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　４ 年　 ５ 月 ２０ 日現在

機関番号：１４４０１
研究種目：若手研究
研究期間：2019～2021
課題番号：１９Ｋ１４５９２

研究課題名（和文）無視不可能な欠測値データを可能にする多重代入法の開発

研究課題名（英文）Multiple Imputation making non-ignorable missing-data mechanism ignorable in a new framework

研究代表者

森川　耕輔（Kosuke, Morikawa）

大阪大学・基礎工学研究科・講師

研究者番号：４０８２４３０５

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）：無視不可能な欠測値データ解析における，モデルの識別性及びセミパラメトリック漸近有効推定量を提案した．また，通常仮定される完全データのモデルへの仮定を観測データのモデルに対し課すことで，欠測メカニズムの無視可能性を観測データから検証可能な手法を提案した．さらに，令和３年度には，観測データのモデルを一般化線形モデルとした下で新たな識別性の十分条件を提案した．提案された新しい識別可能性条件の下，パラメータを効率的に推定可能なFractional代入法と多重代入法を提案した．今後は，これまでに得られた成果を，新たな無視可能性条件の下，セミパラメトリックな多重代入法へ拡張することが課題である．

研究成果の学術的意義や社会的意義
得られるべきデータが得られない欠測値の問題は極めて重要であり，欠測に適切に対処しない解析方法は推定量に重大なバイアスを生じ得る．本研究では，応用上しばしば仮定されるデータから検証不能な欠測メカニズムの無視可能性を，観測データから検証するための十分条件を与え，これまで主観的に課されていた仮定を客観的に吟味することが可能となった．また，無視不可能であるという状況下でセミパラメトリック漸近有効推定量を提案し，新たな欠測値の代入法を開発した．これらの成果により，これまで応用上忌避されてきた無視不可能であるという状況下でのデータ解析手法の数理的基盤が構築され，より有用かつ実用的なものになるだろう．

研究成果の概要（英文）：We proposed a new model-identification condition and a semiparametric efficient estimator to analyze non-ignorable missing data. We also proposed a method for verifying the ingnorability of the missing-data mechanism from observed data. The key idea is to impose assumptions on the observed data model, not imposing on the complete data model. In the last year, we proposed a new sufficient condition for identifiability under the assumption that the observed data model is a generalized linear model. Under the proposed new identification conditions, we proposed Fractional and Multiple imputations, which can efficiently estimate parameters. Extending the results to semiparametric estimation under the new ignorability conditions is future work.

研究分野：数理統計学

キーワード：欠測値データ解析　多重代入法

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

得られるべきデータが得られない欠測値の問題は極めて重要であり，欠測に適切に対処しない解析方法は結果に大きなバイアスを生む．近年はビッグデータや超高次元データの解析需要が増大し，欠測値データ解析の方法論は益々重要になっている．例えば，米国政府による全国健康栄養調査（National Health and Nutrition Examination Survey）では，毎年約 4 万世帯に対して調査を実施する大規模な調査であるが，1997 年から 2004 年での世帯収入の未回答率の平均は 30.75%である．このような欠測値データに対して，データが得られない被験者を除外した解析は，サンプルサイズの減少に伴う有効性の低下だけではなく，推定量に重大なバイアスを生じ得る．そのため，欠測値を考慮した適切な解析法が必要となる．

欠測値データの解析においては，当該データを条件つけたときそのデータが実際に観測される確率を表す，欠測メカニズムの理解が重要となる．欠測メカニズムは，全ての変数ではなく，完全に観測されている変数さえ用いれば，データが観測される確率を説明できる場合，無視可能であると言われる．一方で，データが観測されるかどうかに欠測している変数も依存している場合，無視不可能であると言われる．欠測メカニズムが無視不可能な場合，完全データを解析する際は不要である，完全データと欠測メカニズムの両方のモデリングが必要となる．データが欠測している状況では，我々は一部の観測されているデータしか所持しておらず，仮定した完全データのモデルの妥当性を視覚的にも，また（情報量規準などにより）客観的にも評価することは難しく，已む無く主観的な仮定が必要となる．そのため，欠測値データ解析ではデータの欠測は無視可能であるという仮定がしばしば置かれるが，一般的にはこの仮定は成り立たないため，無視不可能な状況下でできるだけ弱い仮定の下での統計的推測法の開発は喫緊の課題である．

