4 に示す)。また、Abalone データと Wine データ (ともに、UCI Machine Learning Repository [http://archive.ics.uci.edu/ml] より) に適用した (表 4.2)。その結果、ファジィ c-平均法では、おおむね 2.0 倍 (1.2~2.4) の加速が得られ、データサイズが大きくなるほど、また、クラスタの重なりが大きいほど、効果を発揮することがわかった。

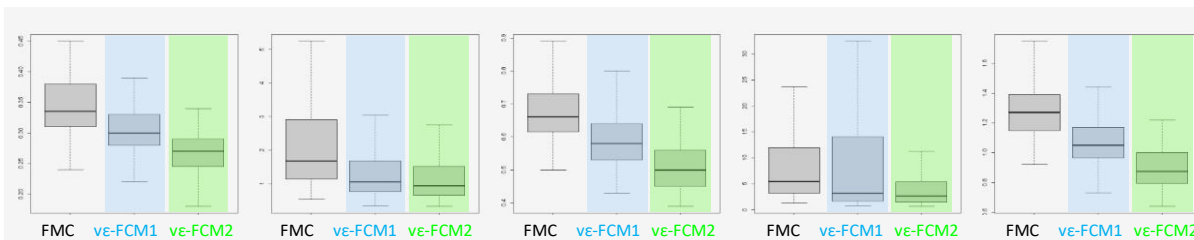


図 4.4 CPU 時間の比較

表 4.3 Abalone data での実施結果

| # of clusters | # of iteration |         | CPU time |         | Speed Up       |          |
|---------------|----------------|---------|----------|---------|----------------|----------|
|               | FCM            | ve-FCM2 | FCM      | ve-FCM2 | # of iteration | CPU time |
| K=3           | 129            | 54      | 7.52     | 3.32    | 2.39           | 2.27     |
| K=4           | 287            | 116     | 20.88    | 8.82    | 2.47           | 2.37     |
| K=5           | 195            | 83      | 17.91    | 7.94    | 2.35           | 2.26     |
| K=6           | 238            | 105     | 26.03    | 11.7    | 2.27           | 2.22     |
| K=7           | 648            | 268     | 82.8     | 34.71   | 2.42           | 2.39     |

表 4.4 Wine quality data での実施結果

| # of cluster | # of iteration |         | CPU time |         | Speed Up       |          |
|--------------|----------------|---------|----------|---------|----------------|----------|
|              | FCM            | ve-FCM2 | FCM      | ve-FCM2 | # of iteration | CPU time |
| K=2          | 561            | 301     | 55.87    | 28.89   | 1.86           | 1.93     |
| K=3          | 270            | 218     | 39.86    | 30.64   | 1.24           | 1.30     |
| K=4          | 613            | 346     | 118.87   | 64.86   | 1.77           | 1.83     |
| K=5          | 670            | 352     | 168.97   | 82.75   | 1.90           | 2.04     |
| K=6          | 651            | 283     | 201.41   | 82.19   | 2.30           | 2.45     |
| K=7          | 698            | 314     | 249.41   | 107.94  | 2.22           | 2.31     |
| K=8          | 769            | 369     | 322      | 146.89  | 2.08           | 2.19     |
| K=9          | 763            | 343     | 343.91   | 149.33  | 2.22           | 2.30     |
| K=10         | 784            | 365     | 394.61   | 171.24  | 2.15           | 2.30     |

Wynn, P. (1962). Acceleration techniques for iterated vector and matrix problems. *Mathematics of Computation* 16: 301–322.  
 Yoshioka, M., Kuroda, M., Mori, Y. (2018). Acceleration of computation for fuzzy c-means clustering. The 23rd International Conference on Computational Statistics (COMPSTAT2018).

## 5. 主な発表論文等

〔雑誌論文〕 計5件（うち査読付論文 5件/うち国際共著 0件/うちオープンアクセス 5件）

|   |                       |
|---|-----------------------|
| 1. 著者名<br>千足南々子・森 裕一  | 4. 巻<br>5             |
| 2. 論文標題<br>ビッグファイブパーソナリティとデザイン選択の関連性について  | 5. 発行年<br>2023年       |
| 3. 雑誌名<br>経営とデータサイエンス   | 6. 最初と最後の頁<br>125-135 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>なし   | 査読の有無<br>有            |
| オープンアクセス<br>オープンアクセスとしている（また、その予定である）   | 国際共著<br>-             |
| 1. 著者名<br>稲田 愛・森 裕一   | 4. 巻<br>5             |
| 2. 論文標題<br>テキストマイニングによる傾向・様相の分析   | 5. 発行年<br>2023年       |
| 3. 雑誌名<br>経営とデータサイエンス   | 6. 最初と最後の頁<br>41-53   |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>なし   | 査読の有無<br>有            |
| オープンアクセス<br>オープンアクセスとしている（また、その予定である）   | 国際共著<br>-             |
| 1. 著者名<br>Kuroda, M., Mori, Y., Iizuka, M.  | 4. 巻<br>38            |
| 2. 論文標題<br>Speeding up the convergence of the alternating least squares algorithm using vector acceleration and restarting for nonlinear principal component analysis | 5. 発行年<br>2023年       |
| 3. 雑誌名<br>Computational Statistics  | 6. 最初と最後の頁<br>243-262 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1007/s00180-022-01225-4   | 査読の有無<br>有            |
| オープンアクセス<br>オープンアクセスとしている（また、その予定である）   | 国際共著<br>-             |
| 1. 著者名<br>赤木辰伎・森 裕一・黒田正博・飯塚誠也   | 4. 巻<br>6             |
| 2. 論文標題<br>数量化と次元縮約を伴ったファジィc-平均法  | 5. 発行年<br>2024年       |
| 3. 雑誌名<br>経営とデータサイエンス   | 6. 最初と最後の頁<br>80-93   |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>なし   | 査読の有無<br>有            |
| オープンアクセス<br>オープンアクセスとしている（また、その予定である）   | 国際共著<br>-             |

