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　２７　年　　４　月　２２　日現在

機関番号：　１４４０１
研究種目：　若手研究(B)
研究期間：　2012～2014
課題番号：　２４７００２７７

研究課題名（和文）正則化法によるスパース推定と超高次元データへの応用

研究課題名（英文）Sparse estimation via regularization and its application to ultra high-dimensional data

研究代表者

廣瀬　慧（Hirose, Kei）

大阪大学・基礎工学研究科・助教

研究者番号：４０６０９８０６

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：L1型正則化法にもとづくスパース推定は，近年取得される超高次元データを計算機を使って高効率に推定することの出来る手法で，ここ数年で急速に発展してきた．本研究で行った成果は主に２つある．１つ目は，回帰モデルにおけるL1正則化法において，モデルに含まれる調整パラメータを効率的に選択することのできるアルゴリズムを提案した．２つ目は，因子分析モデルにおけるL1正則化法を提案し，従来法である因子回転との関係性を明らかにした．どちらの研究においてもソフトウェアパッケージ（Rパッケージ msgps，fanc）として公開しており，だれでもフリーで使うことができる．

研究成果の概要（英文）：Sparse estimation via L1 regularization enables us to analyze the ultra high-dimeniosnal data, and it has been rapidly developed in the recent years. I made two research achievements in this project. First, I developed an efficient algorithm that estimates the tuning parameter in sparse regression modeling. Second, I proposed the L1 regularization procedure in the factor analysis model, and investigated the relationship between the traditional rotation technique and regularization procedure. I showed that the regularization is viewed as the generalization of the rotation method. I have also made a free software package in R (msgps and fanc).

研究分野：　統計科学

キーワード：　スパース推定　L1正則化　モデル選択　因子分析

１．研究開始当初の背景

L_1 タイプのノルム制約を課す正則化推定法は，モデルの推定と変数選択を同時に行うことができ，超高次元･大規模データから有効な情報を高効率に抽出する有効な方法として近年様々な分野で用いられている．本研究では，主に，線形回帰モデルにおける L_1 正則化法のモデル選択と，L_1 正則化法の因子分析・構造方程式モデリングへの応用を考えた．具体的な内容は以下のとおりである．

(1) 正則化項が微分不可能な L_1 型正則化法は，推定値を解析的に導出することが困難となる．それゆえ，このモデリングの過程において，

(i) 推定値を高速に計算するアルゴリズム，

(ii) 適切な調整パラメータの選択

が本質的となる．(i) の推定値を計算するアルゴリズムは，すでに数多く提案されている．一方で，(ii)の調整パラメータの選択は，モデルの選択･評価と捉えることができるが，L_1 型正則化法に対するモデル選択の研究は当時あまりされていなかった．というのも，モデル評価基準を導出するためには，解析的な推定量を求める必要があるが，L_1 型正則化法は解析的な推定値の導出が困難であったためである．

(2) L_1 正則化法は，2000 年代前半までは，主に線形回帰モデルのような比較的シンプルなモデルに用いられていたが，最近は，グラフィカルモデルや主成分分析など，比較的複雑なモデルに対しても使われるようになってきた．一方で，心理学でよく用いられる因子分析モデルに対して，L_1 型正則化法はほとんど研究されていない．実際，申請者がこの研究費を申請する少し前に先行研究を調べたが，当時はそのような論文が１本しか見つからず，そこで使われているソフトウェアを実際に使うと，計算スピードが遅く，かつ解析者が主観的に決めなければいけないパラメータが存在していた．

２．研究の目的

（1）まず，線形回帰モデルというシンプルなモデルに対して，推定値を効率的に計算し，かつ適切な調整パラメータを選択できる方法を提案しようと試みた．それを実現するためには，解析的な推定値を計算せずに，推定値と調整パラメータを同時に計算することのできる高速なアルゴリズムを提案することが重要だと考えた．また，線形回帰モデルは極めてシンプルなモデルで，実データに当てはまらないことも多いので，この方法を非線形回帰モデルや関数データ解析へと拡張することも目標とした．

(2) 因子分析モデルに対する L_1 型正則化法の確立を目指した．そのために，なぜそもそもスパース正則化法を用いなければいけないのかを考えた．具体的にはまず，従来のスパース推定法である因子回転と新しいスパース正則化法の間に，どのような関係性があるのかを明らかにしようと考えた．また，推定値を高速に計算するアルゴリズムを開発し，提案法をだれでも使えるようにソフトウェアRのパッケージとして公開することを目標とした．

３．研究の方法

(1)Generalized Path Seeking(GPS)アルゴリズムに着目した．GPS アルゴリズムは，汎用性があり，様々なロス関数，罰則項に適用でき，推定値を高速に計算することができる．さらに，最急降下法に基づくアルゴリズムなので，１回のステップで，推定値の更新式が陽に書けるという特徴がある．この「陽に書ける」というところが，他のアルゴリズムにない重要なポイントである．この特徴を生かすことにより，モデル評価基準を「陽に」計算することが可能となる．

しかし，このアルゴリズムだと，サンプルサイズの二乗のオーダーに比例して計算時間がかかることがわかった．そこでまず，推定値とモデル評価基準を同時に求めるのでなく，推定値をまず全部計算してから必要な計算結果を行列としてメモリに確保し，その行列を QR 分解することによって，パラメータを効率的に計算することができる方法を提案した．

