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　３ 年　 ６ 月 １０ 日現在

機関番号：１５４０１
研究種目：若手研究(B)
研究期間：2017～2020
課題番号：１７Ｋ１２６５０

研究課題名（和文）補助変数を用いたモデリング法の開発と応用

研究課題名（英文）Development of modeling methods using auxiliary variables and its application

研究代表者
伊森　晋平（Imori, Shinpei）

広島大学・先進理工系科学研究科（理）・准教授

研究者番号：８０７４７３４５

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：本研究では，興味の対象である主要変数のモデリング精度を向上させるために，補助変数を用いた手法の開発に取り組んだ．潜在変数を含む不完全データにおいて，補助変数を活用して構築されたモデルの良さを情報量規準によって測り，有用な補助変数を選択する手法を開発し，他の規準量との関係を示した．また，補助変数の数が多い場合に用いるスクリーニング法の提案や共変量シフト下での補助変数を選択するための情報量規準の導出も行った．

研究成果の学術的意義や社会的意義
情報資源の有効利用の観点から，補助情報の活用は重要な問題である．しかしながら，補助変数の活用は常に主要変数のモデリング精度を向上させるとは限らず，悪影響を与える可能性もあることから，補助変数の適切な活用が求められる．したがって，本研究で行った有用な補助変数の選択手法など，補助変数を活用する手法の開発は大きな意義があると考えられる．また，関連分野でも一定の研究成果を得ており，今後の補助変数の活用に関する研究が発展する手助けになると期待できる．

研究成果の概要（英文）：This study attempted to develop methods exploiting auxiliary variables in order to improve modeling accuracy of the primary variables. A selection method of useful auxiliary variables was developed in incomplete data analysis, where latent variables are included. The goodness of a model constructed by auxiliary variables is measured by an information criterion. A relationship between the proposed criterion and previous criteria was shown. This study also proposed a screening method used when the number of auxiliary variables are large, and derived an information criterion to select useful auxiliary variables under covariate shift.

研究分野：数理統計学

キーワード：補助変数　変数選択　情報量規準　不完全データ解析

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年，大量のデータを収集し，その処理を行うことが可能となり，興味の対象となる変数に加えて，副次的に他の変数も収集できるようになった．このように興味の対象となる変数を主要変数とよび，主要変数に関連するが直接は興味がないような変数を補助変数とよぶ．補助変数自体が解析の対象ではなくても，主要変数に関する有用な情報を持っていることが期待できる場合，補助変数を活用することは非常に重要であるといえる．

たとえば，回帰モデルにおける説明変数のように補助変数を利用してモデルを構築した場合を考える．このとき，テストデータにおいて補助変数が観測されていないと，構築したモデルをテストデータに適用することは難しく，予測などの場面で実際に利用することができない．テストデータへの適用がモデルを構築する動機であるケースはしばしば考えられ，これは大きな問題である．そこで，補助変数と主要変数の同時分布モデルに着目し，その周辺化によって主要変数のモデルを得ることを考える．この方法では補助変数と主要変数の同時分布における未知パラメータの推定を通して，補助変数のもつ有用な情報を主要変数のモデルに組み込むことが可能となる．また，テストデータにおいて補助変数が観測されない場合にも適用することができる．

しかしながら，補助変数を活用した主要変数の統計的モデリングは常に精度の向上に結びつくとは限らず，補助変数を活用せずに主要変数だけでモデリングした場合に比べても，精度が悪くなることも起こりうる．したがって，補助変数を単純に活用するのではなく，適切な対応が必要である．

２．研究の目的

本研究では，上述の問題点を解決するために，補助変数を活用した主要変数のモデリング法の開発を目指す．

３．研究の方法

（１）補助変数の有用性は主要変数のモデリング精度によって決まる．そこで主要変数のモデル（確率分布）に対してカルバック・ライブラー情報量を基にしたリスク関数を構築し，補助変数を活用することで得られる主要変数のモデルの良さをそのリスク関数で測ることで，補助変数の有用性を定量的に示すことができる．つまり，ある補助変数を活用したモデルのリスク関数の値が，主要変数のみで構築されたモデルのそれよりも小さくなっていれば，その補助変数は有用である．一方で，リスク関数は母集団分布に依存して決まるため未知であり，実際の解析では，その推定量である情報量規準が用いられる．そこで，情報量規準の最小化によって有用な補助変数を選択する方法について研究した．

