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 １５ 日現在

機関番号：３４４１６
研究種目：基盤研究(C)（一般）
研究期間：2018～2023
課題番号：１８Ｋ１１２０５

研究課題名（和文）EMアルゴリズムに代わる欠測データを用いたパラメータ推定法の開発

研究課題名（英文）Development of alternative methods of parameter estimation to the EM algorithm using missing data

研究代表者
高井　啓二（Takai, Keiji）

関西大学・商学部・教授

研究者番号：２０５７２０１９

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究の目的は，現在欠測データを用いた統計モデルのパラメタを推定する際に標準的に使われているEMアルゴリズム(以降，EMと略)に代わる計算手法を開発することであった．本研究では第一に，従来のEMの欠点を克服するためにフィッシャースコアリング法を改良した不完全データのフィッシャースコアリング法を開発した．この方法で収束のスピードは一般のEMよりも早くなり，EMでは出せなかった誤差分散も導出することができるようになった．第二に，ガンマ分布及びその混合分布に対して適用できるパラメータ計算の方法を開発した．この方法は初期値を自動で発見でき，絶対に計算を失敗しない特性を持っている．

研究成果の学術的意義や社会的意義
本研究では統計モデルのパラメータを推定するための数値計算の方法を開発した．本研究で得られた成果の一つである不完全データのフィッシャースコアリングは，計算スピードとしては早い部類には入らない．しかし，本研究により，その計算過程は本質的には最急降下法となっていることや，その計算プロセスが既存のEMアルゴリズムに近いことなどの解析を行なうことができた．これにより本研究のようなEMアルゴリズムの単純な改善では超一次収束を達成できないことが示唆されている．

研究成果の概要（英文）：The purpose of this study is to develop an alternative to the EM algorithm (hereafter referred to as EM) that is currently used as the standard method for estimating parameters in statistical models with missing data. First, we developed a Fisher scoring for incomplete data, which is an improvement of the Fisher scoring method, to overcome the shortcomings of conventional EM. This method provides a better convergence speed, faster than that of general EM, and also derive the error variances that couldn't be obtained with EM. Second, we developed another parameter estimation method applicable to the gamma distribution and its mixtures. This method has the property of being able to automatically find initial values. This method has the property that it never fails in the process of computation, unlike ordinary computation methods including the Newton-Raphson method.

研究分野：統計学

キーワード：欠測データ　不完全データ　フィッシャースコアリング　加速法

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

データを収集した結果，当初の計画通り完全なデータが得られないことは多い．このような欠測データをはじめとした不完全データは多くの統計的モデルの推定に大きな障害となる．現在の統計的なモデルの多くが，完全データが得られることを想定しているからである．不完全データからパラメータを推定する最も有用な方法の一つは，EM アルゴリズムである．EM アルゴリズムはその提唱以来，いろいろな分野に拡張され，適用例を増やしながら標準的なパラメータ推定方法となってきた．EM アルゴリズムは計算の安定性という意味では，大変優れている．一方で，EM アルゴリズムの欠点としては，収束が遅いこと，そして統計的モデルにおいて重要な誤差分散が得られないということが指摘されてきた．この二つの欠点を克服するために多くの研究がなされてきたが，EM アルゴリズムをニュートン型の計算方法に帰着させるという手法が取られてきた．確かにニュートン型計算方法は収束が早く，多くの場合は誤算分散も得ることができるという点において従来の EM アルゴリズムの欠点を克服するように見えた．しかし，ニュートン型の計算方法を使うことによって，EM が本来持っていた計算の安定性が失われ，計算量も増加するという問題が起きてしまっていた．

２．研究の目的

本研究の目的は，EM アルゴリズムの二つの欠点を克服した方法を開発することである．その際に，従来の研究とは異なり，計算量が少なく，安定性の高い（つまり，計算に失敗することの少ない）計算方法となることを目指す．

３．研究の方法

(1)従来の EM アルゴリズムとニュートン型の計算方法との関連性を解析的に調べ，EM の安定性を新たな計算方法に移植する．
(2)EM の中で(1)で開発した方法を使うことが大変な場合は，より簡単な方法を適用できないか調べる．

４．研究成果

(1) 不完全データに対するフィッシャースコアリング法の開発

データ解析を行う際，想定しているモデルのパラメータ推定は必ず必要になる．加えて，推定されるパラメータの誤差分散もその後の推測の観点から重要な量である．一般に，データが完全であれば，パラメータ推定も，その誤差分散の推定もそれほど難しいものではない．一方，データが不完全であればパラメータ推定は格段に難しくなる．このような状況に対して従来は EM アルゴリズムが用いられてきた．しかし，EM アルゴリズムは誤算分散を容易には計算できないという問題点があった．そこで本研究では，完全データのフィッシャースコアリング法を不完全データに拡張して不完全データのフィッシャースコアリング法を開発した．この計算法の特徴は，特殊な場合には EM アルゴリズムに帰着する，あるいは EM の更新式に非常に近くなることである．これは開発した手法が EM アルゴリズムのニュートン型更新法の形式になっていることを意味する．両者の大きな違いは，EM アルゴリズムにはステップ幅の調整という概念がないが，フィッシャースコアリング法の場合はステップ幅の調整という概念が存在することである．この利点を使って，目的関数の二次近似から求めたステップ幅を使うと，収束点が近づくとフィッシャースコアリング法は適当に決めたステップ幅よりも大幅に加速する．EM アルゴリズムは計算の初期段階では更新が多く，収束点の近傍では更新が少なくなることが知られている．このフィッシャースコアリング法は EM を特殊な場合として含んでいるので，EM の初期段階で更新量が多くなる性質を維持する一方で，収束点が近づいてもステップ幅を適切に選択していることにより依然として高速な収束が維持されている特徴がある．このような特徴は実際のデータを用いた数値計算においても確認されている．以上の結果は，統計学における数値計算を扱う査読付き雑誌に採択されている．

