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　２ 年　 ６ 月 ２９ 日現在

機関番号：３５３０２
研究種目：基盤研究(C)（一般）
研究期間：2016～2019
課題番号：１６Ｋ００６１

研究課題名（和文）EMアルゴリズムによる混合モデルの初期値選択法の提案とその加速に関する研究

研究課題名（英文）Initial value selection and acceleration of the EM algorithm for finite mixture models

研究代表者
黒田　正博（Kuroda, Masahiro）

岡山理科大学・経営学部・教授

研究者番号：９０２７９０４２

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：EMアルゴリズムは，局所的収束と線形収束をもつ最尤推定のための統計計算法である．このアルゴリズムを混合モデルに適用する場合，推定値が局所最大解である可能性と，その値を得るまでの膨大な反復回数が必要になる．そこで，これら2つの問題を解決するために，大域的最大解を求めるための初期値選択と，EMアルゴリズムの加速法を組み合わせたアルゴリズムの開発を目指した．さらに，パラメータ推測のためのブートストラップ法の計算を加速法を開発した．

研究成果の学術的意義や社会的意義
画像解析と機械学習などにおいて用いられる大規模データに対して統計モデルを考えるとき，混合モデルを仮定することになる．このモデルのもとでのデータ解析のための数値計算では，短い計算時間で最適な推定値を得ることが求められる．本研究では，統計計算法に用いられるEMアルゴリズムに焦点を当て，計算時間の短縮を図る加速法の開発と，最良な初期値を見つける初期値選択法の開発により，この問題の解決を図る．この研究の応用場面としては，医療診断における画像解析や機械学習等の分野のデータ解析であり，実用価値が高い研究である．

研究成果の概要（英文）：The EM algorithm is a general and popular algorithm for finding maximum likelihood estimates from incomplete data due to stability in convergence, simplicity in implementation and applicability in practice, while the algorithm only guarantees local and linear convergence. They are the drawbacks when the EM algorithm is applied to finite mixture models. We tried to develop an initial value selection method to select a suitable initial value such that the EM algorithm starting from the selected value can find an estimate maximizing globally the likelihood function. In order to reduce the total computation time and the number of iterations, we developed an algorithm that accelerates the convergence of the EM algorithm. Moreover, we showed an algorithm to improve the speed of computation of the bootstrap method using the EM algorithm.

研究分野：計算機統計学

キーワード：EMアルゴリズム　加速　初期値選択　混合モデル

様　式　C－１９、F－１９－１、Z－１９（共通）

1. 研究開始当初の背景

混合正規モデルは, 画像解析や機械学習等の統計モデルとして用いられ, EM アルゴリズムによるパラメータの最尤推定が行われる. その理由として, EM アルゴリズムの収束の安定性, プログラミングの容易性等の利点が考えられるが, 収束の遅さが問題である. 特に, 解析データが大規模で混合正規モデルの成分数が多い場合, その問題は顕著になり, EM アルゴリズムの加速は重要な課題である. そこで,

我々は, EM アルゴリズムの収束を加速する ε -accelerated EM アルゴリズムおよびその収束スピードを改善した ε R-accelerated EM アルゴリズムを開発した. 我々の提案したこれらの加速法は, EM アルゴリズムの優れた収束性を失うことなく, 収束の加速を実現しており, 混合モデルに対応した改良を ε R-accelerated EM アルゴリズムに加えることで, 収束スピードの問題の解決を図ることが可能であると考える. また, 混合正規モデルの尤度関数は多峰型 (multimodal) であり, 複数の局所的な最大値が存在する. EM アルゴリズムは局所的収束しか保障しておらず, 最尤推定値は初期値に依存する. これは, 混合正規モデルのパラメータ推定において本質的な問題であり, EM アルゴリズムの初期値選択は極めて重要である. 一般的には, 乱数等により生成された複数の初期値から EM アルゴリズムを実行し, その中で尤度関数値が最大であるパラメータ推定値を初期値に設定する. この方法は, Biernacki et.al, (2003)による emEM や Maitra (2009) による RndEM の初期値選択法でも用いられる.

