科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１２　日現在

機関番号：　３４３１５
研究種目：　基盤研究(C)
研究期間：　2011～2013
課題番号：　２３５３０６３０

研究課題名（和文）社会調査データの分析モデルにおけるランダム行列理論の応用

研究課題名（英文）Application of Random Matrix Theory to sociological data analysis

研究代表者
中井　美樹（NAKAI, MIKI）

立命館大学・産業社会学部・教授

研究者番号：００２４１２８２

交付決定額（研究期間全体）：（直接経費）　　2,900,000円、（間接経費）　　870,000円

研究成果の概要（和文）：　本研究は、近年、社会科学でも応用されてきつつあるランダム行列理論を社会調査データ分析に応用し、それによってより精緻な分析モデルに基づいた社会調査データ分析の新たな手法を提案することを目指すものである。従来は基づき自然科学領域で種々の現象のモデル化に応用されてきたこのアイデアの適用により、データの‘誤差’と‘情報（＝真の値）’の部分の識別が可能となることが示されてきた。そこで実際の社会学データ（SSM2005データおよびSSP2010データ）への適用を検討した結果、特に社会学データで大きな課題の1つである量的（連続）データとカテゴリカル・データの統一的扱いに有効なことが明らかとなった。

研究成果の概要（英文）：We have been applying the novel geometrical framework for computing variance-covariance matrices and linear correlation matrices for sets of heterogenous variables (meaning for system of variables that can be continuous or categorical), introduced first in our paper (forthcoming). We have tested the efficiency and limitations of the approach when applied to statistical analysis of sociological data. In particular, we are advancing a variance-covariance analysis of the 2005 Japanese national survey on social stratification and mobility (SSM2005). Such an analysis gives us the opportunity to apply our recent mathematical tools in conjunction with the application of Random Matrix Theory (RMT) for filtering the covariance matrix of the data, and eliminating the statistical noise due to the complex structure of the questionnaire, and to the finite size of the respondents. Moreover, we apply the analysis to the SSP2010 data and comment on a number of different RMT techniques and tools.

研究分野：　社会学

科研費の分科・細目：　社会学・社会学

キーワード：　社会調査データ　共分散行列　ランダム行列理論　カテゴリカル変数

１．研究開始当初の背景

社会学研究においては、社会調査データの分析を手段として用いた経験的研究に基づいて種々の社会現象の理解やメカニズムの解明がすすめられ、社会学理論を構築・展開するという営みが重要な部分を占めてきた。そこではしばしば社会調査データが用いられるが、社会調査データに特有の分析に伴う課題がいくつかある。たとえば、同時に多数の変数を分析対象としつつ調査環境等の制約から標本数が必ずしも多く得られないこと、とりわけ、回収率が低い場合には標本誤差が大きいこと、連続変数だけではなく社会科学ではしばしばカテゴリカル変数（名義尺度や順序尺度）によって多くの事象が測定されるためカテゴリカル・データと連続データを同時に分析する必要があること、無回答により欠損・欠測データ(不完全データ)が生起すること、などである。社会学研究ではこれまで長きにわたって、社会調査データに伴うこうした課題に対処するため新たな計量モデルが開発・応用され、あわせて調査設計に関する研究も進められてきた。そこでは統計解析手法等の開発と分析ツールやプログラムの開発が並行して取り組まれ、さらに社会学に隣接する社会諸科学領域での分析枠組みや自然科学の諸領域で有用とされる分析枠組みが社会学データの分析に様々な利用可能性を持つことが理解され、応用されてきた。

その一方で、近年、社会科学分野（金融工学、情報学、ネットワーク分析など）で応用研究が広がりつつあるランダム行列理論は、これまで社会学領域で社会調査データ、いわゆる標本調査の分析には適用されたことがない。ランダム行列理論はランダム行列、すなわち確率変数を要素に持つ行列を扱い種々の現象のモデル化を行う。

そこで本研究はランダム行列理論を適用する分析枠組みによる新たな手法を社会学における社会調査データ分析に応用し、それによってより精緻な分析モデルに基づいた社会調査データ分析の新たな手法を提案し、より的確で深い社会学的インプリケーションを得ることを目指した。なぜなら、社会調査データの分析に一般的に利用されている種々の多変量解析もその多くは共分散行列の固有値分析に基づいており、したがってこの新たな枠組みを応用することが可能であるからである。従来の研究からは、ランダム行列理論の適用によってデータの'誤差'と'情報（＝真の値）'の部分の識別が可能となることが示されてきた。また近年はカテゴリカル・データ分析に相関行列に基づく手法を応用した分析が提案されている。したがって、この枠組みの応用により社会学データで一般的に用いられる標本調査データに伴う誤差を除去するという課題への対処を可能とし、結果として従来の手法に基づいて得られた知見よりも確かな社会学的知見が導ける可能性があると考えられたからである。以上の研究背景から、これまで試みられたことのない社会調査データへのランダム行列理論の応用は取り組むに値する研究であり、社会学における計量的研究方法の向上に貢献できるとの考えから研究の必要性を実感した。