(2) 因子回転は，因子負荷行列のスパース推定法として数十年も用いられてきたスタンダードな方法である．実は，因子回転はデータへのフィッティングを行わず，最適化問題を解くだけである．そのため，データへのフィッティングを行う正則化法とデータへのフィッティングを行わない因子回転をそのまま比較しても何も結果が出てこないと考えた．

データへのフィッティングとして，これまでのスタンダードな方法は最尤法である．そのため，従来は，最尤法でモデルを推定し，その後で因子の解釈の後で因子回転を行うという 2 段階法であった．本研究では，この 2 段階推定法と新しいスパース正則化法の関係性を調べた．

また，L_1 正則化法では，パラメータの推定値を解析的に計算することは難しいので，EM アルゴリズムと座標降下法を組み合わせたアルゴリズムが有効であると考えた．さらに，パラメータの初期値の決め方，更新アップデートの計算量の評価など，詳細にわたって効率的に計算するための工夫をこらした．

４．研究成果

(1) 提案法を数値シミュレーションで検証したところ，既存手法よりも，予測するときのリスクを小さくできていることが確認できた．また，実際にサンプルサイズが大きい時（具体的に，説明変数の次元数 40，サンプルサイズ 500），QR 分解を使った方法と使わない方法の計算時間を比較すると，QR 分解を使ったほうが約 150 倍も速くなることがあ

った．

また，この提案法をだれでも使えるように，ソフトウェアRのパッケージmsgpsを公開した．このパッケージを使うことにより，lasso，adaptive lasso，log ペナルティに対する正則化最小二乗推定値とモデル評価基準を容易に計算することができる．

(2) 従来法である2段階推定と，提案法である1段階推定をそのまま比較するのは困難に見えたが，「因子負荷行列が回転の不定性を除いて一意である」という条件を加えると，実は，スパース推定法は2段階推定の一般化であることが理論的にわかった．さらに，スパース推定は，因子回転よりもスパースな因子負荷行列を推定できることもわかった．数値実験を行ったところ，たしかに正則化法は因子回転よりもスパースに推定でき，さらに因子回転よりも安定して高次元データのスパース推定をできることを示すことができた．

しかしながら，この方法を実データに適用すると，適切な因子を抽出されないケースが多かった．その原因を探ったところ，因子間に相関がある場合に因子間に相関がないと仮定して推定すると，第一因子がスパースにならないという現象が起きることが理論的に分かった．そこで，因子間相関を仮定してパラメータを推定する方法を考えた．因子間相関の推定には，準ニュートン法を用いたが，初期値への依存性があるものの，初期値を因子間相関なしの推定値にすれば，安定して推定できることがわかった．

また，提案法をRパッケージとして公開した．このパッケージには，推定した因子負荷行列を分かりやすく可視化したグラフィカルツールがある．このグラフィカルツールは，分析者が主観的にスパースなモデルを選び，適合度を用いてそのモデルの正しさを検証するという，今までにない解析法を可能とする．

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５件）

[1] Hirose, K., Kim, S., Kano, Y., Imada, M., Yoshida, M. and Matsuo, M.
Full information maximum likelihood estimation in factor analysis with a lot of missing values.
Journal of Statistical Computation and Simulation, in press. (査読あり)
DOI:10.1080/00949655.2014.995656

[2] Hirose, K. and Yamamoto, M.
Estimation of an oblique structure via penalized likelihood factor analysis.
Computational Statistics and Data Analysis, **79**, 120-132, 2014. (査読あり)
doi:10.1016/j.csda.2014.05.011

[3] Hirose, K. and Yamamoto, M.
Sparse estimation via nonconcave penalized likelihood in a factor analysis model.
Statistics and Computing, in press. (査読あり)
DOI 10.1007/s11222-014-9458-0

[4] Hirose, K., Tateishi, S. and Konishi, S.
Tuning parameter selection in sparse regression modeling.
Computational Statistics and Data Analysis, **59**, 28-40, 2013. (査読あり)
doi:10.1016/j.csda.2012.10.005

[5] Hirose, K., and Higuchi, T.
Creating facial animation of characters via MoCap data.
Journal of Applied Statistics, **39**(12), 2583-2597, 2012. (査読あり)
DOI:10.1080/02664763.2012.724391

〔学会発表〕（計 15 件）

[1] Hirose, K. and Yamamoto, M.
Extension of Rotation Technique via Penalization in Factor Analysis Model.
International Conference on Advances in Interdisciplinary Statistics and Combinatorics (AISC 2014). Greensboro, USA. October, 2014

[2] Hirose, K., and Yamamoto, M.
Lasso-type penalized maximum likelihood factor analysis.
Joint Statistical Meeting 2013. Palais des congrès de Montréal, Canada.
August, 2013.

[3] Hirose, K., Tateishi, S. and Konishi, S.
Regularization Parameter Selection in Convex and Non-Convex Penalized Least Squares Estimation.
Joint Statistical Meeting 2012. San Diego Convention Centre in San Diego.
August, 2012.

〔その他〕
ソフトウェアパッケージ msgps
http://cran.r-project.org/web/packages/msgps/index.html

ソフトウェアパッケージ fanc
http://cran.r-project.org/web/packages/fanc/index.html

６．研究組織

(1)研究代表者

廣瀬　慧（HIROSE Kei）
大阪大学 大学院基礎工学研究科 助教
研究者番号：４０６０９８０６

(2)研究協力者

山本　倫生（YAMAMOTO, Michio）
京都大学 医学(系)研究科(研究院) 助教
研究者番号：５０７２１３９６