しかし, 大規模データで成分数が多い混合正規モデルでは, EM アルゴリズムの収束が遅くなることが予想され, その時, 初期値選択において多大な総反復回数が必要になる. 例えば, 100 個の初期値から EM アルゴリズムを実行し, 各回の反復回数が 1,000 回であれば, 総反復回数は 100,000 回になる. この時, 計算時間は総反復回数に比例する. 反復回数は, EM アルゴリズムの停止条件により決まるため, 適切な条件を与えることも初期値選択法の開発において重要な課題である. 混合正規モデルの解析では, モデルフィッティングにおいて成分数の決定も重要な要素である. この時, 成分数ごとで初期値選択を行うため, 反復回数は莫大になる. そこで, 初期値選択法の加速は非常に有効な手段となる. 開発する初期値選択法に ε R-accelerated EM アルゴリズムを組込み, 反復回数を大幅に減少させることで更なる計算時間の短縮を図る. これらをまとめると,

■　最良な初期値を見つける
■　適切な条件式を与え, EM アルゴリズムの反復回数を減らす
■　EM アルゴリズムの収束を加速し, 反復回数を減らす

が解決すべき研究課題となる.

2. 研究の目的

混合正規モデルのパラメータ推定に EM アルゴリズムが多用されるが, EM アルゴリズムの局所的収束性のため, 尤度関数値は初期値に依存する. このため, 尤度関数値が大域的に最大となる推定値を得ることは困難である. そこで, 複数の初期値から効率良く EM アルゴリズムを実行し, 得られた推定値を統合し初期値を計算する Bayesian model averaging を取り入れた初期値選択法を開発する. そして, その初期値から EM アルゴリズムが計算する尤度関数値が, 大域的に最大となっているかを数値実験で検証する. さらに, 大規模データで成分数が多い混合正規モデルでは, EM アルゴリズムの収束スピードは非常に遅いことが予想される. そこで, 我々の提案した ε R-accelerated EM アルゴリズムを改良し, これを初期値選択法およびパラメータ推定に適用し, 反復回数の減少と計算時間の短縮を目指す.

そのために, 次に示す 3 つの研究課題をそれぞれ解決することにより, 上記の目的を達成する.

■　第 1 課題：混合正規モデルの EM アルゴリズムの加速

ε R-accelerated EM アルゴリズムの適用を試みる. 統計モデルの尤度関数が単峰型 (unimodal) である場合, この収束性と加速性に関する定理は与えられている (研究業績[1]). 混合正規モデルの場合についても, これらの性質の理論研究を行い, 同時に, 数値実験により収束スピードに関する性能評価を行う.

■　第 2 課題：混合正規モデルの初期値選択法の開発と加速化

EM アルゴリズムの総反復回数の減少と計算時間の短縮を図ると同時に, 大域的に最大となる尤度関数値を得るための最良な初期値を見つけるアルゴリズムの開発を行う. 最良な初期値の決定と総反復回数の減少はトレードオフの関係にあるため, 数値実験で確認しつつ両者のバランスを考慮した EM アルゴリズムの停止条件の設定に取り組む. さらに, Bayesian model averaging のアイデアを取り入れ, 複数の推定値を統合した初期値を計算する初期値選択法の開発を目指す.

■　第 3 課題：画像解析と機械学習への適用による性能評価

本研究で開発する初期値選択法の効果および EM アルゴリズムの加速法による計算時間の短縮と解析性能を検証･評価するため, 画像データおよび機械学習データに適用する.

3. 研究の方法

第 1 課題の混合正規モデルに対する ε R-accelerated EM アルゴリズムの改良を数値実験で検証し, 加速性能を評価する. 並行して, 収束性と加速性に関する理論研究を行う. 第 2 課題の初期値選択法の開発では, EM アルゴリズムの停止条件の決定が課題の 1 つであり, ノルム評価や情報量基準等に

基づく条件式を与え，数値実験により最良の停止条件を発見する．さらに，その加速化を行う．また，複数の推定値を統合し初期値を計算する Bayesian model averaging の初期値選択法への適用を研究する．第 3 課題として，画像解析と機械学習のデータに提案手法を適用し，加速性能と初期値選択の効果の評価を行う．

