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

平成　２９　年　　６　月　１２　日現在

機関番号：　３５３０２
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３００５２

研究課題名（和文）大規模データの発見的特徴把握のための情報縮約・クラスタリング融合手法の研究

研究課題名（英文）A study of joint dimension reduction and clustering for heuristic considerations of large-scaled data

研究代表者

森　裕一（Mori, Yuichi）

岡山理科大学・総合情報学部・教授

研究者番号：８０２３００８５

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：大規模データを念頭におき，データの特徴を把握するための情報縮約とクラスタリングおよびその計算手法について検討することを目的に，先行研究等の情報収集と分析・整理，次元縮約手法（変数選択手法を含む）の検討，あらゆる尺度を考慮した情報縮約とクラスタリングの開発，対話性の導入，計算効率の検討を順に行った。その結果，主成分分析の文脈で，複雑さや尺度混在に対応できる手法の開発・整理ができ，対話的なインタフェースの構築や加速化アルゴリズムの適用により，データサイズや処理回数が膨大となる場面でも情報縮約とクラスタリングにおいて発見的・試行錯誤的な考察が可能な環境を提供できるようになった。

研究成果の概要（英文）：The purpose of this study is to propose joint dimension reduction and clustering to deeply consider the features of data, particularly large-scales data. To do this, the followings are performed: gathering information in previous studies, reviewing dimension reduction (including variable selection), proposing dimension reduction and clustering dealing with mixed measurement level data, developing interactive interface, and improving computational efficiency. The study is successful to provide some methods that handle the complexity including mixture of measurement level in the context of principal components analysis, and that consider the features of data by heuristic thinking and trial and error using developed interactive interfaces and proposed acceleration algorithms in case where a huge number of computation is necessary such as large-scaled data processing and variable selection.

研究分野：　計算機統計学

キーワード：　多変量解析　次元縮約　非計量主成分分析　非計量因子分析　クラスター分析　加速化アルゴリズム　対話的可視化ツール

１．研究開始当初の背景

マーケティング，環境，Web，ゲノムなど，多くの分野において，いわゆるビッグデータに対して，大規模であることを活かした効率的なデータの特徴把握が求められていた。

これらの分析には，次元の縮約（および重要な変数の選択）とクラスタリングが有効で，これまでも大規模データに適用されてきているが，次元を縮約することと観測対象や調査観点を分類するにあたって，無駄な次元を保持したまま分析していたり，一方の手法の特長を他方で活かせていないといった面が見られていた。また，データそのものも複雑な構造をもつようになっており，観測対象がさまざまな属性でグループ化されていたり，観測尺度も量的変数と質的変数が混在したりしていて，従来の手法では，その複雑性に対応できなくなっていた。これらに加え，データが大きいために，通常のマシン・システムでは一定時間内に処理できず，そのために，試行錯誤的なインタフェースの提供も難しいという状況も問題となっていた。特に，計算時間については，マシンパワーだけに頼ることない計算時間短縮の工夫が求められていた。

これらの問題点のうち，次元縮約手法と分類手法を同時に考えていく手法については，質的データや 2 値データを対象としてではあるが，いくつかの分析方法が提案されてきており，データの複雑さに対応した次元縮約手法については，主成分分析における変数選択や，Sparse PCA などがあった。しかしながら，既存の融合手法は質的データのみを対象としていること，その中の次元縮約においては，上に示したような複雑さに対応した新しいアプローチはまだ反映されていないこと，次元縮約では量的データが主たる対象であること，計算時間の問題もあって，対話的な処理ができるインタフェースも整備されていないというのが現状であった。

２．研究の目的

大規模データを念頭におき，データの特徴を把握するための情報縮約とクラスタリングおよびその計算手法について検討することが本研究の目的である。そこでは，データがもつ複雑さを考慮し，あらゆる尺度に対応できるように，手法を開発・整理していくことと，データサイズが膨大，あるいは変数選択のように処理回数が膨大な場面において，処理を効率的に行うための計算環境を提案してくことである。