２．研究の目的

本研究では社会調査データ分析へのランダム行列の応用により、具体的には主として以下の２つの目的を追求する。第一に、ランダム行列理論をカテゴリカル・データ（変数）の分析へ拡張すること、第二に、社会調査データに伴う誤差の除去である。まず第一の点に関して、ランダム行列理論を社会学的データに応用する過程で取り組むべき課題として、カテゴリカル・データの分析への応用がある。社会科学や行動科学領域ではカテゴリカル・データとして得られるものが多く、すでに社会学研究法の中ではカテゴリカル･データの種々の分析手法が提案されているが、新たな枠組みの応用を拡張させカテゴリカル･データと連続（量的）データとを統一的に分析することが可能となれば、より有効なデータ分析に結びつけることが可能となり、その結果、社会事象の新たな社会学的考察が可能となる可能性がある。

第二の点に関して、推定における標本誤差が伴う社会調査データの分析にあたり、ランダム行列理論を適用することにより標本データの誤差を適切に除去し、限られたデータからより的確な知見が導けるならば、調査リソースを効率的に用いてデータを収集し、複雑化する現代社会の諸現象の理解と理論化を行う社会学における計量的研究手法の向上に貢献できると考えられる。

３．研究の方法

上記の研究目的を達成するための研究体制として､主として２つの研究班を組織した。これらは「A．社会学研究法における社会調査データの解析に伴う課題と、それに関連する手法の発展についての先行研究をレビューし検討する研究」､「B．ランダム行列理論の理論的研究や既存の先駆的応用研究をレビューし検討する研究」からなる。研究代表者はＡおよびＢに取り組み、研究協力者は主としてＢ班に参加して研究を進めた。また研究成果を発表し研究を深めるため、研究会を年に数回開催し集中的な議論を行い研究計画の効率的な遂行を行った。全体を通じて研究代表者が統括をつとめ、研究協力者との有益な連携を通じて全体テーマを学際的･総合的に推進した。

A．については、社会学データの分析モデルとその課題の理解を深めるため理論的研究や既存研究について検討を行った。必要に応じて海外の研究協力者との議論を行い、社会学データの分析モデルの基本的課題を整理し、現在までに開発されてきた分析手法と

その特徴や問題点を整理し、今後の課題について検討を行った。B．については、社会科学での応用が比較的新しいランダム行列理論について、まず理論的研究や既存の先駆的応用研究の概要について文献の整理を行い、報告と検討を行った。またそれを可能にする統計ソフトウェアについて知識・情報を収集し最近の研究動向について検討を行った。これらを並行して行うことにより、統計分析モデルの基礎的理解と発展的応用の素地ができ、学際的な研究の深まりが可能となる。

本研究ではすでに収集された社会調査データや、二次データに対してランダム行列モデルを応用したクラスタリング分析を適用し、ランダム行列理論の社会学的研究における実証的有効性を提示することとしていた。具体的には日本での代表的な社会調査データ「社会階層と社会移動全国調査（SSM 調査)」データおよび「階層と社会意識全国調査(SSP 調査)」データを主として用いることを念頭に置く。これらのデータ構造や変数の測定方法についてその特徴と課題を整理した。

４．研究成果

(1) 研究会の開催

本研究は、研究代表者が責任を持って研究を遂行することとあわせて、社会学分野における研究での社会調査データ分析にこれまで応用されることがなかったランダム行列理論について、経済物理学や生態学等での応用研究をレビュー・検討し応用するため、社会科学領域でのランダム行列モデルの応用研究に精通した海外の研究者の協力のもとで、研究会での研究報告や検討などを通じて得られた新たな多様な研究成果を蓄積しながら学際的研究を推進してきた。研究会を開催し最新の研究情報を得るとともに研究交流と今後の継続的な研究体制の構築を図った。研究会の開催を通じて以下のような課題を整理・検討した。

