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　　元 年　　５ 月 ２９ 日現在

機関番号：１２６１３
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ０５２６８

研究課題名（和文）超高次元データに対する説明変数のスクリーニング手法に関する研究

研究課題名（英文）Studies on screening methods for data with ultra-high dimensional covariates

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）　　3,300,000円

研究成果の概要（和文）：本研究では、近年重要性が増している、説明変数の次元が観測値の指数オーダーであるような超高次元データの変数選択問題を考察した。また単純な線形モデルだけでは十分なデータ解析が行えない場合も多いため、構造を持つノンパラメトリックモデルである加法モデルおよび変動係数モデルを扱った。具体的には、加法モデルおよび変動係数モデルの構造を持つCox回帰モデルの変数選択と構造の特定化問題をgroup Lassoにより統一的に扱った。同様に分位点回帰モデルにおいても、一致性を持つadaptive group Lassoによる変数選択と構造の特定化問題も考察した。

研究成果の学術的意義や社会的意義
超高次元データの変数選択問題に関しては多くの研究があるが、変動係数モデルや加法モデルのような構造を持つノンパラメトリックモデルについては研究が遅れていた。特に超高次元の説明変数を持つ変動係数モデルおよび加法モデルから、通常の統計的推測が可能である部分線形変動係数モデルおよび部分線形加法モデルを特定化する問題は未解決であった。この問題を、スプライン基底を、定数部分、線形部分、その他と直交化し、それに応じてgroup Lassoのペナルティーを適宜分割することにより解決した。提案した手法により、通常の回帰、Cox回帰、分位点回帰で、変数選択と構造の特定化問題を扱うことができるようになった。

研究成果の概要（英文）：In this research, we studied variable selection for data with ultra-high dimensional covariates. Recently the importance of such kinds of data has been increasing. There are many data sets of the kind for which linear regression models and their variants are not flexible enough to carry out data analysis. Therefore, we focused on structured nonparametric regression models such as additive models and varying coefficient models. Specifically, we dealt with simultaneous variable selection and structure identification for varying coefficient Cox models by appealing to the group Lasso. Besides, we considered quantile regression models with additive and varying coefficient structures and proposed an adaptive group Lasso method with selection consistency.

研究分野：統計科学

キーワード：超高次元データ　変動係数モデル　加法モデル　Cox回帰モデル　分位点回帰　スプライン基底

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

（１）データ収集技術の進歩により、医学、金融、マーケティングなど様々な分野で、説明変数の次元が観測値の指数オーダーであるような超高次元データが利用可能になっている。説明変数の数が非常に多いとはいえ、実際に有意な説明変数は限られており、変数選択が極めて重要である。しかし説明変数が超高次元であることにより、計算および理論上の制約から、従来の統計的な変数選択の手法は適用できず、新たな手法が数多く提案され、研究開始時点から現在まで、以下の二段階で変数選択を行う形が標準になっている。

第一段階（スクリーニング） Lasso あるいは一変数の marginal model による independence screening などにより、有意である可能性のある説明変数だけ選び、予備推定可能なレベルまで説明変数の次元を減少させる。この段階では有意な説明変数を落とさないことが重要である。

第二段階（一致性を持つ変数選択と推定）予備推定を行い、adaptive Lasso、SCAD などの手法により、実際に有意である説明変数のみを選び、場合により同時にパラメータの推定も行う。

　第二段階においても説明変数は依然として高次元であり、伝統的な統計的変数選択法をそのまま適用することはやはり困難である。以上の手法については、観測値が独立性を持ち、さらに何らかの形で超高次元の線形平均回帰モデルに帰着できるような場合については研究が進み、それらの手法の性質の理解は十分に得られている。またモデルによらないアドホックなスクリーニング手法も数多く提案されていた。

（２）観測値が独立性を持ち、さらに何らかの形で超高次元の線形平均回帰モデルに帰着できる場合についての研究は進んでいたが、生存時間データ、非線形性を持つデータ、時系列データ、分位点回帰モデルなどに関する研究は十分とは言えず、スクリーニング法の研究、一致性を持つ変数選択法の研究とともに、多くの課題が残っていた。

２．研究の目的

（１）本研究では、独立性を持つ通常の超高次元の線形平均回帰モデルに帰着できないモデル、具体的には、加法モデル、変動係数モデルなどを係数部分に持つ、Cox 回帰モデル、分位点回帰モデルに注目して以下（２）および（３）の課題を解決することを目的とした。Cox 回帰モデルは、生存時間解析で最も重要なモデルの一つであり、また分位点回帰モデルは、通常の平均回帰モデルより多くの情報を与える、リスク管理とも関連する重要かつ一般的な分析手法である。

