様式　7（Form 7）

学　位　論　文　要　旨

Dissertation Abstract

学位請求論文題名 Dissertation Title

高次元ビックデータに対する統計的な特徴抽出のための次元縮約法に関する研究

（和訳または英訳） Japanese or English Translation

Study on dimensionality reduction method for High-Dimensional Big Data

人間社会環境学　専　攻（Division）

氏　　名（Name）原田魁成

主任指導教員氏名（Primary Supervisor）寒河江雅彦

（注）学位論文要旨の表紙

Note: This is the cover page of the dissertation abstract.

Abstract

In this paper, I study a statistical analysis method for extracting useful features from high-dimensional big data. When existing statistical models are applied to high-dimensional big data, the models become complex and analysis may not be easy. One of the analysis methods for these data is a dimension reduction method such as principal component analysis or non-negative matrix factorization, which can extract the common structure of the data, and a regularization model that prevents overlearning of the data. By using these models, it is possible to objectively extract useful features from large-sample, high-dimensional big data, as well as to prevent the complexity of the models and extract clear features. For example, by applying the dimension reduction method to an input-output table of regional economic transactions, I was able to identify regional basic industries and industrial clusters, and by applying it to cell phone location data, I was able to analyze changes in human flow before and after the COVID-19 epidemic. In addition, I found a possible selection criterion for the problem of determining the hyperparameters of regularized models, which has been an open problem in existing studies.

概要

本稿では高次元ビックデータから有益な特徴を抽出するための統計的解析法について述べている。高次元ビックデータに既存の統計モデルを適用した場合、モデルの次元数が増加するに伴い、解析が困難になる場合がある。これらのビッグデータに対する有益な構造を取り出せる次元縮約法とデータの過学習を防ぐ正則化モデルについて考える。

次元縮約法は、行列構造を「行」と「列」に分解して解釈できる。例えば、顧客の購買情報が記されたPOSデータに関して、行方向に顧客の属性、列方向に購買した商品コード等が記録されている場合、「顧客の属性」と「購入した商品」の2つの行列に分解され、これらを併せて「どの属性(年齢・性別等)がどのような商品(群)を購入する傾向があるか」という特徴が解析できる。また従来型の分析方法と異なり、明確な目的変数を設定しない解析手法である。したがって、特定の変数に狙いを絞って解析する従来型モデルよりも、データの構造を客観的に解析できる。

本稿では次元縮約法の、「主成分分析」や「非負値行列因子分解」に関する研究を行っている。主成分分析はデータの特徴を、主成分と呼ばれる低次元の空間にデータを写像し、その空間上でデータの持つ構造を解析する方法である。縮約する次元(主成分)を、データのばらつきが最大となるように計算するため、低次元でデータの特徴を捉える解析が可能である。さらに直交性の下で主成分が構築されることから、主成分同士で異なった特徴が抽出できる。一方、非負値行列因子分解は、基底行列と表現行列という、2つの低次元非負値行列に分解することでデータ行列の近似を行う方法である。分解された2つの行列の要素がすべて非負値であることから、正負で特徴が相殺されることなく、加法的な解釈が可能となる。特に画像データや購買データ等の非負値のデータ構造を分析するために望ましい解析手法である。

正則化モデルは、統計モデルに正則化項と呼ばれる制約条件を付与することで、解を安定化させる効果や、解の一部を０にする(この性質をスパース性と呼ぶ)ことでモデルの複雑化を防ぐ効果がある。特に解の安定化に寄与する正則化項を付与した回帰モデルをRidge回帰(L2正則化)、解の一部を０にできる正則化項を付与した回帰モデルをLasso回帰(L1正則化)と呼ぶ。本稿では、次元縮約法と正則化項を組み合わせ、次元縮約法の適用によって得られる結果をさらに簡素で、解釈しやすいものにするため、スパース性を持つLasso回帰型の正則化モデルに着目し、研究を行っている。

本稿における次元縮約法及び正則化モデルに関する研究目的は、大標本かつ高次元なビックデータから有益な特徴を客観的に抽出できること及びモデルの複雑化を防ぎ、複雑な構造から明瞭な特徴を抽出することである。データ解析における従来型のモデルは、分析したい特定の変数を、あらかじめ目的変数として設定した上で解析を行っている。例えばPOS(Points Of Sales)データのような、リアルタイムに購買情報が更新される時系列ビックデータを解析する場合、時系列分析でよく使用されるARIMAモデル(Auto Regressive Integrated Moving Average：自己回帰和分移動平均モデル)等の従来型の解析手法では、ある特定の商品や性別等のカテゴリーに絞って、売上の動向を解析することとなる。こうした分析手法では、特定の商品やカテゴリーに関して解析することが可能であるが、その一方でそれら以外の変数は同モデル内では考慮されない。すなわち同手法では、選択した商品やカテゴリーにおける変数間の関係性を見落とす可能性があり、その関係性が適切に解析されるか否かは、解析者の主観に依存する。他方、次元縮約法は、解析する目的変数を特定せず、全ての変数を含めた解析を行う手法である。そのため、複数の変数間に跨る特徴を抽出することが可能である。また、上記の例について、変数間の関係性を考慮するために、複数の変数を組み込んだ従来型のモデルを使用すると、モデルが複雑となる場合や変数間の組み合わせ的なモデルの

