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　３０　年　　６　月　１４　日現在

機関番号：　３４４１６
研究種目：　若手研究(B)
研究期間：　2014～2017
課題番号：　２６７３００２２

研究課題名（和文）欠測データ解析における新しい無視可能条件の構築と推定量の分布の研究

研究課題名（英文）Study on the creation of a new ignobility condition and the investigation of an estimator distribution in the analysis of missing data

研究代表者

高井　啓二（TAKAI, Keiji）

関西大学・商学部・准教授

研究者番号：２０５７２０１９

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：本科研費課題では，欠測データを分析する理論を開発し，その理論を判別分析に応用してきた．理論面では，単調，非単調両方の欠測データパターンの下における，MARと同等の条件付き独立性を導出した．加えて，選択行列を用いて，パラメータの計算・推定を容易に行う方法を開発した．この方法により，推定量の性質も調べることが可能となった．応用面では，欠測データの理論を用いて判別分析における半教師あり学習に取り組んだ．ここでの半教師あり学習とは，一部のみが観測されているデータから判別規則を推定することである．そこで欠測データの理論を用いると，一定の条件下ではこのような場合でも正確な判別が可能であることを示した．

研究成果の概要（英文）：In this project, I developed the theory for analysis of missing data and applied it to special data in the discriminant analysis. As theoretical study research, I derived conditional independences equivalent to MAR(missing at random) under monotonic and non-monotonic missing-data mechanisms. In addition, I constructed a method to overcome some difficulties in computation and estimation of the parameters of interest with missing data by using the selection matrix. It allows us to investigate properties of the estimator which are necessary for inference. As application research, I tackled a semi-supervised learning problem in discriminant analysis using the missing-data analysis theory. The semi-supervised learning is an estimation of the parameters from the partially observed data. I showed that the use of the missing-data analysis theory makes it possible to obtain the correct discriminant rule even with the such partially observed data.

研究分野：　統計科学

キーワード：　欠測データ　MAR　条件付き独立　EMアルゴリズム　判別分析　漸近理論

１．研究開始当初の背景

データを収集すると，当初の意図や計画に反して観測できない値が存在することがある．そのような観測できなかった値を欠測値と言う．例を以下の表1に与える．表中の?が欠測値である．欠測値がある場合に通常の統計手法を用いるためには (平均値を求める場合さえも)様々な工夫が必要である．通常の統計手法が，欠測値のないデータ(完全データ)のために開発されてきたからである．社会科学など人を扱うデータの多くは欠測値を含むため，そのようなデータへの対処方法の研究が必要である．欠測値データ解析の理論は，欠測値と同様の性質を持つと見なすことができる潜在変数モデルや，判別分析モデル，因果推論，EMアルゴリズムの理論にも応用できる点で重要である．

表 1 欠測データの例

id	変数1	変数2
1	?	30.8
2	14.5	?
…	…	…
n	13.9	32.1

一般的なデータの解析の手順をおおまかに言えば，データの収集，モデルの構築，モデルを規定するパラメータの推定，パラメータの推定量を用いた推測，となる．欠測データはデータ収集段階で不可避的に生じる．モデルの構築は，データの基本的な特性値(平均値，分散など)を推定して，それをもとに行われるため，欠測データからそれらの値が推定できることが実用上必要である．モデルを規定するパラメータの推定段階における，推定法の重要性は言うまでもない．最後の推測段階においても，その前段階での推定の影響が大きい．推定量の性質は，どのような推定法によって得られたに大きく依存するからである．従って，欠測データを分析する限りにおいてはどの段階においても，いかにパラメータを推定するかという問題が存在している．本研究課題においては，以下のような推定問題が重要であると考えている．

