様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月 １９ 日現在

機関番号：３３９０８
研究種目：基盤研究(C)（一般）
研究期間：2020～2023
課題番号：２０Ｋ０３３０３

研究課題名（和文）カテゴリカルデータのための正規直交主成分分析の開発とその応用

研究課題名（英文）Development of Orthonormal principal component analysis for categorical data and its applications

研究代表者
村上　隆（Murakami, Takashi）

中京大学・文化科学研究所・特任研究員

研究者番号：７００９３０７８

交付決定額（研究期間全体）：（直接経費）　　1,700,000円

研究成果の概要（和文）：本研究の主たる目的は、多重対応分析（multiple correspondence analysis; MCA)を改良することであった。MCAの解釈は主にグラフィカルな表現にもとづいて行われてきたために、３次元以上の解を解釈することが困難であった。我々は、解釈の手掛かりをグラフから軸に移すために、独立クラスター回転を伴う負荷行列を導入した。心理測定、ならびに社会調査の実データを用いたいくつかの研究は、新たな知見と複数の理論的仮説を検証することにより、我々の「カテゴリカルデータのための正規直交主成分分析」（OPCA）の有効性を明らかにした。

研究成果の学術的意義や社会的意義
質的、あるいは順序的応答カテゴリーを用いる心理測定、社会調査は広く実施され、膨大なデータが得らつつある。そうしたデータの分析には、因子分析や各種の数量化の方法（多重対応分析はそのひとつ）が用いられてきた。心理測定データについては、項目反応理論や構造方程式モデルといった潜在変量を用いられた分析が近年台頭しているが、モデルの前提条件や複雑性が、心理測定モデルに不適合であることが多く、社会調査に関しては、本来心理測定以上に多次元性が想定されるデータの分析が、無理に少数の次元に落とし込まれる傾向があった。本研究は、それらの問題点を解決する一助となることが期待できる。

研究成果の概要（英文）：The main objective of this study was to improve Multiple Correspondence Analysis (MCA). Because interpretations of MCA results have mainly been based on graphical representations, reading the solution with more than three dimensions took a lot of work. We introduced the loading matrix and its independent cluster rotation to shift the main clue of interpretations from the diagrams to axes. Orthonormal polynomials and the design matrices transformed to orthonormality, used to give numerical values to response categories, were consistent with the traditional quantification in MCA. Several application studies using real psychometric and social survey data demonstrated the efficiency of our Orthonormal Principal Component Analysis (OPCA) for categorical data by giving some new findings and confirming a few theoretical hypotheses.

研究分野：社会科学

キーワード：多重対応分析　主成分分析　因子の回転　心理測定　社会調査　カテゴリカルデータ

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

(1) 評定尺度を用いた心理測定においては、直接観測不可能な、(多変量) 正規分布する潜在変量の存在を仮定し、回答者は潜在変量上に固定した位置を占めるものとする。潜在変量とカテゴリカルな反応の次元（これも連続変量）の関係を線形モデル（構造方程式モデル）で仮定し、実際の反応は反応の次元を全回答者に共通に設定された境界点をによってカテゴリー化されているものとする。こうした前提が非現実的であることは明らかであり、実際、記述的方法である多重対応分析（Multiple Correspondence Analysis (MCA)）の結果は、上記のモデルでは説明できない反応傾向を示していた（村上，2019）。

(2) 社会調査データの分析においては、反応をカテゴリカルなものとみなした MCA の適用が多くみられたが、その解釈は主に散布図によるグラフィカルな表現にもとづいており、３次元以上の解の解釈は難しかった。また、空間表現を言語的表現に移す過程も主観性が大きく働き、これをロールシャッハ検査のような投影法と陰口をたたく向きもあった。かなり強い理論的仮定にもとづく心理測定と比較して、より探索的な要素が大きく、必然的に多次元になりがちな社会調査データの分析・解釈において、これは不都合な性質であった。

(3) 村上（2016）、Murakami (2020) は、MCA の結果の解釈の手掛かりを空間表現から軸（への射影）に移すことを目指し、「カテゴリカルデータのための正規直交主成分分析」（Orthonormal Principal Component Analysis, OPCA）を開発した。この方法により、心理測定データでは、構造方程式モデルによっては発見できない反応の変動が検出でき、社会調査データでは、解釈の手掛かりを空間表現から複数の軸の座標に移すことにより、多次元の解が解釈しやすくなることが示唆されていた。