想定が必要となり、解釈が困難となる。例えばある特定の商品の売上を推定するため、他の多くの商品や気温等の外的情報を変数としたモデルを使用した場合、ある商品の売上予測の精度は上昇するが、その予測に多数の変数情報が含まれるため、有益な変数の特定が困難となる。この時、従来型のモデルにL1正則化項を付与する正則化モデルを適用することで、売上予測にあまり寄与しない特徴を０にすることができ、結果として予測に有益な特徴を際立たせることができる。本稿では次元縮約法に正則化モデルを組み込んだ新しいモデリングを開発し、実際のビックデータ解析に適用することで、提案したモデルの正当性を実証的に明らかにする。

本稿で示した研究分野に対する新規性として、縮約する次元数及び正則化項を制御するハイパーパラメータの決定法について1つの決定基準を実証的に示していることが挙げられる。本稿で使用したいずれの次元縮約法においても、縮約する次元数の決定法は未解決な問題である。また、ハイパーパラメータの決定法において、回帰モデルに対する決定法に関しては複数提案されている一方で、次元縮約法が属する行列分解モデルでは、有力な決定法は提案されていない。特に本稿では、回帰モデルにおける正則化項のハイパーパラメータの決定に使用されているCV法(Cross Validation法:交差検証法)のアイデアを用いて行列分解モデルに拡張し、さらにCV値を評価するためのテストデータの選定基準に、実験計画の分野で使用されるBIBD法(Balanced Incomplete Block Design:つり合い不完備ブロック計画)を用いることで、モデルの学習データに対する過学習を防ぎつつ、データとの誤差を小さくできる最適解が得られることを示した。この結果は次元縮約法における縮約する次元数の決定問題にも応用できる可能性があり、今後の次元縮約法を用いた解析全般に、大きく貢献できる可能性がある。

以下では、本研究で行った実証研究の概要を述べる。次元縮約法及び正則化モデルを適用するデータとして、経済学の分野において地域経済分析を行う上でよく使用される産業連関表と、人の移動行動をリアルタイムに記録した携帯電話の位置情報を用いて解析を行った。産業連関表は、一定期間に行われた財・サービスの産業間引を一つの行列に集約した統計表であり、対象年次の産業構造や産業部門間の相互関係など、マクロな経済活動を総合的に把握できる。この統計表を用いた分析の多くは、特定の産業における最終需要が変化したときの経済波及効果及び経済損失の推定や、特定の項目(産業、粗付加価値、最終需要項目等)における地域比較等の経済活動に関する分析である。本稿では特定の項目に絞らず、地域の基盤産業や産業クラスター等を次元縮約法を用いた客観的な分析を試みている。また、携帯電話の位置情報は、国勢調査やパーソントリップ調査のような既存統計資料よりもリアルタイム性と小地域区分での分析が可能な点が優れている。本稿で使用するモバイル空間統計では1時間単位24時間365日の最小500m×500mメッシュ規模の人口滞留データが記録されており、非常に詳細な人流まで解析できる。一方で、大標本かつ高次元データとなるため、次元縮約法を用いて人流のパターンを解析することを試みている。また、次元縮約法にL1正則化項を付与した正則化モデルによって、より解釈しやすい人流パターンの解析に成功した。

主成分分析を産業連関表に適用した研究では、全国の産業連関表に主成分分析を適用することで得られる主成分に、47都道府県の産業連関表データを射影することで、同一固有空間上で47都道府県の産業構造の類似度を測定する工夫を行った。産業連関分析において、こうした主成分分析の適用による地域比較の分析例は本研究が初めてである。また、2005年及び2011年の産業連関表に主成分分析を適用し、産業構造の経年変化を解析した研究では、対象とした福島県、東京都、石川県のいずれも2005年から2011年の間で地域内に大きな経済構造の変化があり、その特徴を主成分分析によって捉えることができた。特に福島県では東日本大震災における医療・交通の特需効果、東京都