（２）加法モデルの場合には、有意な変数を選ぶ変数選択だけでなく、有意な変数が線形な効果をもつか、あるいは非線形な効果も持つかという、構造の特定化問題（部分線形加法モデルの特定化問題）がある。変動係数モデルにおいても、有意な変数を選ぶ変数選択だけでなく、各説明変数の係数関数が定数であるか否かの構造の特定化問題（部分線形変動係数モデルの特定化問題）が重要である。これらの問題はその重要性にも拘わらず未解決であった。本研究では、この問題に注目し、超高次元の加法モデルおよび変動係数モデルからの、部分線形加法モデルおよび部分線形変動係数モデルの特定化問題の解決を目的とした。以上の部分線形加法モデルおよび部分線形変動係数モデルの特定化問題に関しても、第一段階のスクリーニングと第二段階の一致性を持つ推定法の二つの問題がある。

（３）さらに近年、高次元の構造を維持したまま統計的推測を行うことの重要性が認識されてきた。変数選択では、興味のある変数についても、選択されたかされないかの結果しか得られない。しかしながらこの場合には、興味のある係数の推定を行うことにより信頼区間を構成することができ、その係数の有意性を検討することもできる。当初の目的だけでなく、以上の高次元の構造を維持したままの変動係数モデルに関する統計的推測（de-biased or de-sparsified Lasso など）の問題も研究の目的に加えた。

３．研究の方法

（１）変動係数 Cox 回帰モデルにおける変数選択および構造の特定化に関する研究：本研究は矢部竜太氏との国内共同研究であった。この研究はある意味では研究代表者の継続的な研究テーマであり、研究のアイデアを出すこと、研究計画の立案、および理論面の研究は、主として研究代表者が行った。研究計画の立案および理論研究の過程で矢部氏の意見を求めながら、矢部氏と共同してシミュレーションおよび実証研究を行った。

（２）セミパラメトリック分位点回帰モデルにおける変数選択および構造の特定化に関する研究：本研究は、台湾の Ching-Kang Ing 教授および Wei-Ying Wu 教授との国際共同研究であり、また（１）の研究とも密接に関連する。まず研究代表者が Ching-Kang Ing 教授に分位点回帰におけるセミパラメトリックモデルの重要性と関連する問題を提起し、Ing 教授との直接の面談による議論および電子メールでの密接な意見交換をもとに主として研究代表者が理論面の研究を行った。そして論文のとりまとめなどの責任著者は Ing 教授が務め、シミュレーションおよび実証研究は Wei-Ying Wu 教授が行った。

（３）De-biased Lasso による変動係数モデルの推定に関する研究：本研究は、国際学会に参加する中で研究代表者がアイデアと研究の構想を得たものであり、理論研究およびシミュレーション等は研究代表者が、一部の研究補助をのぞき単独で行ったものである。

４．研究成果

（１）変動係数 Cox 回帰モデルにおける変数選択および構造の特定化に関する研究：Lasso による超高次元 Cox 回帰モデルの説明変数のスクリーニングに関する重要な結果は、Huang et al.(2013)に与えられている。しかしながら、Cox 回帰の文脈でよく用いられている、変動係数モデル、加法モデル、時変係数モデルの場合の理論的な結果は、その重要性にも拘わらずあまり得られていなかった。雑誌論文②において、研究代表者と共同研究者の矢部氏は、group Lasso による説明変数のスクリーニングを提案し、以下の重要な研究成果を挙げた。そしてこれらの研究成果を国際学会において発表した（学会発表⑤、⑥）。

① 上記の三つのモデルを理論面で統一的な手法で扱い、オラクル不等式などの、Lasso に関するの標準的な理論的結果を得た。この理論的な結果は、これら三つのモデルについて、個々のモデルについて別々の論文を書く研究者も多いが、その必要はないことを示唆している。

② スプライン基底を、定数部分、線形部分、その他と直交化し、それに応じて group Lasso のペナルティーを適宜分割することにより、変数選択だけでなく、構造の特定化に関するスクリーニングも同時に行うことができることを示した。

③ シミュレーションおよび実データへの応用を行い、提案した手法の小標本での挙動の良さと実用性を示した。

（２）セミパラメトリック分位点回帰モデルにおける変数選択および構造の特定化に関する研究：線形分位点回帰モデルに Lasso を適用した場合の、オラクル不等式などの基本的かつ重要な結果は、Belloni and Chernozhukov(2011)に与えられている。変動係数モデル、加法モデルに対する Lasso に関する結果もある程度は得られているので、本研究では第二段階目の、変数選択および構造特定化に関して一致性を持つ推定量について考察し、研究代表者と共同研究者の Ing 氏および Wu 氏は、以下の重要な結果を得た。そしてこれらの研究成果を国際学会において発表した（学会発表②、③、④）。

① adaptive group Lasso よりも広いクラスである、一般的な重み付きの group Lasso について、変動係数モデルと加法モデルを統一的な手法で考察し、変数選択および構造の特定化に関する一致性を証明した。

② ここでも、スプライン基底を、定数部分、線形部分、その他と直交化し、それに応じて group Lasso のペナルティーを適宜分割することにより、変数選択と構造の特定化が同時に行えることを示した。