（２）補助変数が複数ある場合，有用な補助変数を活用するか否かを選択するだけでなく，通常の回帰分析などにおける変数選択問題と同様に，有用な補助変数の組み合わせも選択対象となる．したがって，補助変数の数が多い場合，（１）の情報量規準を利用した有用な補助変数の選択は計算コストの観点から問題が生じる．すなわち，補助変数の数が多いデータに対しては，考えられる補助変数の組み合わせが莫大になるため，全ての組み合わせを計算することは計算時間の観点から困難となる．そこで，補助変数の数が多い場合でも利用可能な計算効率のよい補助変数の選択法の開発を試みた．

（３）補助変数は，未知パラメータの推定に用いる訓練データにおいて観測され，推定したモデルの評価で用いる，あるいは実際にそのモデルを適用するテストデータにおいては欠測する変数として捉えることもできる．このように，訓練データとテストデータの違いに着目した研究分野として，共変量シフト（Shimodaira, 2000）が知られている．共変量シフトは回帰問題の一種であり，共変量（説明変数）の確率分布が訓練データとテストデータの間で変化することを許容している．ただし，共変量を与えた下での目的変数の条件付き分布は変化しないことを仮定している．このような共変量シフトと補助変数の活用を組み合わせた新たな手法を提案した．

４．研究成果

（１）混合分布におけるラベルのような潜在変数を含む，不完全データの解析を考える．潜在変数も主要変数の一部とすると，リスク関数は潜在主要変数と観測主要変数を組み合わせた完全データに対して定められる．このとき，リスク関数の漸近不偏推定量として情報量規準が導出さ

れ，この情報量規準を用いた補助変数の活用手法について研究を行った．
この情報量規準は潜在変数や補助変数がない，通常のモデル選択においてよく用いられている赤池情報量規準(AIC; Akaike, 1974)や不完全データに対してAICを拡張した規準量(Shimodaira and Maeda, 2018)を含む一般的な結果である．また，潜在変数や補助変数がない状況において，汎用的なモデル選択手法として知られるクロスバリデーションと AIC が漸近同等であることが知られている（Stone, 1977）．本研究では，真のモデルの一部を既知とすることで，本研究の設定においてもクロスバリデーションと情報量規準が定数項とスケール倍を除いて，漸近同等であることを示した．しかしながら，真のモデルの一部が既知であるという仮定は実際の利用時には極めて限定的であり，クロスバリデーションの汎用性は失われてしまう．したがって，この設定においては情報量規準の方が有用であると考えられる．

（２）補助変数を用いない通常のモデル選択における計算効率のよい手法として，スクリーニング法がある．（１）の情報量規準を用いた候補モデルのスクリーニング法について研究を行った．具体的には，補助変数をひとつ用いたモデルに対して情報量規準を計算し，その値が小さいものほど有用な補助変数とみなすことで変数に順序づけを行うことを考える．情報量規準の主要項だけでも推定に悪影響を与えるような補助変数を排除することは可能だと考えられるが，補助変数と主要変数が独立である場合には，その補助変数は本来不要であるにもかかわらず，検出することが難しいと懸念される．そのため情報量規準を用いた順序づけを提案している．補助変数を順序づけすることにより，入れ子型の候補モデル集合が構築でき，すべての変数の組み合わせを計算するより格段に候補モデルの数が減少する．したがって，本手法を用いることで計算コストの低下が見込まれる．

（３）共変量シフトにおいては，訓練データとテストデータ間で共変量（説明変数）の分布が異なっているため，最尤法や最小二乗法などの妥当性は保証されない．そこで，訓練データとテストデータでの分布の違いを考慮した重み付き推定法がしばしば用いられる．また，リスク関数はテストデータの従う確率分布に基づき構築される．したがって，（１）で求めた情報量規準を共変量シフト下で利用することは望ましくないと考えられる．そこで，分布の差異を考慮した重み付き推定を行ったときのリスク関数の漸近不偏推定量として情報量規準を導出した．この情報量規準は共変量シフト下での有用な補助変数を選択に利用できる．

上記の研究成果に加えて，補助変数を用いた主要変数のモデリングへの適用には至っていないが，多変量解析や変数選択などの関連分野においてもいくつかの成果を得ている．例えば，多変量線形回帰モデルにおける Cp（Mallows, 1973）タイプの変数選択規準について，目的変数の次元が大きくなるような高次元データに対し，漸近有効性を持つための条件を導出している．さらに，目的変数の次元の大きさに応じて Cp タイプ規準の漸近的性質が変化しうることが示されている．すなわち，目的変数の一部が補助変数であるような状況を考えれば，補助変数を活用することによりモデル選択手法の漸近的性質に影響を与えることが示唆されたものとなった．このように，補助変数を用いたモデリングと関連分野のさらなる融合が期待できる．

【文献】

Akaike, H. A new look at the statistical model identification. IEEE Trans. Autom. Control 1974, 19, 716-723.