２．研究の目的

本研究の目的は，欠測値データ解析において，欠測メカニズムが無視不可能である際必要とされる主観的な仮定をできるだけ緩めることである．その際，次の 3 つの意味で仮定を緩める．

設定 1．完全データへのモデリングを，より客観的な観測データへのモデリングに変更；
設定 2．完全データのモデリングを必要としないセミパラメトリック推測法；
設定 3．完全データのモデリングを必要とせず，さらに無視可能性条件を[1]で発見された新しい無視可能性条件への拡張．

以上の３つの設定下で，パラメータの識別性の十分条件を与え，さらに代表的な欠測値データ解析手法の１つである“多重代入法”の開発を行う．多重代入法とは，図１で表されている通り，欠測している値を埋め，擬似的な完全データを複数個作成することにより，欠測値データ解析を行う手法である．通常の重み付き法と比べると，完全データさえ作成してしまえば，通常の完全データを解析するためのプログラムが使用でき，欠測値データ解析に詳しくない応用の研究者も使用できるといったメリットがある．

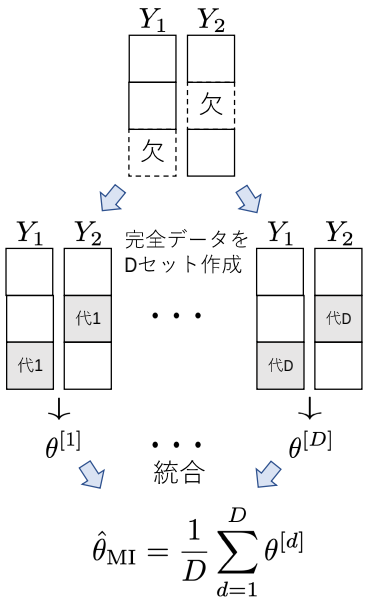

図 1. 多重代入法の概要.

３．研究の方法

主に，次の 3 つの問題を解決することで，この目的を達成する．

研究①（設定 1：完全データへのモデリングを，より客観的な観測データへのモデリングに変更）

多重代入法を行うためには，完全データのモデリングが必要である．しかし，完全データのモデリングには主観的な仮定を要するため，この完全データのモデリングを観測データへのモデリングに変更することで，データから検証可能な無視可能性条件に対する十分条件を構築する．また，識別可能と判断されたモデルに対して，多重代入法を実行するアルゴリズムを開発する．

研究②（設定 2：完全データへのモデリングが不要なセミパラメトリック推定量）
無視不可能な欠測値データにおいて完全データのモデリングを必要としないセミパラメトリック推定量を構築する．特に，推定量の漸近分散が最小となるセミパラメトリック漸近有効推定量を提案する．また，その推定量と漸近的に同等なセミパラメトリック多重代入法を構築する．

研究③（設定 3：新しい無視可能性条件の下での多重代入法の開発）
近年発見された新たな無視可能性条件[1]に基づき，問題点②で提案したセミパラメトリック漸近有効推定量を改善する．さらに，そのセミパラメトリック漸近有効推定量と同等の有効性を持つ推定量を提案する．

４．研究成果
研究①（完全データへのモデリングが不要なセミパラメトリック推定量）に関して．
採択論文[4]で設定 1 と 2 における一般的なモデルの識別可能条件を導出し，現在投稿中の論文[2]で，具体的に観測データのモデルが指数型分布属で，欠測データメカニズムがロジスティック分布に従う場合の識別可能性条件を導出した．さらに，多重代入法および Fractional 代入法を提案し，その推定量の漸近的性質を導出した．

研究②（設定 2：完全データへのモデリングが不要なセミパラメトリック推定量）
採択論文[4]で無視不可能な欠測値データにおいて完全データのモデリングを必要としないセミパラメトリック推定量を提案した．図 2 で本提案手法による推定量と他の推定量(CK，RR，RRC，RKI)との比較を箱ヒゲ図により示している．提案推定量は，他の推定量に比べてバイアスおよびばらつきが小さいことが分かる．本研究成果は数理統計学のトップジャーナルである Annals of Statistics 誌に採択された．