(2) 不完全データの AIC

データ解析を行うとき，採用候補となるモデルは通常複数存在する．その中でどれがもっともよいのかを選ぶために利用するのが情報量規準である．ところが不完全データに対してはこのような情報量規準は利用できない．それは不完全データでは想定モデルと，推定できるモデルが一般には同一ではないからである．いくら良いモデルを想定しても，データの不完全性のためにパラメータ推定が出来ないことは多い．そこで，本研究では，選択対象となる想定モデルの集合と，

推定に使われるモデルが違う状況での情報量規準を開発した.

推定に使うモデルはできるだけ大きいデータを考える．これには二つの利点がある．一つは，大きいデータであればデータの不完全性を導く要因となる変数が含まれる可能性が高くなり，その結果，パラメータが一致推定できることである．二つ目は，興味あるパラメータの誤差分散が小さくなることである.一般に,大きなモデルの中の興味あるパラメータの一部の誤差分散は，単に興味ある部分のデータだけを使ったパラメータの誤差分散よりも小さくなることが知られている．このことは，一部の興味あるパラメータが推定可能であっても正しい.

本研究では，できるだけ大きいデータでパラメータ推定を行い，興味あるモデルのパラメータに代入を行い，その評価を行うことのできる情報量規準(不完全データの AIC)を作った．この不完全データの AIC はいくつかのシミュレーションで，これまで開発されてきた他の不完全データの AIC よりも優れたパフォーマンスを示した．これらの結果は査読付き論文として公表した.

(3) ガンマ混合分布から得られる分布のパラメータ推定法の開発

ガンマ混合分布はその形状を多様に変えることができるため，さまざまな分野で現在でも，使用されている．代表的には機械の故障率や地下水の化学物質の含有量などのデータによく適合することが知られている．ガンマ分布はかなり歴史のある分布であるが，最尤推定値を得ることは簡単ではない．正規分布のように明示的な形で最尤推定値を得ることができないため，何らかの数値計算法に頼らざるを得ない．特に形状パラメータが小さいときにはニュートン型の計算が破綻しうることが報告されている．このようなパラメータ計算の問題は単にガンマ分布だけにとどまらない.

ガンマ分布は有用な混合分布を作ることでも知られている．例えば，負の二項分布はポアソン分布をガンマ混合した結果として得られる．消費者データの分析では，個人の買い物個数の分布をポアソン分布，個人間の異質性をガンマ分布として仮定して，負の二項分布を導くという定式化がなされている.交通データに対しても同様の定式化がなされている.負の二項分布の問題は，ガンマ分布と同様に最尤推定値が明示的な形で得られないことである．そこで，負の二項分布に従うと考えられるデータをポアソン分布とガンマ分布の同時分布から得られる不完全データとみなすことで EM アルゴリズムを構築する．ところがこの場合も，EM アルゴリズムの明示的な更新式を得ることができない．しかし,EM の更新式を導く計算式を変形するとガンマ分布の場合に得られるのと同じ形の方程式が得られる.同様の方程式は,指数分布のガンマ混合である Pareto 分布についても得られる．つまり，ガンマ分布の推定方法を開発することは，ガンマ混合によって導かれる分布の推定方法を開発することに等しい.

ところで，ガンマ分布の最尤推定値に非常に近いものとして一般化ガンマ分布による推定値が存在する．一般化ガンマ分布はパラメータを特定の値に固定することで，ガンマ分布に帰着する．この推定値の特徴は，ガンマ分布とは異なり明示的な形の推定量が得られることにある．負の二項分布や Pareto 分布をガンマ分布の混合ではなく一般化ガンマ分布の混合と考えることで，パラメータの推定値を得ることができるようになった．実際のデータでもこの方法の有効性は確認できた．本研究の結果は，国内会議で発表した．さらにこの結果を拡張し一般化したものを国際会議でも発表している.

５．主な発表論文等

〔雑誌論文〕　計2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　1件）

１．著者名 Keiji Takai	４．巻 30
２．論文標題 Incomplete-data Fisher scoring method with steplength adjustment (article) Author	５．発行年 2020年
３．雑誌名 Statistics and Computing	６．最初と最後の頁 871-886
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s11222-020-09923-z	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 Keiji Takai	４．巻 47
２．論文標題 On the use of the selection matrix in the maximum likelihood estimation of normal distribution models with missing data	５．発行年 2018年
３．雑誌名 Communications in Statistics - Theory and Methods	６．最初と最後の頁 3392-3407
掲載論文のDOI（デジタルオブジェクト識別子） 10.1080/03610926.2017.1353631	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3件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 高井啓二
２．発表標題 負の二項分布モデルによるチラシ掲載効果の検証
３．学会等名 関西 大学 RISS(Research Institute for Socionetwork Strategies of Kansai University) セミナー「広告効果測定，視線追跡データとパスデータの融合」
４．発表年 2021年

１．発表者名 高井啓二
２．発表標題 欠測データを用いたフィッシャースコアリング法
３．学会等名 科研費シンポジウム「高次元複雑データの統計モデリング」
４．発表年 2019年

１．発表者名 高井啓二
２．発表標題 パラメータ分割による不完全データフィッシャースコアリング
３．学会等名 統計関連学会連合大会 2018
４．発表年 2018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