具体的には，主成分分析の文脈で，データの複雑さが取り扱えること，あらゆる尺度が混在したデータが処理できること，対話的なインタフェースを採用し，発見的な情報の表現と把握ができる環境を提供すること，計算アルゴリズムや分散処理システムの工夫により，高速な計算環境を提供するといったことに焦点をあて研究する。

これらを通して，大規模データを対象に，データの複雑さを考慮し，あらゆる尺度のデータを処理できる情報縮約・クラスタリングを核とした手法を提案するとともに，発見的情報表現が可能なインタフェースと高速な計算環境の提供をめざすものである。

３．研究の方法

研究は，次の 5 つに分けて行った。

(1) 先行研究等の情報収集と分析・整理

情報縮約とクラスタリングの融合手法の先行研究や公開されているソフトウェアなどを入手し，分析・整理する。集めた手法については，その性能を評価し，処理可能サイズと処理の関係および現在のソフトウェアの適用範囲について，その現状と改善点を明らかにする。また，実用場面や事例を収集し，そこに求められている機能を整理する。

(2) 次元縮約手法（変数選択手法を含む）の検討

複雑な構造をもつデータの情報縮約について最新の研究成果を集め，トレンドの技法を明らかにするとともに，どの程度まで複雑さが扱えているかを整理する。また，パラメータや有効な変数の数の推定について，発見的探索手法の利用で実現が可能かを明らかにする。

(3) あらゆる尺度を考慮した情報縮約とクラスタリングの研究

尺度混在データの扱いについて，既存の質的データに対する主成分分析（非計量主成分分析）の数量化の手法を利用し，連続量に変換した後に，情報縮約やクラスタリングに臨む手順を検討する。

(4) 対話性の導入

発見的，試行錯誤的な考察ができるように，元データの操作，データグループの選択，パラメータの値指定，変数や情報の取捨選択，結果出力後の表示指定，結果らのクラスターの拾い上げなどに，対話的操作を導入する。また，それに適したプラットフォームを選び，システムを構築する。

(5) 計算効率の検討

加速化を含めた計算を効率的に行う計算アルゴリズムを検討する。また，Apache Hadoop を用いた分散処理システムでの運用を検討し，データの大きさや反復計算の頻出による膨大な計算量の軽減を図る。

４．研究成果

３の(1)(2)は先行研究の整理であるので，(3)～(5)について，以下，(1)～(3)として報告する。

(1) あらゆる尺度を考慮した情報縮約とクラスタリングの研究

①主成分分析における質的データの扱い

PRINCIPALS や homals をはじめとする交互最小二乗法（ALS）による最適尺度法など，主成分分析における数量化には先行研究があり，また，われわれの過去の研究成果から，

通常の主成分分析（PCA）と，変数の一部から主成分を抽出する拡張主成分分析（M.PCA）で質的データを扱えるようにした（非計量 PCA，非計量 M.PCA）。

②因子分析における質的データの扱い

因子分析においても ALS を利用した FACTALS があり，これを用いて，われわれの環境下において質的データが因子分析（FA）を扱えるようにした（非計量 FA）。

③尺度混在のデータの分析

①や②によって，尺度混在のデータ行列は，目的とする手法に適した形で連続量のデータ行列に変換されることになるので，元のデータの様相が量的データとして統一的に扱えるようになる。すなわち，旧来の連続量に対して開発されてきた分析手法がそのまま適用できるようになったわけである。

主成分分析の場合について，結果を示す。図 1 は 19 変数からなる量的データに PCA を適用して得られたバイプロット（主成分数 2）である。このデータの 1 変数を名義尺度に，8 変数を順序尺度に変換した人工データを作成し，これに尺度混在の PCA を適用し，バイプロットを描いたものが図 2 である。提案の非計量 PCA は，元のデータの様相を大きく損なうことなく，推定が行えており，4 つのクラスターも保持できていることがわかる。