２．研究の目的

本研究課題の目的は以下の２点にまとめられる。

(1)　OPCA の理論面での整備をはかる。具体的には、直交多項式によるカテゴリーの数量化が、MCA の当てはまりのよさを変えないことの証明、独立クラスター回転の負荷行列の列方向の平方和が、当該次元によって説明されるデータの分散の大きさ、あるいは、数量化得点の分散に対応することの証明、完全単純構造の負荷行列を得るための主クラスター成分分析（Principal Cluster Component Analysis, PCCA）の制約付き主成分分析としての定式化の導出である。

(2) データの分析・解釈を複数行うことにより、OPCA の実用的価値が示されるようなデータの分析例を蓄積する。心理測定データにおいては、構造方程式モデルによっては説明できない項目反応のルールの個人差の類型を発見すること、社会調査データでは、従来の MCA によっては見出されなかった個人差の次元を見出すことを目指す。

３．研究の方法

(1) 理論面の整備に関しては、主として線形代数にもとづく式の証明と、コンピュータ・シミュレーションによる確認。

(2) 実データの分析に関しては、手持ちのデータ、ならびにインターネットからダウンロードできる公開されたデータを用いた分析を行う。特に、本研究のための調査等は実施していない。

データ分析、コンピュータシミュレーションに用いたソフトウェアは、Matlab (2020 以降 2023 まで)、M*plus* 8.6 以降のバージョン、SPSS 25、Excel 2014 等である。

(3) 使用したデータは、心理測定関係では、申請者がかかわってデータの使用を許可されている中村（2000）による、14 種のパーソナリティ尺度を 1645 名の大学生に実施することによって得られたもの、社会調査関係では、申請者が勤務先大学で学生実習として実施したプロ野球観戦者を対象とした調査、および所属する研究所によるスポーツ博物館観覧者に実施した調査のデータである。また、一部インターネットに公開されている海外のデータをダウンロードして使用したものもある。

４．研究成果

(1) 理論面における成果

①正規直交多項式と正規直交化したデザイン行列が MCA の最適化基準を変えないことの証明

カテゴリーに $c-1$ 次元の正規直交な数値を割り当てる限り、MCA の最適化基準を変えることがない。このことは経験的事実としては確認していたが、正式の証明を行った（Murakami, 2020）。

②独立クラスター回転によるパターン行列の列ごとの平方和の意味

独立クラスター回転（Harris & Kaiser, 1964）は、重み行列の直交回転による斜交回転と呼ばれるように、本質的に斜交回転（数量化得点＝主成分得点間に非ゼロの相関が発生する）である。通常の斜交回転ではパターン行列の列方向の平方和には意味付けができない。しかしながら、独立クラスター分析によって得られるパターン行列の列方向の平方和は、当該次元の数量化得点の分散、あるいは、その数量化得点が説明するデータの変動の大きさと解釈できることを示す

（村上、2023）。

③主クラスター分析（PCCA）の理論面の整備

20 世紀半ばに旧ソ連の研究者によって開発された主クラスター分析という方法は、観戦単純構造をなす負荷行列をクラスター分析によって得る方法であるが、方法の定式化もアルゴリズムもヒューリスティックに導出されており、理論的根拠がはっきりしなかった。本研究は、これを制約付き主成分分析として数理的に導出したものである。それにより、全変数を１つのクラスターとすることから出発するアルゴリズムに、空のクラスターが残らないことなど、若干の有用な性質も求められた（Murakami, 2024）。なお、この方法の適用は、通常の量的データにとどまっており、

（２）実データの分析における成果

① 評価項目の反応ルールの個人差

心理測定、マーケット調査等における質問への３～７段階評定形式の項目（Likert 型項目）への反応については、従来、連続的な潜在変量上で正規分布するの個人の位置が、全回答者に共通の線形モデル（構造方程式モデル）によって各項目反応への連続体に写像され、それがやはり全回答者に共通の境界点によって離散的なカテゴリー反応につながるというモデルで分析されることが一般的であった。しかしながら、実データと同じ平均値、標準偏差、項目間相関をもつ人工データを、上述のモデルに従って生成し、それを OPCA によって実データと同じ手続きで分析してみると、いくつかの点で異なった特徴が認められることがわかった。たとえば、図１に見られるように、通常の相関行列では、実データとシミュレーションデータの固有値にはまったく違いが見られないのに対し、MCA の基準化 Byrt 行列では大きな差が見られる。すなわち、第２から第４固有値において、実データの固有値が大幅に上回る。