(1) 欠測データにもとづく推定では，Rubin (1976) が導入した「MAR」（あるいは「無視可能」）という概念が極めて重要な役割 を果たしてきた.欠測データを生じさせるメカニズム（以下，欠測メカニズムと言う）がMAR である とは，欠測メカニズムをモデリングすることなく，最尤推定できるということである．欠測メカニズムをモデリングすることは実際にはかなり困難なため，MARであれば分析上非常に有用である.ところが，このMARという概念は非常に理解しづらく使いづらい．そしてSeaman et al. (2013)が主張したように，多くの人(統計学者にさえ)に誤解されてしまっている．それゆえ，MARを数学的に扱いやすい条件付き独立性などを用いて表すことが必要となっている．

(2) 一般に，推定値は何らかの方程式の解として得られるが，その方程式が明示的に解けるとは限らない．場合によっては，明示的にその形を示すことができず，単に解けない方程式の解としてしか得られないことがある．一方，明示的に解けた場合には，その推定値（あるいは推定量）の性質を詳細に調べることができる．従って，推定値は明示的な形で得られることが望ましい．

しかし，欠測データを用いる時，明示的に推定値が得られることはほとんどない．欠測データの場合にも，完全データの場合と同じく最尤推定値を得るには尤度方程式を解く．しかし，欠測データの尤度方程式は，欠測値を積分で消した結果として得られる．そのため，推定値を明示的な形で得ることは一般には不可能であった．多変量解析における基本的な分布である多変量正規分布においてさえも，推定値を明示的に得ることはできていない．そのため欠測データから計算される推定値（推定量）の性質を調べることが困難となっていた．そこで，欠測データにもとづいて推定値を明示的な形で表現する必要性がある．

(3) 欠測データの理論は，欠測データと同じ構造を持ったデータの分析にも利用できるはずである．欠測データは一般にはデータ収集の際に意図していない形で出現する．しかし，場合によっては，欠測データは意図的なデータ収集の結果として得られることがある．例えば，健康診断をすると血圧，肝臓の状態を表す値などが得られる．このような値が悪い一部の人に対しては，更に細かい検査が行われて，健康などうかの判断がなされる．このとき，健康診断の結果はすべての人に対して観測されているが，健康かどうかについては一部の人にだけ観測されている．つまり，この健康診断の結果を表すデータと最終的に健康か病気かを表すデータを合わせたものは欠測データとなっている．それゆえ，欠測データの理論を用いることができる．

このようなデータに対して実用上重要なのは，健康診断の値がどのような値であれば健康であり，どのような値であれば健康でないのかという判別ルールを構築することである．データは欠測データであるから，通常の判別ルールの推定方法は使えない．そこで，欠測データの理論を用いて，欠測データから正しい判別ルールを構築する方法を開発する必要がある．

２．研究の目的

完全データを用いる時，統計解析における推定の基本的な方法は，最尤推定法である．最尤推定法によって得られる推定値にもとづく最尤推定量は，サンプルサイズを大きくするとき，真の値に収束し，その分布が正規分布で近似できるという好ましい特性を持っている．もちろん実際の調査においてサン

プルサイズを大きくすることはできないので,「サンプルサイズを大きくする」とは「大きなサンプルサイズの標本をとるという望ましい状況ならば」という意味である．そのとき，標本は母集団に非常に近いので，標本の特性量（最尤推定量）と母集団の特性値（パラメータ値の真値）も近いことが望ましい．完全データにもとづくとき，最尤推定量が真の値に収束するというのは，最尤推定量が対応する（その推定量によって推定したい母集団の）特性値に近いことを言っている．近いというだけでは，どれくらい近いのかわからない．どれくらい近いのか，推定量が母集団の値と一致しているとみなしていいのかといった疑問が生じる．そこで，最尤推定量の分布を用いて推測することが必要となる．上で述べた,「その分布が正規分布で近似できる」というのは，完全データの最尤推定量の分布は，上と同じ意味でサンプルサイズが大きいとき，正規分布で近似できるという性質を持っているということである．

本研究における究極の目的は，欠測データにおいても，完全データの場合と同様に，最尤推定法を用いた推定を行うこと，である．そしてその結果得られた，最尤推定量が完全データと同じか，あるいは近い性質を持っていることがわかることが望ましい．従って，本研究課題では，欠測データにもとづいて得られた最尤推定量が，そのような望ましい性質を持っていることを確認することを目的としている．