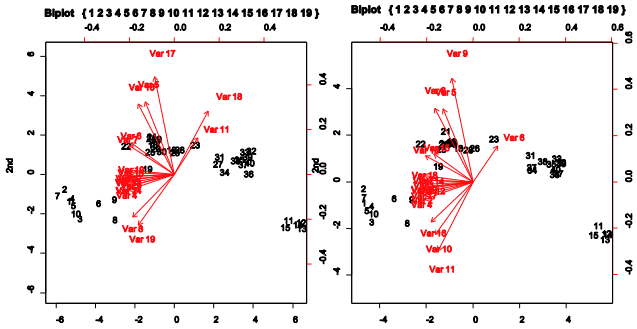

図 1 元データのバイプロット

図 2 尺度混在データのバイプロット

④尺度混在データの変数選択

連続量として等質な変数群として扱えるので，変数選択も統一規準で行うことができるようになった。

主成分分析の文脈では，M.PCA がその計算過程に変数選択手順が含まれるので，M.PCA を変数の数ごとに実行すれば，自然と変数選択となる。したがって，非計量 M.PCA により変数選択を行えばよいが，数量化と変数選択をどのタイミングで行うかで，3 つのタイプが考えられる（図 3）。すなわち，全変数を用いて数量化したものを 1 つの量的データとして変数選択を行うか（*Type* 1），変数の数に対応した最適な変数群が見つかるごとに数量化を行うか（*Type* 2），あるいは，最適な変数群を見つけるために，テンポラリな 1 変数を削除または追加するごとに数量化を行うか（*Type* 3）である。

③と同じ人工データに対して *Type* 3 による変数減増法を適用した結果が図 4 である。

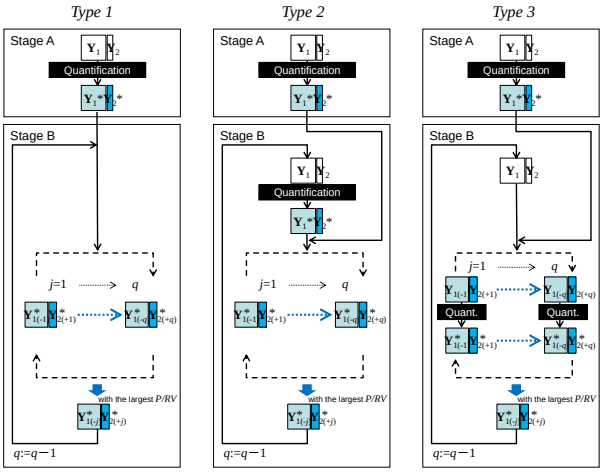

図 3 M.PCA の選択過程（数量化のタイミングによる 2 タイプ）

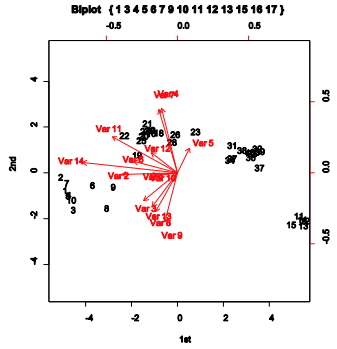

図 4 尺度混在データのバイプロット（M.PCA により選ばれた 14 変数）

⑤次元縮約とクラスタリングの同時推定

これについては，既存の手法を整理することにとどまった。GROUPALS，多重コレスポンデンス分析と K-means 法を同時に適用する方法，2 値データの次元縮約とクラスタリングを組合せた手法などである。これらに，③の手順を適用することで，尺度混在データへの同時推定が拡張できることを確認した。

(4) 対話性の導入

計算エンジンとして R を用いることが多いことから，R パッケージである Shiny を利用し，データ，パラメータ，出力の表示範囲などを対話的に指定するインタフェースの作成を試みた。図 5 は項目反応理論の一連の分析過程を対話的に進めるための Shiny アプリケーションである。ラジオボタン，チェックボックス，スライダーなどでパラメータや出力が制御できるので，考察の手段が増え，発見的・試行錯誤的に問題解決できる環境が実現できることがわかった。