[平成 23 年度]
第１回研究会（H23.5） 社会調査データの解析手法およびデータ解析の課題の整理
第 2 回研究会（H23.6）社会調査データにおけるカテゴリカル・データの分析の整理
第 3 回研究会（H23.7）ランダム行列理論の応用(1)
第 3 回研究会（H23.8）ランダム行列理論の応用(2)
[平成 24 年度]
第 1 回研究会（H24.8）カテゴリカル・データ解析分析の応用、研究成果の発表
[平成 25 年度]
第 1 回研究会（H25.8）社会学的分析モデルと幾何学的分析モデルの検討(1)、研究成果のとりまとめと公表に向けた準備(1)
第 2 回研究会（H26.2）社会学的分析モデルと幾何学的分析モデルの検討(2)、研究成果のとりまとめと公表に向けた準備(2)

最終年度の研究発表とそれに基づく議論をうけ、さらに継続的に本研究を発展・深化させる必要性が研究代表者および研究協力者の間で認識されるに至った。

本研究期間を通じて達成した研究成果は多く、以下はその主要な成果である。

(2) カテゴリカル・データと連続(量的)データ間の分散共分散分析へのランダム行列理論の応用

第一の課題については、まず行列理論や線形代数によってカテゴリカル･データと連続(量的)データの間の分散共分散行列を定義することについて検討した。従来、連続(量的）データとカテゴリカル･データの相関や連関を表現するためにはそれぞれ固有の計量モデルや測度が考案されてきたが、それらを整理し検討した上で、従来の測度ではカバーされてこなかった、カテゴリカル－連続変数間の分散共分散行列を計量的に導出する統一的なフレームワークを定式化した。また、カテゴリカル変数間の共分散を議論した先行研究の検討をもとに、変数間の連関構造を多次元空間内における幾何学的オブジェクト（シンプレックス）によって表現する方法を考案した。

このアイデアに基づいて、実際の大規模社会調査データを分析するにあたって、数値計算パッケージである Scilab 言語を用いて共分散行列計算のためのアルゴリズムを構築した。従来の種々の連関指標の特徴をまず検討し、本研究より得られたデータの連関構造の情報との共通点と相違点を明らかにした。

実際の社会調査データとして用いた「社会階層と社会移動全国調査（SSM 調査)」は 800 以上の変数と 5743 人の回答からなるデータセットである。まず主要な変数を含んだ欠損値のないサブセット（変数 148、2215 名のデータ）を作成した。導出・提案した手法を用いてこのサブデータセットについて相関行列(実対称行列）を計算した。これを視覚化したのが図１である。

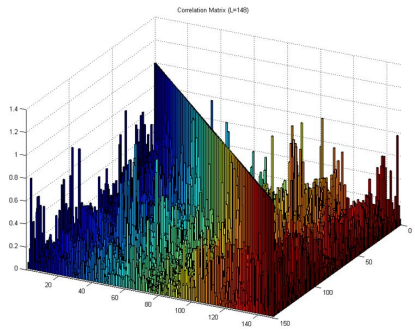

図１　相関行列

従来のさまざまな連関指標はそれぞれに固有の文脈や意味を持ち、その限りでは有用であるが、カテゴリカル･データと連続(量

的)データの間の連関構造を統一的な見地から評価することを困難にさせてきた。本研究で定式化した新たな指標により、カテゴリカル・データ間についても連続－カテゴリカル・データ間についても分散共分散の評価が可能になる。多くの多変量解析は共分散行列をもとに解析が進められるが、こうした多変量解析の基礎となる共分散行列の新たな定式化により既存研究での分析結果の再検討を可能にした。

この成果、すなわち拡張的な新たな手法の定式化と社会調査データ分析への応用については、英文論文としてまとめ海外の社会学ジャーナルに投稿し、掲載決定済みである。

(3) 社会調査データに伴う誤差の除去

相関行列の固有値(スペクトル)は種々の多変量解析(例えば、主成分分析、コレスポンデンス分析、多次元尺度構成法など）の基礎となる重要なものである。上記(2)で得た相関行列の固有値の分布は、図 2 に示すとおりである。

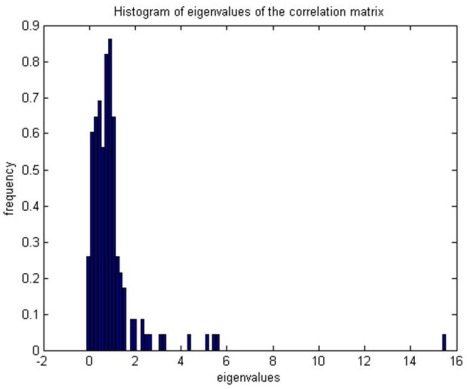

図２　社会調査データ(SSM2005)の相関行列の固有値のヒストグラム

ここで 0 に近い小さな固有値に対応する成分は統計的ノイズであるとの仮説を社会学的データに応用し、社会調査データの統計的ノイズはランダム行列理論により説明できることを示すことを試みた。