３．研究の方法

「１．研究開始当初の背景」における番号に対応した研究の方法を述べる．

(1) 本研究では，グラフィカルモデルにおいて知られている独立性に関する様々な性質を用いる．

(2) 数理統計学における基本的な知識と確率変数によってその要素が表される行列に関する知見を用いる．

(3) まず欠測データの理論を用いて，漸近的な（サンプルサイズを大きくしたとき）相対効率の表現を調べる．その後，シミュレーション技法を持いて，実際のパラメータ値を代入して漸近的な相対効率の実際の値を調べる．

４．研究成果

「１．研究開始当初の背景」における番号に対応した研究の成果を述べる．

(1) 欠測データメカニズムを条件付き独立性で表すための研究

本研究では，MAR の同値条件として，いくつかの条件付き独立性を導出した．条件付き独立性の利点は，様々な知見が統計学において知られていること，そして条件付き独立性はしばしばグラフィカルモデル（GM）と呼ばれるグラフとして表現できることにある．すでにいくつかの研究では GM を用いて欠測データメカニズムをモデリングすることが行われている．そのため，どのような GM が基本的な欠測データメカニズムである MAR に対応しているのかを知ることは応用上も重要である．

この研究の結果，単調な欠測パターンのとき，MAR と同等の条件付き独立性を導出することができた．それにより，これまで様々な研究者が MAR だと同等だと主張してきた独立性についての別証明や簡潔な照明を与えることができた．非単調な欠測パターンの場合には，独立性は十分条件であった．しかし，特殊な構造の欠測データにおいては，MAR と同等の独立性が存在した．それにより，欠測データ解析と密接な関係にある因果推論の分野における「強い無視可能性」と MAR の関係性を示すことができた．これらの結果の応用として，特定の欠測データパターンにおいて MAR を表すグラフィカルモデルの集合を特定することができた．

(2) 欠測データ解析における選択行列の利用に関する研究

本研究では，選択行列を用いて多変量正規分布にもとづいて推定・推測を行う場合に限り，最尤推定値の明示的な表現，その性質などを導出した．多変量正規分布は，多変量解析を行う際に使われる標準的な分布であるため，この研究結果は多変量解析のかなり広い部分をカバーしている．多変量正規分布からデータが取られている場合，どのような欠測値もどの要素が欠測しているかを表す欠測指標にもとづく選択行列と，一部は観測されていないかもしれない完全データのベクトルの積で表現することができる．選択行列の使用により，尤度は共通のパラメータで表現することができるため，推定値の計算が簡単になる．それによって最尤推定値を明示的に導出することができる．このとき推定値は，完全データのときと異なり，欠測指標とデータ両方の関数になる．

欠測指標とデータの両方を定数とすると，得られる推定値は最尤推定値である．このとき，明示的な形の最尤推定値だけでなく，尤度関数と EM アルゴリズムの関係が明らかとなった．具体的には，尤度関数の変形が EM アルゴリズムであり，EM アルゴリズムはある種のフィッシャースコアリング法と同等であることが示された．

一方，欠測指標とデータの両方を確率変数とすると，最尤推定量が得られる．このとき最尤推定量の一致性（サンプルサイズが大きい時，真の値に収束する性質）が従来とは異なり簡単な証明で得られる．推定量が EM アルゴリズムにより更新されるとき，その推定量の列が one-step 推定量としてどのような性質を持つのかを示すことができた．また，二つの代表的な欠測データメカニズム(MCAR, MAR)の下での情報行列（分散共分散行列）の明示的な形を与えることができた．