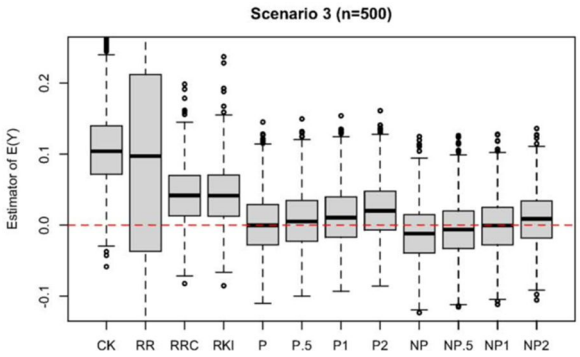

図 1．提案推定量の有効性[4]．
P および NP が提案推定量で，赤色の線が真値．

研究③（設定 3：新しい無視可能性条件の下での多重代入法の開発）
この設定下での多重代入法の構築が，本研究プロジェクトの最終目標であったが，研究期間内に達成することは叶わなかった．今後も継続して研究を進める．

また，本プロジェクト以外にも無視不可能な欠測値データ解析の応用研究として，医学統計データにおけるメタアナリシスに関する研究[3]，地球物理学における本震直後の余震の頻度分布の推測法に関する研究[5]に関する論文も出版した．

以上より，研究③に関する成果を除けば，概ね計画通り研究を進めることができた．

[参考文献]

[1] 狩野 裕 (2014). 日本統計学会誌, **43**, 359-377.
[2] Beppu, K., Morikawa, K. and Im, J. (2022). Imputation with verifiable identification condition for nonignorable missing outcomes. arXiv: 2204.10508.
[3] Huang, Ao, Morikawa, K., Friede, T. and Hattori, S. (2021). Adjusting for publication bias in meta-analysis via inverse probability weighting using clinical trial registries. arXiv: 2109.12526.
[4] Morikawa, K. and Kim, J.K. (2021). Semiparametric optimal estimation with nonignorable nonresponse data. *Annals of Statistics*, **49**, 2991-3014.
[5] Morikawa, K., Nagao, H., Ito, S., Terada, Y., Sakai, S. and Hirata, N. (2021). Forecasting temporal variation of aftershocks immediately after a main shock using Gaussian process regression. *Geophysical Journal International*, **226**, 1018-1035.

５．主な発表論文等

〔雑誌論文〕　計8件（うち査読付論文　4件／うち国際共著　0件／うちオープンアクセス　5件）

１．著者名 Morikawa Kosuke、Nagao Hiromichi、Ito Shin-ichi、Terada Yoshikazu、Sakai Shin’ichi、Hirata Naoshi	４．巻 226
２．論文標題 Forecasting temporal variation of aftershocks immediately after a main shock using Gaussian process regression	５．発行年 2021年
３．雑誌名 Geophysical Journal International	６．最初と最後の頁 1018～1035
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/gji/ggab124	査読の有無 無
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 Tanaka Kenta、Morikawa Kosuke、Katayama Yusuke、Kitamura Tetsuhisa、Sobue Tomotaka、Nakao Shota、Nitta Masahiko、Iwami Taku、Fujimi Satoshi、Uejima Toshifumi、Miyamoto Yuji、Baba Takehiko、Mizobata Yasumitsu、Kuwagata Yasuyuki、Matsuoka Tetsuya、Shimazu Takeshi	４．巻 8
２．論文標題 G20 Summit and emergency medical services in Osaka, Japan	５．発行年 2021年
３．雑誌名 Acute Medicine & Surgery	６．最初と最後の頁 e661
掲載論文のDOI（デジタルオブジェクト識別子） 10.1002/ams2.661	査読の有無 無
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 Morikawa Kosuke、Kim Jae Kwang	４．巻 49
２．論文標題 Semiparametric optimal estimation with nonignorable nonresponse data	５．発行年 2021年
３．雑誌名 The Annals of Statistics	６．最初と最後の頁 2991～3014
掲載論文のDOI（デジタルオブジェクト識別子） 10.1214/21-AOS2070	査読の有無 無
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 Sugasawa Shonosuke、Morikawa Kosuke、Takahata Keisuke	４．巻 31
２．論文標題 Bayesian semiparametric modeling of response mechanism for nonignorable missing data	５．発行年 2021年
３．雑誌名 TEST	６．最初と最後の頁 101～117
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s11749-021-00774-y	査読の有無 無
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 Morikawa, K. and Kim, J. K.	4．巻 未定
2．論文標題 Semiparametric optimal estimation with nonignorable nonresponse data	5．発行年 2021年
3．雑誌名 Annals of Statistics	6．最初と最後の頁 未定
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 Morikawa, K., Nagao, H., Ito, S., Terada, Y., Sakai, S. and Hirata, N.	4．巻 未定
2．論文標題 Forecasting temporal variation of aftershocks immediately after a main shock using Gaussian process regression.	5．発行年 2021年
3．雑誌名 Geophysical Journal International.	6．最初と最後の頁 未定
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/gji/ggab124	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 Sugasawa, S., Morikawa, K. and Takahata, K.	4．巻 未定
2．論文標題 Bayesian Semiparametric Modeling of Response Mechanism for Nonignorable Missing Data.	5．発行年 2021年
3．雑誌名 TEST	6．最初と最後の頁 未定
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 Im, J., Morikawa, K., and Ha, H.-T.	4．巻 144
2．論文標題 A least squares-type density estimator using a polynomial function	5．発行年 2020年
3．雑誌名 Computational Statistics and Data Analysis	6．最初と最後の頁 不明
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.csda.2019.106882	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9件（うち招待講演　0件／うち国際学会　5件）