では地価バブルによる不動産業の取引増加の特徴が解析された。

非負値行列因子分解を47都道府県の産業連関表データに適用した結果、愛知県を中心とした製造圏、東京都を中心とした大都市の第3次産業、神奈川県や千葉県等の製油所機能を持つ地域を中心とした鉱業の特徴が特徴空間上に抽出された。産業連関表に対する非負値行列因子分解の適用例は国内外含め研究例が極めて少ない。また、産業連関表には、需要増加に伴う地域経済全体への波及効果を示す逆行列係数表や、産業間の距離を計測した「APL(Average Propagation Length:平均波及世代数)」など、その他多数の観点からの解析が可能である。さらにいずれのデータにおいても、地域経済分析をより詳細に行う上で、県間の産業間取引を記録した地域間産業連関表や国際間産業連関表へと解析対象を拡大するにつれて大標本かつ高次元なビックデータとなるため、本分析手法の応用可能性がある。

スパース主成分分析を、金沢市近郊含む石川中央都市圏4市2町におけるメッシュ内滞在人口データに適用した結果、石川県庁や大学機関、兼六園等に関する「就業・就学・観光」の特徴や、大型ショッピングセンター、繁華街である片町等に関する「買い物・飲食等」の小地域における行動の特徴抽出に成功した。いずれの行動パターンも、2020年4月から5月にかけてCOVID-19流行に伴い、移動の自粛が要請された緊急事態宣言期に大きく人流が低下し、解除後に概ね回復したが、COVID-19流行前である2019年の水準までには回復していないことが示された。また通常の主成分分析と比較して、スパース制約の効果が明確に現れた。この分析において、正則化項のハイパーパラメータの決定にCV法とBIBD法を組み合わせたモデルを使用し、過学習を防ぎつつ、データとの誤差を小さくできる最適解を推定した。

スパース非負値行列分解をCOVID-19流行期前後における県間移動データ(携帯電話の位置情報)に適用した結果、COVID-19第1波流行期には、政府の自粛に関する呼びかけに従い、緊急事態宣言中の県間移動行動量は概ね0として抽出された。一方スパース効果によって設定した次元数内で抽出しきれなかった特徴は、元データとの残差項で解析することができ、COVID-19流行期の行動パターンは、大型台風が発生した場合の異質さと同等規模であった。携帯電話の位置情報データの解析にスパース非負値行列因子分解法を用いた分析例は本研究が初めてであり、正則化モデルを用いた解析がリアルタイムに蓄積される動的データに対しても有効であることを示した。

今後の展開として、縮約する次元数や正則化項のハイパーパラメータの選択に関する先行研究をまとめている。また、行列データを対象とした分析法であった次元縮約法を、高次元データのテンソル分解を用いた解析へと展開させ、今後さらに高次元なデータを解析するための予備的な研究を、中部圏地域間産業連関データに対する適用例と併せて議論している。

学位論文審査報告書

2022 年　2 月　1 日

１　論文提出者

金沢大学大学院人間社会環境研究科

専　攻　　人間社会環境学

氏　名　　原田　魁成

２　学位論文題目（外国語の場合は，和訳を付記すること。）

高次元ビックデータに対する統計的な特徴抽出のための次元縮約法に関する研究

※　提出者の実績

・査読付き論文 計 8 編

・研究報告 計 16 件(いずれも提出者が報告したもの)

・受賞歴　応用統計学会 2021 年度年会　優秀発表賞

３　審査結果

判　　定（いずれかに〇印）　（合 格）・　不合格

授与学位（いずれかに〇印）　博士（ 社会環境学・文学・法学・（経済学）・学術 ）

４　学位論文審査委員

委員長　　寒河江　雅彦　　㊞

委　員　　佐藤　清和

委　員　　佐無田　光

委　員　　柳　在圭

委　員　　藤生　慎

委　員

（学位論文審査委員全員の審査により判定した。）

５　論文審査の結果の要旨

本論文ではビックデータから有益な特徴を抽出するための統計的な解析手法に関する研究について述べている。既存の高次元ビックデータ解析の多くは、解析対象となる目的変数を設定した上で解析が行われる。その場合、設定した変数以外の項目との関係性が見落とされる可能性がある。また、変数間の関係性を考慮する場合、複数の変数を組み込んだモデルを使用することになり、モデルの複雑性に起因した解釈の困難性を生じることとなる。こうしたビックデータを解析する上で生じる課題に対し、解析する目的変数を特定せず、全ての変数を含めた解析を行う「次元縮約法」に加えて、解析に大きく寄与しない変数の影響を取り除くことでモデルの複雑化を防ぐ「正則化モデリング」について研究し、実証研究を通して、有効性を示している。

次元縮約法については主成分分析と非負値行列因子分解を研究対象とした。主成分分析は低次元下でも元データの情報を最も効率的に次元圧縮ができ、さらに直交分解された特徴(主成分)ごとに解釈が可能である。非負値行列因子分解は、正値を要素に持つ 2 つの行列に分解するため、正値のみで特徴を解釈できる利点がある。

