様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　６　月　　３　日現在

機関番号：　１２６１３
研究種目：　基盤研究(C)（一般）
研究期間：　2013～2015
課題番号：　２５４００１９７

研究課題名（和文）高次元変動係数モデルにおける変数選択に関する研究

研究課題名（英文）Research on variable selection for high-dimensional varying coefficient models

研究代表者
本田　敏雄（HONDA, Toshio）

一橋大学・大学院経済学研究科・教授

研究者番号：３０２６１７５４

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）：高次元変動係数モデルに関する変数選択問題について大きな成果を挙げた。特に、変動係数をもつ高次元longitudinal dataに対しては、スクリーニングにより説明変数を十分に減らした上で、group SCADにより半変動係数モデルを特定化する一貫した手法を提案し、理論的な性質と数値的な性質を詳しく調べた。さらにその半変動係数モデルの定数係数の有効推定の問題も扱った。その他には、変動係数モデルに対する前進型のスクリーニング法の提案、変動係数をもつCox回帰モデルの変数選択問題も扱った。

研究成果の概要（英文）：I considered variable selection for high-dimensional varying coefficient models for longitudinal data and obtained some significant results. Specifically, I proposed with collaborators a useful two-step procedure to identify the true semi-varying coefficient model of the longitudinal data. The procedure consists of the nonparametric independence screening and the group SCAD method. Besides, I proposed an efficient estimator of the constant coefficients of the true semi-varying coefficient. I proved its desirable theoretical properties and some numerical properties were also investigated in our collaborative papers. In addition, I also obtained some theoretical results for a forward variable selection procedure for high-dimensional variable coefficient models and Cox models with variable coefficients. Numerical properties were also examined in our collaborative papers.

研究分野：　統計学

キーワード：　変数選択　高次元データ　変動係数モデル　セミパラメトリックモデル　longitudinal data

１．研究開始当初の背景

(1) 研究開始以前より、データ収集の技術の飛躍的な発展により、医学、金融、マーケティングなどの分野で、説明変数の数が非常に多い高次元データが得られるようになっており、それに伴い、高次元データの解析手法の研究も重要になっていた。従来の統計的手法では、説明変数の数が非常に多い高次元データをそのままの形で扱うことはできないが、そのような場合でも、実際には有意な説明変数の数は少数であることが経験的に知られている。
従って様々な形の回帰モデルで、以下のような二段階の変数選択が提案されていた。
第一段階：スクリーニングを行うことにより明らかに有意でない説明変数の数を除く。
第二段階：残された説明変数から有意な説明変数を選択する。

(2) (1)で述べた第二段階においても、AIC や BIC などの伝統的なモデル選択基準をすべての場合に計算し、その結果により変数選択を行うことがかなり困難であるくらい多くの説明変数が含まれる場合が多い。そこで第二段階の方法としては、SCAD や adaptive Lasso のようなペナルティー(Fan and Li(2001)、Zou(2006)など)を尤度関数などに加えた上で回帰係数を推定し、その回帰係数の推定値が 0 か否かなどにより変数を選択するという方法を用いることが提案され、その有効性も確かめられている。この十数年にわたって盛んに研究されている、SCAD や adaptive Lasso を含む Lasso などによる変数選択法の研究はかなりの進展をみせているが、スクリーニング法の研究はまだ始まったばかりであった。

２．研究の目的

(1) 本研究は、高次元変動係数モデルに対するスクリーニング法の提案とその理論的性質の研究を主な目的とした。変動係数モデルは、線形モデルの回帰係数が、低次元の index variable とよばれる年齢、時間などの重要な変数に従って変化するという、線形モデルの基本的な拡張であり、線形モデルの簡潔性とノンパラメトリックモデルの柔軟性という長所を併せ持つモデルである。そして高次元データに対しても、有効な分析ツールになると思われ、このモデルに対するスクリーニング法を提案することは、高次元データ解析への大きな貢献となると考えられた。そのスクリーニング法としては、近年 independence screening 法が提案された。これは一つの説明変数だけに注目し他の変数をすべて無視した周辺モデルでの推定結果により、有意な説明変数の候補を選んでいくという方法で、パラメトリックモデルに対しては Fan and Song(2010)、加法モデルについては Fan, Feng, and Song(2011)の結果がある。本研究に深く関係するのは後者である。変動係数モデルは加法モデルと類似性をもち、加法モデルと同様の結果の成立が予想できた。さらに、予想される定理の証明についても、加法モデルにおけるアイデアとテクニックがある程度適用可能と考えられた。

(2) 通常の高次元変動係数モデル以外に、変動係数を持つ高次元 longitudinal data における変数選択問題、変動係数をもつ高次元 Cox 回帰モデルの変数選択問題についても研究を行うことも目的とした。これらのデータ、モデルは、医学統計で非常によく見られるデータ、モデルであり、近年は遺伝子データの解析の必要性から、高次元の場合の手法の研究が大変重要なものとなっている。また可能であれば、１.(1)で述べた第二段階の手法の研究および変動係数から半変動係数モデルを特定化する方法の研究を行うことも目的とした。