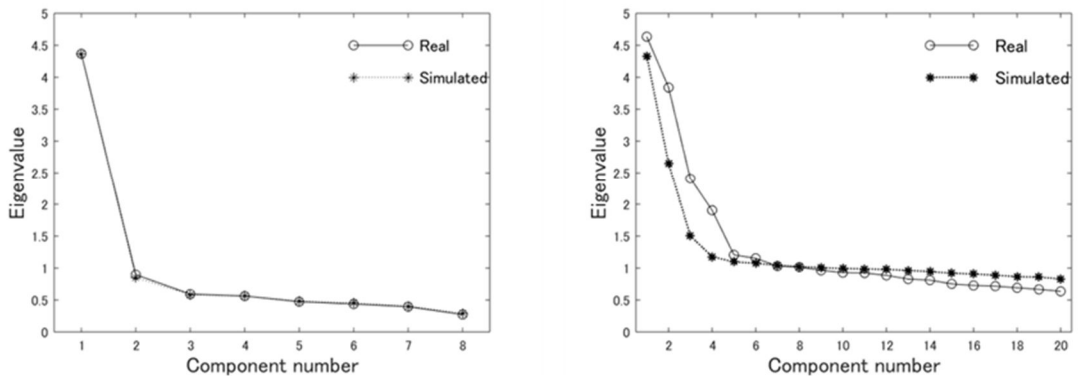

図１　変数間相関行列の固有値と MCA の基準化 Byrt 行列の固有値。実データ（Real）とシミュレーションデータ（Simulated）の比較。データは、2017 年６月実施のプロ野球観戦者に対する調査から、球団・球場に対する８つの評価項目（5-point）。*N* = 811（村上、2019）。

対応する数量化得点について見ると、第１と第 2 の次元の間の関係についても、図２に見られるような差がある。すなわち、実データもシミュレーションデータも、全体として２次関数をいわば包絡線とする分布（馬蹄現象と呼ばれる）を示しているが、実データにはすべての項目に対して同一の反応をする停留点（anchoring points）への集中と、放物線から大きく逸脱する回答者が比較的多く見れらる点が異なっている。

これについては、近年台頭しつつある person-oriented approach（たとえば、Bergman & Lundh, 2015）の観点、すなわち、回答者が異なる反応ルールをもつ下位グループに分類できるといった観点から解釈できる可能性もあり、これについては、今後の検討課題である。いずれにしてもこれらの結果は、現在心理学的データ分析の定番となっている構造方程式モデルの適用について、根本的な再検討を要請する事実であると考えられ、改めて記述的方法である MCA や OPCA による心理測定データの分析の有効性を示唆するものである。

② 心理測定型項目を含む社会調査データの分析

前述のように OPCA 開発の目的は、3 次元以上の解の解釈を容易にすることであった。これによって、多様な内容・形式をもつ質問項目を同時に１つの方法により一度に分析することが目触れることになる。しかしながら、もともと少数の次元上での個人差をできるだけ高い信頼性で測定することを目指すため、比較的等質的な（相互相関の高い）比較的多数の項目からなる心理測定尺度項目を、回答者のなるべく多くの側面を記述するために多様性の高い（相互に連関する可能性の低い）項目からなる社会調査の項目群に含めて実施し、それらを同時に分析しようとすれば、説明される分散の大部分は心理測定型項目に偏り、社会調査項目の座標を覆い隠してしまうことになる。

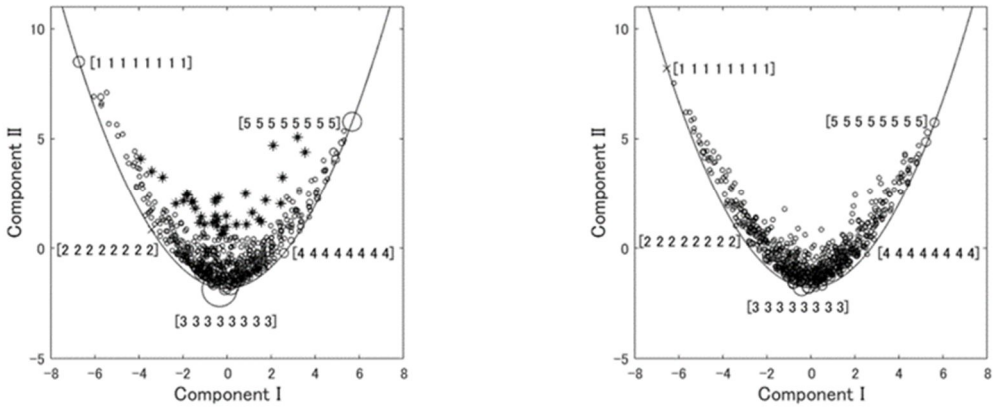

図２　数量化得点の分布の比較。左が実データ、右がシミュレーションデータ。こうした違いはほとんどの Likert 型項目への反応においてみられるが、その特徴は多様である（村上,2019）。