③ BIC 型の情報量基準を提案し、重み付き group Lasso のチューニングパラメーターの選択問題を解決した。この結果は Lee et al.(2014)のセミパラパラメトリックモデルへの拡張である。ここでは分位点回帰を扱っているため、通常よりも丁寧な理論的な解析が必要であった。

④ シミュレーションおよび実データへの応用を行い、提案した手法の小標本での挙動の良さと実用性を示した。

（３）De-biased Lasso による変動係数モデルの推定に関する研究：国際学会に参加し、多くの講演を聞き、de-biased or de-sparsified Lasso（Javanmard and Montanari(2014)、van de Geer et al.(2014)、Zhang and Zhang(2014)）等により、高次元の構造を維持したまま統計的推測を行うことの重要性を認識し、この de-biased Lasso による変動係数モデルの推定に関する研究を開始した。変動係数モデルに関しては、この種の結果は得られていなかった。そして一定の成果を得て、〔その他〕の項のディスカッションペーパーにまとめた上で、国際学会において発表した（学会発表①）。この研究は、開始したばかりで現在進行中のものである。変数選択では、興味のある変数においても、選択されたかされないかの結果しか得られないが、興味のある係数関数を含むすべての係数関数の推定を行うことが可能で、それらに関する信頼区間を構成することにより、その係数関数の有意性を検討することもできる。具体的には以下の成果を得て学会発表をした。

① 変動係数モデルの係数関数の de-biased group Lasso 推定量を構成した上で、その理論的な性質を示した。

② シミュレーションを行い、提案した手法の小標本での挙動の良さと実用性を示した。

＜引用文献＞

Belloni, A and Chernozhukov, V. ℓ1-penalized quantile regression in highdimensional sparse models. Ann. Statist. 39(2011), pp.82-130.

Huang, J., Sun, T., Ying, Z., Yu, Y., and Zhang, C.-H. Oracle inequalities for the Lasso in the Cox model. Ann. Statist. 41(2013), pp. 1142-1165

Javanmard, A. and Montanari, A. Confidence intervals and hypothesis testing for

high-dimensional regression. J. Machine Learning Research 15(2014), pp.2869-2909.

Lee, E. R., Noh, H., and Park, B. U. Model selection via Bayesian information criterion for quantile regression models. J. Amer. Statist. Assoc. 109(2014), pp.216-229.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Statist. 42(2014), pp.1166-1202.

Zhang, C.-H. and Zhang, S. S. Confidence intervals for low dimensional parameters in high dimensional linear models. J. Royal Statist. Soc. Ser. B 76(2014), pp.217-242.

５．主な発表論文等

〔雑誌論文〕（計２件）

① Toshio Honda, Ching-Kang Ing, Wei-Ying Wu. Adaptively weighted group Lasso for semiparametric quantile regression models. Bernoulli, 印刷中(2019). 査読有
http://www.bernoulli-society.org/index.php/publications/bernoulli-journal/bernoulli-journal

② Toshio Honda, Ryota Yabe. Variable selection and structure identification for varying coefficient Cox models. Journal of Multivariate Analysis 161(2017), pp. 103-122. 査読有
DOI:10.1016/j.jmva.2017.07.007

〔学会発表〕（計 10 件）

① Toshio Honda. The de-biased group Lasso estimation for varying coefficient models. CMStatistics 2018, 2018 年 12 月 15 日, ピサ（イタリア）
② Toshio Honda. Adaptively weighted group Lasso for semiparametric quantile regression models. The 5th IMS-APRM, 2018 年 6 月 29 日, シンガポール（シンガポール）
③ Toshio Honda. Adaptively weighted group Lasso for semiparametric quantile regression models. CMStatistics 2017, 2017 年 12 月 16 日, ロンドン（イギリス）
④ Toshio Honda. Adaptively weighted group Lasso for semiparametric quantile regression models. European Meeting of Statisticians 2017, 2017 年 7 月 24 日, ヘルシンキ（フィンランド）
⑤ Toshio Honda.Variable selection and structure identification for varying coefficient Cox models. EcoSta 2017, 2017 年 6 月 16 日, 香港（中国）
⑥ Toshio Honda.Variable selection and structure identification for varying coefficient Cox models. CMStatistics 2016, 2016 年 12 月 11 日, セビリア（スペイン）

〔その他〕

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

Toshio Honda. The de-biased group Lasso estimation for varying coefficient models Discussion Paper #2018-04, Graduate School of Economics, Hitotsubashi University.

６．研究組織

(1)研究協力者

研究協力者氏名：Ing, Ching-Kang
所属研究機関：National Tsing Hua University(Taiwan)
部局名：Institute of Statistics
職名：Director
研究協力者氏名：Wu, Wei-Ying
所属研究機関：National Dong Hua University(Taiwan)
部局名：Department of Applied Mathematics
職名：Assistant Professor
研究協力者氏名：矢部　竜太
ローマ字氏名：YABE, Ryota
所属研究機関：信州大学
部局名：経法学部
職名：講師

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。