〔学会発表〕 計9件（うち招待講演 2件 / うち国際学会 6件）

|   |
|---|
| 1. 発表者名<br>Mori, Y., Akaki, T., Kuroda, M.                    |
| 2. 発表標題<br>Acceleration of Computation in Fuzzy Clustering    |
| 3. 学会等名<br>The IASC-ARS Interim Conference 2022 (招待講演) (国際学会) |
| 4. 発表年<br>2022年   |

|  |
|--|
| 1. 発表者名<br>片山浩子, 森 裕一                        |
| 2. 発表標題<br>質的データの項目精選 - 項目反応理論を利用した項目選択の検討 - |
| 3. 学会等名<br>日本計算機統計学会第36回シンポジウム               |
| 4. 発表年<br>2022年                              |

|  |
|--|
| 1. 発表者名<br>Katayama, H., Nishiyama, C., Mori, M.   |
| 2. 発表標題<br>Item Selection for qualitative data   |
| 3. 学会等名<br>The 11th Conference of the Asian Regional Section of the International Association for Statistical Computing (IASC-ARS2022)<br>(国際学会) |
| 4. 発表年<br>2022年  |

|  |
|--|
| 1. 発表者名<br>Nishiyama, C., Katayama, H., Mori, M.   |
| 2. 発表標題<br>Partial least squares for qualitative data  |
| 3. 学会等名<br>The 11th Conference of the Asian Regional Section of the International Association for Statistical Computing (IASC-ARS2022)<br>(国際学会) |
| 4. 発表年<br>2022年  |

|                             |
|-----------------------------|
| 1. 発表者名<br>片山浩子, 西山ちとせ, 森裕一 |
| 2. 発表標題<br>質的データの項目選択       |
| 3. 学会等名<br>日本消費経済学会第46回全国大会 |
| 4. 発表年<br>2021年             |

|   |
|---|
| 1. 発表者名<br>Akaki, T., Mori, Y., Kuroda, M., Iizuka, M.                |
| 2. 発表標題<br>Clustering with quantification and dimension reduction     |
| 3. 学会等名<br>The 8th Japanese-German Symposium on Classification (国際学会) |
| 4. 発表年<br>2023年   |

|   |
|---|
| 1. 発表者名<br>Akaki, T., Mori, Y., Kuroda, M., Iizuka, M.  |
| 2. 発表標題<br>Dimension-reduced fuzzy clustering for categorical data  |
| 3. 学会等名<br>The 12th Conference of the Asian Regional Section of the International Association for Statistical Computing (IASC-ARS2023) (国際学会) |
| 4. 発表年<br>2023年   |

|  |
|--|
| 1. 発表者名<br>Akaki, T., Mori, Y., Kuroda, M., Iizuka, M.   |
| 2. 発表標題<br>Acceleration and quantification with dimension reduction in fuzzy clustering                  |
| 3. 学会等名<br>Statistical Computing And Robust Inference For High Dimensional Data (SCRI2023) (招待講演) (国際学会) |
| 4. 発表年<br>2023年  |



|  |
|--|
| 1. 発表者名<br>赤木辰伎・森 裕一・黒田正博・飯塚誠也                         |
| 2. 発表標題<br>ファジィクラスタリングにおける高次元質的データの扱い                  |
| 3. 学会等名<br>北海道大学情報基盤センター萌芽型共同研究集会「第43回大規模データ科学に関する研究会」 |
| 4. 発表年<br>2024年  |

〔図書〕 計0件

〔産業財産権〕

〔その他〕

-

#### 6. 研究組織

|       | 氏名<br>(ローマ字氏名)<br>(研究者番号)                | 所属研究機関・部局・職<br>(機関番号)              | 備考 |
|-------|--|------------------------------------|----|
| 研究分担者 | 黒田 正博<br>(KURODA Masahiro)<br>(90279042) | 岡山理科大学・経営学部・教授<br><br>(35302)      |    |
| 研究分担者 | 飯塚 誠也<br>(IIZUKA Masaya)<br>(60322236)   | 岡山大学・全学教育・学生支援機構・教授<br><br>(15301) |    |

#### 7. 科研費を使用して開催した国際研究集会

〔国際研究集会〕 計1件

|  |                    |
|--|--------------------|
| 国際研究集会<br>岡山理科大学マネジメント学会 第9回研究会「Deep learning and Computational aspects」 | 開催年<br>2022年～2022年 |
|--|--------------------|

#### 8. 本研究に関連して実施した国際共同研究の実施状況

|         |         |
|---------|---------|
| 共同研究相手国 | 相手方研究機関 |
|---------|---------|