第 1 課題では，ε R-accelerated EM アルゴリズムによる混合正規モデルの EM アルゴリズムの加速を実現する．数値実験による加速性能の検証を行い，加速法が機能しないケースの原因を分析し，混合正規モデルに対応した改良を行う．その後，様々な成分数と変数の組み合わせによる数値実験を行い，この加速法の収束スピードを評価する．ε R-accelerated EM アルゴリズムの収束性と加速性の理論は，統計モデルの尤度関数が単峰型である場合の EM アルゴリズムに対し構築しているため，混合正規モデルにおける理論的検証は行っていない．そこで，混合正規モデルにおける EM アルゴリズムの収束定理に関する研究成果をもとに，改良した ε R-accelerated EM アルゴリズムの収束性と加速性の理論研究を行う．

第 2 課題の初期値選択法の開発では，EM アルゴリズムの停止条件の設定が研究課題の 1 つである．この条件の理論的導出とその最適性の評価を行う方法はない．このため，数値実験において従来用いられているノルム評価に加え情報量基準等に基づく条件式を与え，推定値と尤度関数値の振舞いを確認し，初期値の最良性と反復回数のバランスを考慮した停止条件の決定を行う．さらに，数理的考察によるこの条件の妥当性の検討を行う．次に，初期値選択法の加速化に取り組む．加速の目的は，反復回数の減少とそれに伴う計算時間の短縮である．第 1 課題で開発した加速法を適用し，数値実験による収束と加速性能の評価を行う．また，初期値選択で得られる複数の推定値を統合し初期値とする Bayesian model averaging を導入した方法の研究を行う．

第 3 課題では，最後の第 1 課題および第 2 課題で改良した EM アルゴリズムの加速法と初期値選択法を画像解析と機械学習のデータに適用し，その収束と加速性能を評価する．さらに，MATLAB を用いて得られる解析結果と我々の結果の比較を行い，本研究課題で開発した方法の解析性能を評価する．

4．研究成果

混合正規分布モデルのパラメータ推定において，emEM を拡張した初期値選択法の開発を試みた．このアルゴリズムでは，初期値の候補選択のための計算ステップがキーであり，より少ない計算回数と時間で最適な初期値を見つけることが必要となる．そこで，このステップにおける EM アルゴリズムの反復の停止条件を緩く指定することにより，総反復回数の減少を目指した．数値実験では，その停止条件の精度をどのように設定するかを様々な混合正規モデルのもとでおこなった．**これらの結果は，上記の目的を達成したものである考え，研究会などでも発表した．しかし，その妥当性についての理論的な検証までには至らなかった．現在も，その点を解決すべく引き続き研究をおこなっている．**

混合正規分布モデルのパラメータ推定における初期選択の研究過程において，この手法が非計量データの主成分分析（非計量主成分分析）のための交互最小二乗法にも応用できることがわかった．したがって，本研究の申請期間の途中の年度から，この研究も同時におこなうこととした．交互最小二乗法のための初期値選択法の基本的アイデアは，EM アルゴリズムのそれと同じであるが，交互最小二乗法の計算ステップに合わせて修正をおこなった．混合正規分布モデルと異なり，非計量主成分分析のモデルはシンプルであり，そのため性能検証のための数値実験のおこないやすかった．この研究でも，最適な停止条件の理論的根拠を数値から導き出すことはできなかったが，指針を与えることはできた．そして，この初期値選択法に ε -accelerated EM アルゴリズムで用いたベクター ε 法による交互最小二乗法のための加速法を用いたことで，平均において反復回数で 1/6.6，計算時間で 1/10 程度の短縮を実現した．この研究成果は，論文として発表し，2020 年 7 月に掲載される予定である．

本研究課題とは直接関係はないが，開発した EM アルゴリズムの 2 つの加速法である ε -accelerated EM アルゴリズムと ε R-accelerated EM アルゴリズムの論文が日本計算機統計学会の和文誌「計算機統計学」に掲載されることが決まった．また，EM アルゴリズムの専門書籍として共立出版の one point シリーズから 2020 年 7 月に「EM アルゴリズム」というタイトルで出版される．

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Uno, K., Satomura, H. and Adachi, K.	４．巻 94
２．論文標題 Fixed factor analysis with clustered factor score constraint	５．発行年 2017年
３．雑誌名 Computational Statistics and Data Analysis	６．最初と最後の頁 265-274
掲載論文のDOI（デジタルオブジェクト識別子） j.csda.2015.08.010	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 －