1．発表者名
森川耕輔，長尾大道，伊藤伸一，寺田吉壱，酒井慎一，平田直

2．発表標題
ガウス過程回帰を用いた本震直後における余震分布の推定

3．学会等名
統計関連学会連合大会

4．発表年
2021年

1．発表者名
森川耕輔，Jae Kwang Kim

2．発表標題
標本調査における包含確率の情報を用いたセミパラメトリック漸近有効推定量の提案

3．学会等名
統計関連学会連合大会

4．発表年
2021年

1．発表者名
森川耕輔，長尾大道，伊藤伸一，寺田吉壱，酒井慎一，平田直

2．発表標題
ガウス過程回帰を用いた本震直後における余震分布の推定

3．学会等名
日本地震学会秋季大会

4．発表年
2021年

1．発表者名
Morikawa, K., Nagao, H., Ito, S., Terada, Y. Sakai, S. and Hirata, N.

2．発表標題
Forecasting temporal variation of aftershocks immediately after a main shock using Gaussian process regression.

3．学会等名
Japan Geoscience Union Meeting（国際学会）

4．発表年
2021年

１．発表者名
Beppu, K. and Morikawa, K.

２．発表標題
On the Verifiable Identification Condition in NMAR Missing Data Analysis.

３．学会等名
The 3rd International Conference on Econometrics and Statistics（国際学会）

４．発表年
2021年

１．発表者名
Beppu, K. and Morikawa, K.

２．発表標題
On the Verifiable Identification Condition in NMAR Missing Data Analysis.

３．学会等名
10th World Congress in Probability and Statistics（国際学会）

４．発表年
2021年

１．発表者名
Morikawa, K., Nagao, H., Ito, S., Sakai, S. and Hirata, N.

２．発表標題
Forecasting temporal variation of aftershocks immediately after a main shock using Gaussian process regression.

３．学会等名
Asia Oceania Geosciences Society（国際学会）

４．発表年
2021年

１．発表者名
Kosuke Morikawa

２．発表標題
Prediction of aftershocks with Gaussian process regression: application to the 2004 Chuetsu Earthquake

３．学会等名
Prediction of aftershocks with Gaussian process regression: application to the 2004 Chuetsu earthquake（国際学会）

４．発表年
2020年

１．発表者名 森川耕輔，寺田吉壱，Jae Kwang Kim
２．発表標題 経験尤度法を用いたNMAR データに対するセミパラメトリック多重代入法
３．学会等名 日本統計関連学会連合大会
４．発表年 2019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			
米国	Iowa State University			
韓国	Yonsei University			