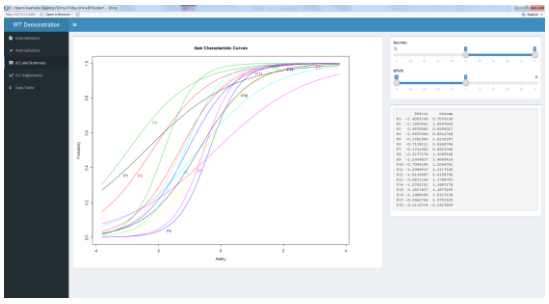

図 5 Shiny アプリケーションによるインタフェース

(5) 計算効率の検討

ハードウェア的な計算時間の短縮は，Apache Hadoop を用いた分散処理システムの運用を模索した。また，収束計算をともなう計算の加速化のアルゴリズムは，これまでの研究で定式化できているので，ここでは，(3)の各手法（非計量 PCA，非計量 FA，非計量 M.PCA）への実装と評価，および，より速いアルゴリズムとしての re-start 法を提案・考察することが主となる。前者では，大きなデータサイズでの実行と計算の繰り返しが膨大となる変数選択場面への適用で，いずれの状況においても，2 倍から 5 倍の速度向上が見られ，後者では，さらに 1.3 倍程度の加速の効果が見られることがわかった。

以上より，次元縮約の文脈で，尺度混在のデータが扱えるようになったことより，分類やクラスタリングでも同時に各種の尺度を扱えるようになり，それらの実行や結果の考察が対話的に行える環境が提供できるようになった。また，加速化アルゴリズムを中心に効率的に計算を行う手だてが提案できた。これらの成果は，2 つの図書にまとめている。なお，(3)で示したような計算方法にいくつかのパターンが出てくること，解の推定には収束判定をともなうことから初期値選択の問題と収束判定による精度の問題などが明確になり，これらは今後の課題に位置付けられる。

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 3 件）

①Kuroda, M., Geng, Z., Sakakihara, M. (2015). Improving the vector epsilon acceleration for the EM algorithm using a re-starting procedure. Computational Statistics, 30, 1051 – 1077［査読有］

②黒田正博 (2014). 相関・連関分析から因果分析へ. 数学セミナー増刊 統計学ガイダンス，日本評論社, 120 - 125［査読無］

③Mori, M., Kuroda, M., Izuka, M., Sakakihara, Y. (2014). Performance of acceleration of ALS algorithm in nonlinear PCA. Proceedings of COMPSTAT 2014 (ISBN 978-2-8399-1347-8, CD-ROM), 257-263［査読有］

〔学会発表〕（19 件）

①Kuroda, M., Mori, Y., Iizuka, M., Sakakihara, M. (2016). Acceleration of convergence of the alternating least squares algorithm for mixed measurement level multivariate data, The 10th ICSA International Conference, 2016 年 12 月 19 日～22 日,.上海 (中華人民共和国).

② Mori, Y., Kuorda, M., Iizuka, M. (2016). Variable selection in nonlinear principal component analysis, 2016 IASC-ARS Interim Conference (with KSS Fall Conference), 2016 年 11 月 4 日～5 日，大田 (韓国).

③朝原広喬，片山浩子，水谷直樹，森　裕一 (2016). shiny アプリを用いた対話的統計分析．日本行動計量学会第 44 回大会，2016 年 8 月 30 日～9 月 2 日，札幌学院大学 (北海道江別市).

④ Katayama, H., Asahara, H., Mizutani, N., Kuroda, M., Mori, Y. (2016). Interactive Statistical Analysis for Item Response Theory using Shiny. 2016 International Conference for JSCS 30th Anniversary, 2016 年 10 月 16 日～17 日，シアトル (アメリカ合衆国).

⑤Kubota, T. (2016). Visualizing and Exploratory Data Analysis for Small Area Suicide Data, 2016 International Conference for JSCS 30th Anniversary, 2016 年 10 月 16 日～17 日，シアトル (アメリカ合衆国).

⑥Sakakihara M., Kuroda, M., Mori, Y., Iizuka, M. (2016). Acceleration of iterative methods for Nonnegative Matrix Factorization, 2016 年 8 月 23 日～26 日，オビエド (スペイン).

⑦Kuroda, M. (2016). Fast estimation using the EM algorithm for Gaussian mixture models. The 4th Institute of Mathematical Statistics Asia Pacific Rim Meeting, 2016 年 6 月 27 日～6 月 30 日，香港 (香港).