１．著者名 黒田正博	４．巻 30(2)
２．論文標題 vector ε アルゴリズムによるEM アルゴリ ズムの収束の加速化とその改良	５．発行年 2019年
３．雑誌名 計算機統計学	６．最初と最後の頁 －
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 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 Kuroda, M., Mori, Y. and Iizuka, M.	４．巻 －
２．論文標題 Initial value selection for the alternating least squares algorithm	５．発行年 2020年
３．雑誌名 Advanced Studies in Classification and Data Science	６．最初と最後の頁 －
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 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10件（うち招待講演　1件／うち国際学会　9件）

１．発表者名 Kuroda, M., Mori, Y., Iizuka, M.
２．発表標題 Improvement of computation for nonlinear multivariate methods
３．学会等名 The 10th Conference of the IASC-ARS/68th Annual NZSA Conference（国際学会）
４．発表年 2018年

１．発表者名
Kuroda, M., Mori, Y

２．発表標題
Speed-up bootstrap computation for the covariance matrix of MLEs from incomplete data

３．学会等名
2nd International Conference on Econometrics and Statistics EcoSta2018（国際学会）

４．発表年
2018年

１．発表者名
Yoshioka, M., Kuroda, M., Mori, Y.

２．発表標題
Computational efficiency for fuzzy clustering

３．学会等名
The 11th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2018)（国際学会）

４．発表年
2018年

１．発表者名
Kuroda, M., Mori, Y.

２．発表標題
Speed-up of bootstrap computation to incomplete data.

３．学会等名
The IASC-ARS 25th Anniversary Conference & CASC 2nd Annual Conference.（招待講演）（国際学会）

４．発表年
2018年

１．発表者名
Kuroda, M., Mori, Y., Iizuka, M.

２．発表標題
Initial value selection for the alternating least squares algorithm.

３．学会等名
Conference of the International Federation of Classification Societies 2017（国際学会）

４．発表年
2017年

１．発表者名
Mori, Y., Iizuka, M., Kuorda, M.

２．発表標題
Variable selection for mixed measurement level data in dimension reduction methods and its computation.

３．学会等名
The 1st International Conference on Econometrics and Statistics (EcoSta 2017)（国際学会）

４．発表年
2017年

１．発表者名
黒田正博，森 裕一，飯塚誠也

２．発表標題
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

３．学会等名
日本計算機統計学会第30回大会

４．発表年
2016年

１．発表者名
Kuroda, M.

２．発表標題
Fast estimation using the EM algorithm for Gaussian mixture models

３．学会等名
The 4th Institute of Mathematical Statistics Asia Pacific Rim Meeting（国際学会）

４．発表年
2016年

１．発表者名
Mori, Y., Sakakihara, M., Kuroda, M. and Iizuka, M.

２．発表標題
Acceleration of iterative methods for nonnegative matrix factorization

３．学会等名
COMPSTAT 2016（国際学会）

４．発表年
2016年

１．発表者名 Kuroda, M., Mori, Y., Iizuka, M. and Sakakihara, M.
２．発表標題 Acceleration of convergence of the alternating least squares algorithm for mixed measurement level multivariate data
３．学会等名 The 10th International Chinese Statistical Association (ICSA) international conference（国際学会）
４．発表年 2016年

〔図書〕　計3件

１．著者名 森 裕一・黒田 正博・足立 浩平	４．発行年 2017年
２．出版社 共立出版	５．総ページ数 120
３．書名 最小二乗法・交互最小二乗法（統計学One Point 3）	

１．著者名 Mori. Y., Kuroda, M., Makino, N.	４．発行年 2016年
２．出版社 Springer	５．総ページ数 80
３．書名 Nonlinear Principal Component Analysis and Its Applications	

１．著者名 Adachi, K.	４．発行年 2016年
２．出版社 Springer	５．総ページ数 301
３．書名 Matrix-Based Introduction to Multivariate Data Analysis	

〔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
研究分担者	足立　浩平 (Adachi Kohei) (60299055)	大阪大学・人間科学研究科・教授 (14401)	
研究分担者	飯塚　誠也 (Iizuka Masaya) (60322236)	岡山大学・全学教育・学生支援機構・教授 (15301)	
研究分担者	森　裕一 (Mori Yuichi) (80230085)	岡山理科大学・経営学部・教授 (35302)	