この問題に関する完全な解決策はまだ見出されていないが、村上(2022)においては、そこで分析した調査に含まれた①で分析の対象とした 8 項目が含まれており、この影響はあまりにも大きい。しかしながら、これらの項目の高次の効果（特に 2 次の効果）も捨てがたい。これらの 8 項目を分離して OPCA にかけ、その数量化得点を他の社会調査項目とあわせて分析することも考えられたが、連続変量を離散化する方法など、主観が介在せざるを得ない手続きが回避できない。そこで、これらの 8 項目を５-class の潜在クラス分析にかけ、各回答者が割当てられたクラスを「球場・球団評価」という１項目として扱った。OPCA の 6 次元解の第 2 と第 4 次元間の数量化得点と数量化主座標（各カテゴリーに割り当てられた数量）の関係を図示したのが図３である。

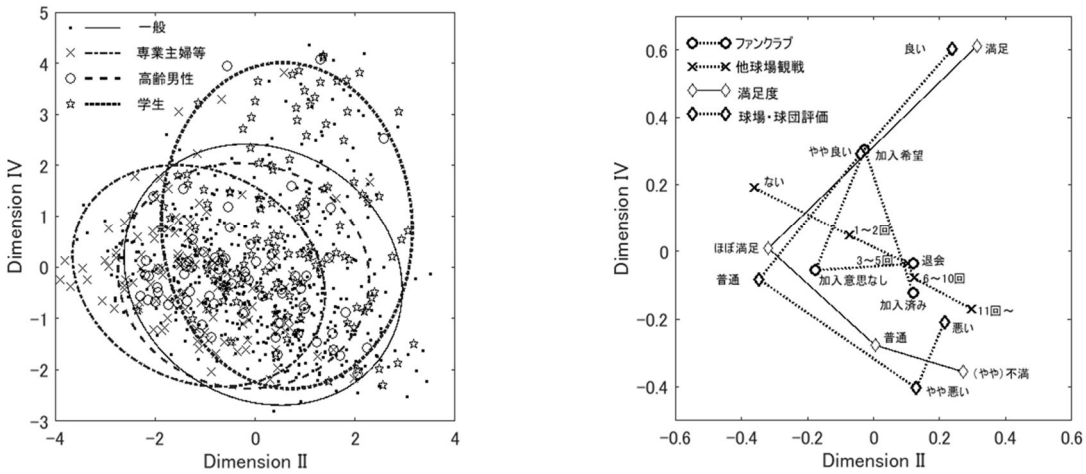

図４　次元 2 と 4 の間の散布図（左）とカテゴリーの主座標。散布図には 4 つの職業カテゴリーに属する個体の 87%集中楕円を描いた（村上,2022a）。

右の図で球場・球団評価と観戦満足度の主座標は、横向きの 2 次関数の形状を示し、左の散布図はほぼ横倒しの馬蹄の形をとっている。縦軸（次元４）は評価と満足度の高さを、横軸は、ほぼ観戦回数の１次関数である。観戦未経験者にとって、まだ明確な評価は定まっていないが、観戦経験が増えるほど、評価は高低 2 方向に分かれていくことが見てとれる。

③　好みにもとづく文化資本の階層差の分析

これはイギリスにおいて行われたかなり大規模な調査から、Le Roux & Rouanet (2010) が MCA の適用例として抜き出した４つの質問項目に対する 1215 名の回答者のデータである。４つの質問は、「テレビ番組」、「映画」、「絵画」、「外食」についてのそれぞれ６～８個のカテゴリーから１つ選ぶというものである。OPCA ３次元解の数量化得点の分布は、図４のようになった。