⑧黒田正博，森　裕一，飯塚誠也 (2016). リスタートを用いた加速化交互最小二乗法による非計量主成分分析の変数選択法について．日本計算機統計学会第 30 回大会，2016 年 5 月 19 日～20 日，ハートピア京都 (京都市).

⑨Mori, Y., Kuroda, M., Sakakihara, M., Iizuka, M. (2015). Acceleration of the alternating least squares algorithm for nonlinear multivariate analyses. The 9th Conference of the Asian Regional Section of the IASC, 2015年12月17日～19 日，シンガポール (シンガポール).

⑩Kuroda M. (2015). Initial value selection of the EM algorithm for Gaussian mixture models, The 9th Conference of the Asian Regional Section of the IASC, 2015 年 12 月 17 日～19 日，シンガポール (シンガポール).

⑪森　裕一，黒田正博，榊原道夫，飯塚誠也（2015）．尺度混在データの多変量手法．日本計算機統計学会第 29 回シンポジウム，2015 年 11 月 27 日～28 日，まなぼっと幣舞 (北海道釧路市).

⑫黒田正博 (2015). EM アルゴリズムの加速化と応用．日本行動計量学会第 43 回大会，2015 年 9 月 1 日～9 月 4 日，首都大学東京 (東京都八王子市).

⑬Sakakihara, M., Kuroda, M., Mori, Y., Iizuka, M. (2015). On acceleration methods for Alternating Least Squares algorithm. IFCS2015, 2015 年 7 月 6 日～9 日，ボローニャ (イタリア).

⑭松田航平，飯塚誠也，黒田正博，森　裕一 (2015). 非計量因子分析と計算の加速化．日本計算機統計学会第 29 回大会，2015 年 5 月 14 日～15 日，山梨県立図書館 (山梨県甲府市).

⑮黒田正博 (2015). 正規混合モデルの EM アルゴリズムの初期値選択と加速. 日本計算機統計学会第 29 回大会, 2015 年 5 月 14 日～15 日, 山梨県立図書館 (山梨県甲府市).
⑯Kuroda M. (2014). An initial value selection method of the EM algorithm for mixture models. Kyoto International Conference on Modern Statistics in the 21st Century, 2014 年 11 月 17 日～11 月 18 日, 京都国際会議場 (京都市).
⑰森　裕一, 黒田正博, 飯塚誠也, 榊原道夫（2014）. 最小交互二乗法の加速化. 日本行動計量学会第 42 回大会, 2014 年 9 月 3 日～5 日, 東北大学 (宮城県仙台市).

〔図書〕（計 2 件）

①Mori. Y., Kuroda, M., Makino, N. (2017). Nonlinear Principal Component Analysis and Its Applications. JSS Research Series in Statistics, Springer, 2017 年発行, 80 ページ.
②森　裕一, 黒田正博, 足立浩平（2017 年）. 最小二乗法・交互最小二乗法, 統計ワンポイントシリーズ 3, 共立出版, 2017 年発行, 112 ページ.

〔その他〕

ホームページ等

http://mo161.soci.ous.ac.jp/vasmm/

６．研究組織

(1) 研究代表者

森　裕一（Mori, Yuichi）
岡山理科大学・総合情報学部・教授
研究者番号：80230085

(2) 研究分担者

飯塚　誠也（Iizuka, Masaya）
岡山大学・アドミッションセンター・教授
研究者番号：60322236

黒田　正博（Kuroda, Masahiro）
岡山理科大学・総合情報学部・教授
研究者番号：90279042

水谷　直樹（Mizutani, Naoki）
岡山理科大学・総合情報学部・准教授
研究者番号：30330533

久保　田貴文 Kubota, Takafumi）
多摩大学・経営情報学部・准教授
研究者番号：30379705

(3) 連携研究者

足立　浩平（Adachi, Kohei）
大阪大学・人間科学研究科・教授
研究者番号：60299055

中野　純司（Nakano, Junji）
統計数理研究所・統計計算開発センター・教授
研究者番号：60136281