正則化モデルについては L1 正則化項を制約条件に課す正則化モデル(Lasso)を研究対象とした。例えば、このモデルは回帰分析における回帰係数の一部をハイパーパラメータの調整により０にできるスパース効果があるため、モデルの複雑化を防ぐことが出来る。本稿では主成分分析と非負値行列因子分解に L1 正則化項を付与した正則化モデルとして、スパース主成分分析、スパース非負値行列分解に関して研究を行った。

いずれの解析手法においても、データの変数間で共通する特徴を客観的に抽出することに成功した。また、通常の次元縮約法と正則化項を付与した次元縮約法では、正則化モデルの方が特徴の強弱が際立ち、明瞭な解釈が可能となった。具体的な実証研究では、財・サービスの取引を 1 つの行列に集約した産業連関表と、携帯電話の位置情報データに基づく 500m×500m メッシュ内滞留人口データを使用している。実施法分析の主要な結果を以下に述べる。

主成分分析を全国の産業連関表に適用した研究では、47 都道府県の産業連関表データを射影し、同一固有空間上で 47 都道府県の産業構造の類似度を測定した。また 2005 年及び 2011 年の産業連関表に主成分分析を適用し、産業構造の経年変化を解析した研究では、対象とした福島県、東京都、石川県のいずれも 2005 年から 2011 年の間で地域内に大きな経済活動の変化があり、その影響があった産業の特徴を主成分分析によって捉えることができた。

非負値行列因子分解を同じく 47 都道府県の産業連関表データに適用した結果、愛知県を中心とした製造圏、東京都を中心とした大都市の第 3 次産業、神奈川県や千葉県等の製油所機能を持つ地域を中心とした鉱業の特徴が特徴空間上に抽出された。

スパース主成分分析を金沢市近郊含む石川中央都市圏 4 市 2 町におけるメッシュ内滞在人口データに適用した結果、石川県庁や大学機関、兼六園等に関する「就業・就学・観光」の特徴や、大型ショッピングセンター、繁華街である片町等に関する「買い物・飲食等」の特徴が抽出された。また通常の主成分分析と比較して、スパース制約の効果が明確化された。

スパース非負値行列分解を COVID-19 流行前後における県間移動データ(携帯電話の位置情報)に適用した結果、COVID-19 第 1 波流行期には、政府の自粛に関する呼びかけに従い、緊急事態宣言中の県間移動行動量はほとんどなかったことが明らかになった。一方スパース制約による特徴は、元データとの残差項で解析することができ、COVID-19 流行期の行動パターンは、大型台風が発生した場合の計画運休等による移動行動の制限と同等規模であった。

本研究の新規性として、縮約する次元数及び正則化項を制御するハイパーパラメータの決定法について 1 つの決定基準を実証的に示していることが挙げられる。本稿で使用したいずれの次元縮約法においても、縮約する有効な次元数の決定法は未解決な問題である。また、ハイパーパラメータの決定法において、次元縮約法が属する行列分解モデルでは、有力な決定法は提案されていない。本稿では、CV 法(Cross Validation 法:交差検証法)のアイデアを用いて行列分解モデルに拡張し、さらに CV 値を評価するためのテストデータの選定基準に、実験計画の分野で使用されるつり合い不完備ブロック計画（BIBD : Balanced Incomplete Block Design）を用いることで、モデルの学習データに対する過学習を防ぎつつ、データとの誤差を小さくできる最適解が得られることを示した。この結果は次元縮約法における縮約する次元数の決定問題にも応用でき、様々な次元縮約法全般に大きく貢献できる可能性がある。

今後の展開として、本稿では行列データに対する行列因子分解モデルを基に、次元縮約法及び正則化モデルに関して議論したが、これらの解析手法は n 次元テンソルデータに対しても拡張可能である。すなわち、さらなる高次元データの解析のために、「テンソル分解」を用いた解析へと拡張できる。本稿では予備的に、中部圏地域間産業連関データを 3 次元テンソルデータへと再構築し、テンソル分解を用いて特徴抽出を試みた実証例を記している。

本審査論文は、4 編の査読付き論文誌に採択と現在進行中の研究成果（投稿準備中、学会発表済）で構成されている。応用統計学会 2021 年度年会では優秀発表賞を受賞し、高い評価を得ている。経済分析への新しいアプローチを導入し、方法論の開発と実証分析の両面から期待される若手研究者である。博士後期課程の 3 年間に学会発表、シンポジウム等での研究発表も 16 件を越え、活発な研究活動に邁進している。また、論文審査、審査委員との質疑でも的確な返答をしている。以上の研究業績、審査結果を踏まえ、博士（経済学）を授与することを認めます。