３．研究の方法

(1) longitudinal data に対する変数選択と有効推定法の研究：本研究は当初からの計画通り、国立台湾大の Ming-Yen Cheng 教授、香港バプティスト大の Heng Peng 准教授との国際共同研究として進めた。まず Fan, Feng, and Song(2011)で提案された independence screening 法を、どのような方法で変動係数を持つ高次元の longitudinal data に適用することが可能であるか、ということの検討から研究を開始した。基本的には、相互訪問、電子メールのやりとりにより共同研究を進めた。independence screening 法の理論研究については主として研究代表者が行い、数値的な研究は、途中から研究に加わった国立シンガポール大の Jialiang Li 准教授が主に担当した。その後さらに研究を進展させ、スクリーニング後の第二段階の変数選択の研究、一部の説明変数が定数係数をもつ半変動係数モデルの特定化の研究、半変動係数モデルの定数係数の有効推定の研究を行った。相互訪問、電子メールのやりとりにより共同研究を進めたが、研究代表者は基本的に理論的な研究を担当した。そしてこれらの研究成果をまとめて国際誌に投稿し、国際学会でも発表を行った。

(2) 変動係数モデルに対する前進型変数選択法の研究：本研究は、国立台湾大の Ming-Yen Cheng 教授と、途中からこの国際共同研究に参加した、国立シンガポール大の Jin-Ting Zhang 准教授との共同研究である。Wang(2009)にあるような前進型のスクリーニング法を変動係数モデルに応用することから研究を開始した。相互訪問、電子メールのやりとりにより共同研究を進めたが、研究代表者は基本的に理論的な研究を担当し、Jin-Ting Zhang 准教授が主として数値的な研究を担当した。この研究成果をまとめて国際誌に投稿し、国際学会でも発表を行った。

(3) 変動係数 Cox 回帰モデルにおける変数選択法の研究：高次元の時変動係数 Cox 回帰モデルに関する継続研究であり、フンボルト大の Härdle 教授より電子メールでコメント、アドバイスを受ける形で、研究代表者が理論研究、数値実験等を行った。この研究成果は国際誌に投稿し、国際学会で発表した。

４．研究成果

(1) longitudinal data に対する変数選択と有効推定法の研究：医学統計の分野で重要な longitudinal data の統計解析に関しても、近年のデータ収集技術の向上により、高次元の説明変数を持つデータが多くみられるようになった。また、通常の線形モデルは制約が多いため、より柔軟なデータ解析が可能である変動係数モデル、半変動係数モデルを考察することにした。本研究で扱った longitudinal data は、観測時間、個体あたりの観測数が一様でないもので、理論的な取り扱いが容易でないものである。また longitudinal data は、計量経済学ではパネルデータとよばれ、多くの研究がなされている。本研究では変動係数をもつ高次元の longitudinal data に対して、以下に詳述する三つの重要なテーマについて研究を行った。以下の結果を二本の論文を統計学のトップジャーナル論文（雑誌論文①、③）にまとめ、学会発表①、③の国際学会でも発表した。

①変数選択におけるスクリーニング法の研究：Fan, Feng, and Song(2011) では、independence screening 法を加法モデルに適用しているが、理論的な解析は十分ではなかった。ただし independence screening 法自体は、説明変数一つの marginal model を用いるもので、計算および実際の取り扱いは非常に容易なものである。この研究では、変動係数モデルの特性とスプライン基底の局所的な性質を巧妙に使うことにより、スクリーニング法としての一致性を証明し、選択される変数の上限に関する結果にも厳密な証明を与えた。さらに数値実験によりこの手法の性質を検証した。

②半変動係数モデルの特定化の研究：①で述べたスクリーニングにより、予備推定が可能であるレベルまで十分に説明変数を減らすことができた場合の課題は、真のモデルを特定化することである。特に変動係数モデルでは、一部の説明変数は、関数でなく定数を回帰係数に持つことも十分に考えられる（半変動係数モデル）。そこで、関数を近似するスプライン基底を定数部分と関数部分に分解し、group SCAD の手法と組み合せて、モデル選択について一致性を持つ group SCAD 法を提案した。さらに数値実験によりこの手法の性質を検証した。

③半変動係数モデルの定数係数の有効推定の研究：longitudinal data に対する半変動係数モデルを特定した後は、係数の再推定を行う必要がある。その際には、定数係数の効率の良い推定を行うことが重要である。ここでは、各個体内の分散共分散行列が観測時間のみに依存するという仮定のもとでの有効推定量を提案し、その有効性を証明した。具体的には、各個体内の分散共分散行列を局所線形推定量で推定し、その分散共分散行列の推定量を用いてスプライン回帰型の推定を行うという斬新な手法により有効推定量を構成した。さらに数値実験によりこの手法の性質を検証した。