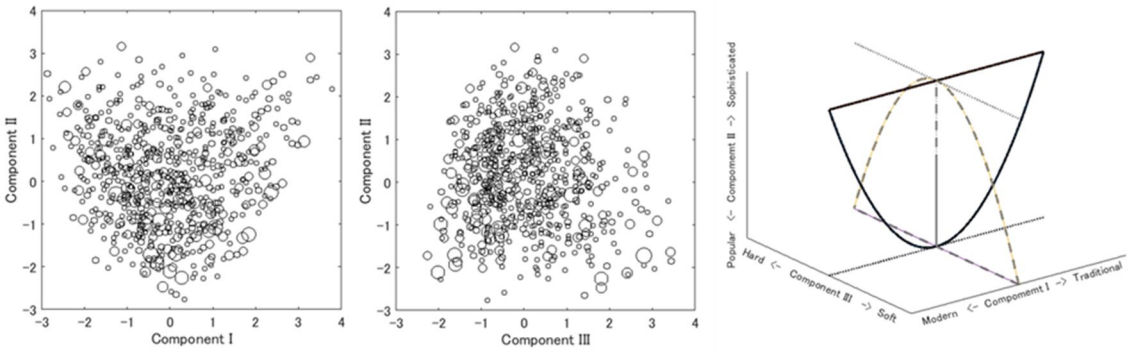

図４　イギリス人の好みに関する OPCA の 3 次元解。左は第 1 と第 2、中央は第 3 と第 2 次元の間の散布図。右は、3 次元の関係を模式的に表現したもの（村上、2022b）。

図４の３つのグラフにおいて、縦方向の第２次元は、文化資本の量を示すと考えられ、左の図の横軸である第１次元は、伝統対モダンの対立軸、中央の図の横軸である第３次元は、ソフト対ハードの対立軸と考えられた。そして伝統対モダンの対立は、文化資本の多い階層で、ソフト対ハードの対立は文化資本の少ない階層で主に生じていることは、それぞれの位置での分散の大きさが示している。すなわち、ここでの馬蹄は、次元１と次元３の分散が次元２によって増減していることを示している。つまり、馬蹄には明確な経験的な意味があることを示している。

④　心理尺度構成の基礎

現在、前述のような構造方程式モデルの台頭により、データのモデルへの適合度が必要以上に重視されるようになった結果、多数の項目を収集・実施して、得られたデータにもとづいて、適切な複数の（通常、反応カテゴリーの単純和、sum score）を構成するステップが、ていねいに行われず、特に我が国では海外で作成された尺度をそのまま翻訳して使用、論文化するという、あまり好ましくない状況が出来しているように思われる。また、独自の尺度が構成される場合にも、量的データへの最尤法による探索的因子分析（Exploratory Factor Analysis, EFA）の負荷行列から項目選択が行われることが多い。

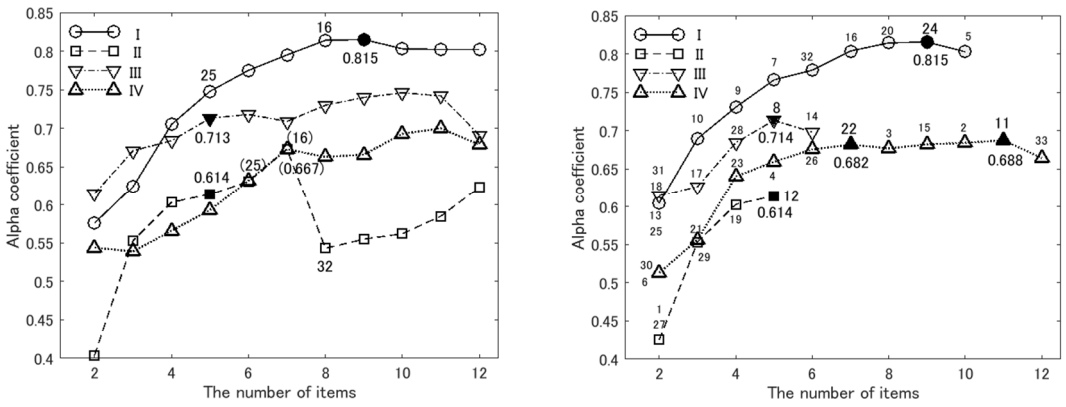

図５ EFA による項目選択（左）と PCCA による項目選択（右）の比較。EFA では、α係数の突然の下降が起きることがある。

Murakami(2024)は、(2)の③で述べた主クラスター主成分分析により、項目選択を行う手続きを、EFA を用いる方法と比較した。尺度の quality の評価基準は、sum score の内的整合性の指標であるα係数とし、ここではあえて機械的に、負荷量の高い順に項目を付け加えていくという手続きをとる。この経過を示したのが図５である。