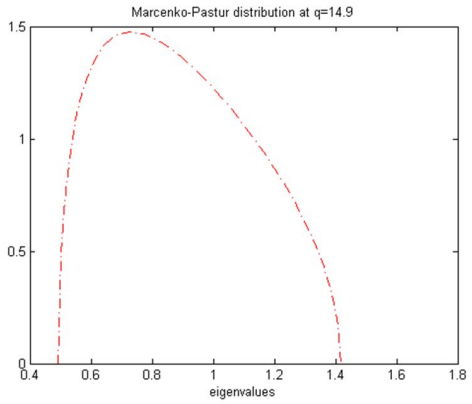

図３　Marcenko-Pastur 分布

図 3 は Marcenko-Pastur によって理論的に計算されたランダム行列の固有値が従う確率密度関数であり、図 2 と図 3 のズレに着目して相関行列のスペクトルから下限の部分のスペクトルを除去することにより、より意味のある固有値の部分のみをより詳細で深い分析に用いることができると考えられる。

この結果得られた相関行列に基づき主成分分析およびクラスター分析を行い詳細に検討すると、先行研究で示唆される主成分と相関の高い変数がいくつか得られていることが明らかとなった。このことから、ランダム行列理論を社会調査データに応用することは有効であると期待される。これまでの検討においてランダム行列理論を社会学的データに応用する際には最も単純な手法を適用している。今後、より詳細な分析や種々の解析手法に応用して比較検討を行う必要がある。

(4)社会調査データにともなう欠測(欠損）データの扱いの問題について

本研究を進める過程で、社会調査データ分析に本研究で導出した新たな手法を応用する際に、欠損値データの扱いが問題となる。社会調査データを扱う際に伴う欠損値データを考慮にいれて本手法を適用するためには、さらに慎重な検討が必要となる。こうした新たな課題に気づいた点も本研究の成果といえ、こうした課題も含めてさらに継続的に本研究を発展・深化させる必要性が研究代表者および研究協力者の間で認識されるに至った。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 5 件）

① Vernizzi, G. and Miki Nakai, “A Geometrical Framework for Covariance Matrices of Continuous and Categorical Variables,” *Sociological Methods & Research*, 査読有, 掲載決定済.

②Nakai, Miki, 2013, “Socio-economic and Gender Differences in Voluntary Participation in Japan,” Wolfgang Gaul, Andreas Geyer-Schulz, Yasumasa Baba, and Akinori Okada (Eds.) *German-Japanese Interchange of Data Analysis Results*, Springer, Heidelberg-Berlin: 225-234. DOI: 10.1007/978-3-319-01264-3_20. 査読有.

③ Nakai, Miki, 2011, “Social Stratification and Consumption Patterns: Cultural Practices and lifestyles in Japan,” S. Ingrassia, R. Rocci, M. Vichi (eds.) *New Perspectives in Statistical Modeling and Data Analysis*, Springer, Heidelberg-Berlin: 211-218. DOI: 10.1007/978-3-642-11363-5_24. 査読有.

〔学会発表〕（計 4 件）
① Nakai, Miki, 2013, “Patterns of Cultural Practices and Characteristics of the Cultural Omnivore,” IFCS2013, University of Tilburg, The Netherlands, July 14-17.
②Nakai, Miki, 2011, “Class and Gender Differences in Cultural Participation: Asymmetric Multidimensional Scaling of Cultural Consumption,” The 8th International Meeting of the Classification and Data Anaysis Group of the Italian Statistical Society (CLADAG), Universita degli Studi di Pavia, September 7-9.

〔図書〕（計 1 件）
①岡太 彬訓・中井 美樹・元治 恵子, 2012, 『データ分析入門―基礎統計―』共立出版. 全 176 頁.

６．研究組織
(1)研究代表者
中井　美樹（NAKAI, Miki）
立命館大学・産業社会学部・教授
研究者番号：　００２４１２８２

(2)研究分担者
なし

(3)連携研究者
なし

(4)研究協力者
Graziano Vernizzi
Siena College・Department of Physics and Astronomy・Associate Professor