(2) 変動係数モデルに対する前進型変数選択法の研究：変動係数モデルは、構造を持つノンパラメトリックモデルの中で最もよく用いられるものの一つで、変動係数モデルのような線形モデルの長所とノンパラメトリックモデルの長所を持ったモデルである。この研究では、高次元の説明変数と index variable があるという設定の下で、前進型のスクリーニング法を提案してその理論的な性質を調べ、高次元変数選択のスクリーニング法としての一致性を持つことを示した。この研究で提案した手法は、線形モデルに対して Wang(2009)において提案された手法のアイデアを、変動係数モデルに適用したものである。本研究では Wang(2009)での理論的な結果を改良し、BIC または EBIC を停止ルールに用いた場合の結果、グラム行列の最大固有値と最小固有値の影響、重要な変数グループとそうでない変数グループが分かれる場合について詳しく調べた。またさまざまな数値実験を行い、停止ルールとしては通常の BIC を用いるのがよいことを示した。この研究結果は、変動係数モデル以外の構造を持つノンパラメトリックモデルの中で最もよく用いられるものの一つである加法モデルにもそのまま応用できる。以上の結果は、統計学のトップジャーナル論文（雑誌論文②）にまとめられ、学会発表②の国際学会等で発表された。

(3) 変動係数 Cox 回帰モデルにおける変数選択法の研究：Cox 回帰モデルは、医学統計の重要なテーマである生存時間解析におけるもっとも一般的なモデルの一つである。そのほかに、経済学、経営学においても、ミクロ計量経済学における求職期間などの duration time の分析、企業の存続期間の分析、クレジットカード保有者に関する分析などでも広く用いられている。現在では、遺伝子解析技術の進歩、データ収集技術の進歩などにより、Cox 回帰モデルが用いられるような場合においても、高次元の説明変数を持つデータが得られるようになっており、高次元の説明変数から適切な説明変数を選択する手法が極めて重要になっている。通常の線形の Cox 回帰モデルの場合には、Lasso、

adaptive Lasso、SCAD などを用いた手法が研究されてきた。しかしながら、より柔軟なデータ解析をするためには、変動係数モデルのような、線形モデルの長所とノンパラメトリックモデルの長所を持ったモデルが必要となる。そこでこの研究では、時間変動型の変動係数モデルと index variable による変動係数モデルについて、SCAD と adaptive Lasso の理論的性質を詳細に調べ、変数選択について一致性を持つことを証明した。数値実験により理論的な結果を検証した。以上の結果は、雑誌論文④の論文にまとめられ、学会発表④の国際学会でも発表された。

＜引用文献＞
Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96, pp.1348-1360.

Fan, J. Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse Ultra-High-Dimensional Additive Models. Journal of the American Statistical Association 106, pp.544-557.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. The Annals of Statistics 38, pp.3567-3604.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. Journal of the American Statistical Association 104, pp.1512-1524.

Zou, H. (2006). The adaptive lasso and its oracle properties, Journal of the American statistical association 101, pp.1418-29.

５．主な発表論文等

〔雑誌論文〕（計４件）

① Ming-Yen Cheng, Toshio Honda, Jialiang Li. Efficient estimation in semi-varying coefficient models for longitudinal/clustered data. The Annals of Statistics, 印刷中(2016). 査読有
http://imstat.org/aos/

② Ming-Yen Cheng, Toshio Honda, Jin-Ting Zhang. Forward variable selection for sparse ultra-high dimensional varying coefficient models. Journal of the American Statistical Association, 印刷中(2016). 査読有
DOI : 10.1080/01621459.2015.1080708

③ Ming-Yen Cheng, Toshio Honda, Jialiang Li, Heng Peng. Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data. The Annals of Statistics, 42(2014), pp.1819-1849. 査読有
DOI: 10.1214/14-AOS1236

④ Toshio Honda, Wolfgang Karl Härdle. Variable selection in Cox regression models with varying coefficients. Journal of Statistical Planning and Inference, 148(2014), pp.67-81. 査読有
DOI: 10.1016/j.jspi.2013.12.002

〔学会発表〕（計８件）

① Toshio Honda. Efficient estimation in semivarying coefficient models for longitudinal/clustered data, IASC-ARS 2015、2015 年 12 月 18 日, シンガポール(シンガポール)

② Toshio Honda. Forward variable selection for sparse ultra-high dimensional varying coefficient Models, Waseda International Symposium "High Dimensional Statistical Analysis for Spatio-Temporal Processes & Quantile Analysis for Time Series", 2015 年 11 月 11 日, 早稲田大学理工学部(東京都・新宿区)

③ Toshio Honda. Nonparametric independence screening and structural identification for ultra-high dimensional longitudinal data, Joint Statistical Meetings 2014, 2014 年 8 月 6 日, ボストン(アメリカ)

④ Toshio Honda. Variable selection in Cox regression models with varying coefficients, The 59th World Statistics Congress, 2013 年 8 月 30 日, 香港(中国)

〔その他〕
ホームページ等
https://hri.ad.hit-u.ac.jp/html/449_profile_ja.html

６．研究組織

(1)研究代表者
本田　敏雄（HONDA, Toshio）
一橋大学・大学院経済学研究科・教授
研究者番号：３０２６１７５４