Mallows, C. L. Some comments on Cp. Technometrics. 1973, 15, 661-675.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. J. Stat. Plan. Inference 2000, 90, 227-244.

Shimodaira, H. and Maeda, H. An information criterion for model selection with missing data via complete-data divergence. Ann. Inst. Stat. Math. 2018, 70, 421-438.

Stone, M. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. J. R. Stat. Soc. Ser. B Methodol. 1977, 39, 44-47.

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　1件／うちオープンアクセス　2件）

１．著者名 Imori Shinpei、von Rosen Dietrich	４．巻 36
２．論文標題 On the mean and dispersion of the Moore-Penrose generalized inverse of a Wishart matrix	５．発行年 2020年
３．雑誌名 The Electronic Journal of Linear Algebra	６．最初と最後の頁 124～133
掲載論文のDOI（デジタルオブジェクト識別子） 10.13001/ela.2020.5091	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 該当する

１．著者名 Tanabe, R., Kamo, K., Fukui, K., Imori, S.	４．巻 49
２．論文標題 Statistical inference for estimating the incidence of cancer at the prefectural level in Japan	５．発行年 2019年
３．雑誌名 Japanese Journal of Clinical Oncology	６．最初と最後の頁 481～485
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/jjco/hyz033	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 Imori, S., Shimodaira, H.	４．巻 21
２．論文標題 An Information Criterion for Auxiliary Variable Selection in Incomplete Data Analysis	５．発行年 2019年
３．雑誌名 Entropy	６．最初と最後の頁 281～281
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/e21030281	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 －

〔学会発表〕　計11件（うち招待講演　6件／うち国際学会　7件）

１．発表者名 Shinpei Imori
２．発表標題 Asymptotic efficiency of Cp-type criterion in high-dimensional multivariate linear regression models
３．学会等名 3rd International Conference on Econometrics and Statistics (EcoSta 2019)（招待講演）（国際学会）
４．発表年 2019年

１．発表者名
Shinpei Imori

２．発表標題
Convergence rate of importance weighted orthogonal greedy algorithm

３．学会等名
The 11th ICSA International Conference（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
Imori, S., von Rosen, D., Oda, R.

２．発表標題
Growth curve model with bilinear random coefficient

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

４．発表年
2018年

１．発表者名
Imori, S.

２．発表標題
Orthogonal greedy algorithm under covariate shift

３．学会等名
The International Conference on Trends and Perspectives in Linear Statistical Inference（招待講演）（国際学会）

４．発表年
2018年

１．発表者名
Imori, S.

２．発表標題
Convergence rate of OGA under covariate shift

３．学会等名
The 5th Institute of Mathematical Statistics Asia Pacific Rim Meeting（招待講演）（国際学会）

４．発表年
2018年

１．発表者名
前田 篤刀，伊森 晋平，下平 英寿

２．発表標題
回帰モデルにおける補助変数を活用した推定精度の向上

３．学会等名
2018年度 統計関連学会連合大会

４．発表年
2018年

１．発表者名
伊森 晋平

２．発表標題
モデル選択手法とその漸近的性質

３．学会等名
RIMS研究集会「高次元量子雑音の統計モデリング」

４．発表年
2018年

１．発表者名
伊森 晋平，寺田 吉壱，下平 英寿

２．発表標題
補助変数のスクリーニング法について

３．学会等名
2017年度 統計関連学会連合大会

４．発表年
2017年

１．発表者名
井戸 貴大，伊森 晋平，下平 英寿

２．発表標題
共変量シフトにおける補助変数を用いた予測と情報量規準

３．学会等名
2017年度 統計関連学会連合大会

４．発表年
2017年

１．発表者名 Imori, S.
２．発表標題 Screening procedure for auxiliary variables in the Gaussian mixture model
３．学会等名 Conference of the International Federation of Classification Societies（招待講演）（国際学会）
４．発表年 2017年

１．発表者名 Ido, T., Imori, S., Shimodaira, H.
２．発表標題 An information criterion for prediction with auxiliary variables under covariate shift
３．学会等名 The 10th Conference of the IASC-ARS（国際学会）
４．発表年 2017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			
スウェーデン	SLU			
その他の国・地域　台湾	National Tsing Hua University			