(3) 欠測データ解析の理論を使った，判別分析に関する結果

欠測データがどのように発生するかによって，判別ルールの構成規則を変える必要がある．上で出た健康診断の例では，健康か病気かを判別する検査を受ける人は，健康診断の結果が良くなかった人に限られている．つまり，この場合，観測された健康診断の結果にもとづいて，更に健康診断を受けるかどうかが決まっている．よって，このような欠測データメカニズムは MAR である．一方，金融会社のどの顧客が優良かを判断するには，各顧客の属性にもとづいて判断するのではなく，その顧客の中からランダムに顧客を選んで優良かどうかを判断する必要が生じることがある．この場合には，やはり欠測データとなるが，欠測データメカニズムは MCAR と呼ばれる．以上のように，欠測データにおける判別ルールの構成については，MAR と MCAR の二つの状況を考える必要がある．

そこで本研究では，この二つの場合に，どのように判別ルールを構築すればよいのか，そして，その判別ルールの正確性について，欠測データの解析法を用いて研究した．その結果，結果がわかっているデータの存在が，全体に大きな影響を及ぼすことがわかった．実際には，健康診断の結果や優良顧客かどうかを知るためには，多大な金銭的コストや時間的コストがかかるため，少数のデータについてのみ結果を得ることができる．この結果は，その結果についてはできるだけ正確にすることで全データを用いたときの誤判別率をコントロールできることを意味している．

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）

① (査読あり) Keiji Takai (Accepted for publication). “On the use of the selection matrix in the maximum likelihood estimation of normal distribution models with missing data.” Communications in Statistics : Theory and Methods, Taylor & Francis.

② (査読あり) Kenichi Hayashi and Keiji Takai (2017). “Finite-sample analysis of impacts of unlabelled data and their labelling mechanisms in linear discriminant analysis.” Communications in Statistics : Simulation and Computation, 46(1), 184-203. Taylor & Francis.

③ (査読あり) Keiji Takai and Kenichi Hayashi (2014). “Effects of unlabeled data on classifi- cation error in normal discriminant analysis.” Journal of Statistical Planning and Inference , 147, 66-83. Elsevier.

〔学会発表〕（計 9 件）

① Keiji Takai. (2017). “An incomplete-data Fisher scoring method with an acceleration method.” The 10th Conference of the IASC-ARS/68th Annual NZSA Conference (IASC-NZSA 2017). The University of Auckland, New Zealand.

② Keiji Takai. (2017). “An incomplete-data Fisher scoring.” Hangzhou international statistical symposium (日中統計会議).Hangzhou Huanzhang HNA Resort, China.

③ 高 井 啓 二 .(2017).“Parameter estimation with incomplete-data Fisher scoring method.”（英語セッション）統計関連学会連合大会 2017 , 南山大学, 愛知県.

④ 林賢一・高井啓二.(2016).“MARデータにおける変数の部分集合に対する 情報量規準.” 統計関連学会連合大会 2016, 金沢大学, 石川県.

⑤ 高井啓二.(2016). ”欠測データにもとづく正規分布モデルの最尤推定にお ける選択行列の使用について. ”2015 年度科学研究費補助金シンポジウム「事象時間データ解析に関する理論と方法論およびその応用.」 大阪大学, 大阪.

⑥ Keiji Takai and Keinichi Hayashi.(2015). “An information criterion for a subset of MAR data.” 8th International Conference of the ERCIM Working Group on Computational and Methodological Statistics (CMStatistics 2015), University of London, England.

⑦ 高井啓二.(2105).‘非単調欠測データに対する正規分布モデルの最尤推定量について” 日本行動計量学会第 43 回大会.首都大学東京, 東京.

⑧ 高井啓二.(2105).“欠測データ解析入門” 日本行動計量学会第 17 回春の合宿セミナー.東京大学, 東京.

⑨ 高井啓二.(2014).‘MARと独立性の関係”2014年度統計関連学会連合大会. 東京大学, 東京.

〔図書〕（計　1　件）

① 高井啓二，星野崇宏，野間久史.(2016).
欠測データの統計科学-医学と社会科 学への応用-.岩波書店.総ページ 232 ページ，担当 23-101,196-213 ページ.

〔産業財産権〕

○出願状況（計　0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
高井 啓二（TAKAI, Keiji）
関西大学・商学部・准教授
研究者番号：20572019