データとしては、Mehrabian & Epstein (1972)による情動的共感尺度（Emotional Empathy Scale; EES）の日本語版を 1645 名の大学生に実施した結果である。この尺度は、100 名前後の大学生の反応にもとづいて、事前の因子分析等を一切実施しないまま、１次元の尺度として公刊されており、尺度構成の初期の段階の適切なモデルとなっていると考えられる。

EFA において通常用いられる斜交回転によって、負荷行列の要素は相関係数でなく、標準偏回帰係数となるため、他の項目群と逆相関である項目が比較的高い正の係数をもつことがある。この点が考慮されていないため、機械的に項目選択を行っていくと図５左に見られるように、１つの項目を加算したとき、α係数の突然の下降が起きることがある。PCCA の負荷量は基本的に sum score と項目との相関係数であるから、こうした現象は起こらない。

５．主な発表論文等

〔雑誌論文〕　計7件（うち査読付論文　2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 村上 隆	４．巻 49
２．論文標題 多重対応分析と正規直交主成分分析　―プロ野球の観客調査の分析―	５．発行年 2022年
３．雑誌名 行動計量学	６．最初と最後の頁 43-52
掲載論文のDOI（デジタルオブジェクト識別子） 10.2333/jbhmk.49.43	査読の有無 有
オープンアクセス 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 ―

１．著者名 村上 隆	４．巻 50
２．論文標題 多重対応分析における回転の効果　Le Roux & Rouanet のデータの再分析	５．発行年 2022年
３．雑誌名 日本行動計量学会第50回大会抄録集	６．最初と最後の頁 243～246
掲載論文のDOI（デジタルオブジェクト識別子） 10.20742/pbsj.50.0_243	査読の有無 無
オープンアクセス 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 ―

１．著者名 村上 隆	４．巻 49
２．論文標題 多重対応分析の因子分析的使用　社会調査データの記述的多変量解析	５．発行年 2021年
３．雑誌名 日本行動計量学会第49回大会抄録集	６．最初と最後の頁 4-7
掲載論文のDOI（デジタルオブジェクト識別子） 10.20742/pbsj.49.0_4	査読の有無 無
オープンアクセス 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 ―

１．著者名 村上　隆	４．巻 32
２．論文標題 中京大学スポーツミュージアムの来館者による博物館評価　ーカテゴリカルデータのための正規直交主成分分析によるデータの再分析ー	５．発行年 2021年
３．雑誌名 文化科学研究	６．最初と最後の頁 53-80
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス 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 ―

1．著者名 村上　隆	4．巻 1
2．論文標題 潜在プロフィール分析における前提条件が解に及ぼす効果	5．発行年 2020年
3．雑誌名 日本行動計量学会第48回大会抄録集	6．最初と最後の頁 222-225
掲載論文のDOI（デジタルオブジェクト識別子） 10.20742/pbsj.48.0_222	査読の有無 無
オープンアクセス 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 －

1．著者名 村上　隆・谷岡　謙・堀　兼大朗	4．巻 22
2．論文標題 大学博物館の来館者による評価	5．発行年 2021年
3．雑誌名 中大学文化科学研究所叢書（大学教育と博物館）	6．最初と最後の頁 163-198
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Takashi Murakami	4．巻 1
2．論文標題 Orhonormal principal component analysis for categorical data as a transformation of multiple correspondence analysis. In Tadashi Imaizumi, Atsushi Nakayama, & Satoru Yokoyama (eds.) Advanced study in behaviormetrics and data science Springer Nature Singapore, 211-231..	5．発行年 2020年
3．雑誌名 Advanced Studies in Behaviormetrics and Data Science, Springer	6．最初と最後の頁 211-231
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-981-15-2700-5_13	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計4件（うち招待講演　1件／うち国際学会　0件）

1．発表者名 村上　隆
2．発表標題 多重対応分析における回転の効果
3．学会等名 日本行動計量学会第50回大会
4．発表年 2022年

１．発表者名
村上 隆

２．発表標題
多重対応分析の因子分析的使用　社会調査データの記述的多変量解析

３．学会等名
日本行動計量学会（招待講演）

４．発表年
2021年

１．発表者名
村上隆

２．発表標題
潜在プロフィール分析における前提条件が解に及ぼす効果

３．学会等名
日本行動計量学会

４．発表年
2020年

１．発表者名
村上隆

２．発表標題
Independent cluster rotationに関する位置考察　斜交回転なのになぜ因子ごとの説明力が定義できるのか

３．学会等名
日本行動計量学会

４．発表年
2023年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

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0件

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