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

平成　３０　年　　６　月　　４　日現在

機関番号：　１２６０８
研究種目：　若手研究(B)
研究期間：　2015～2017
課題番号：　１５Ｋ１５９４６

研究課題名（和文）スパース正則化法による複雑高次元データ解析法の確立

研究課題名（英文）Analysis of complex and high dimensional data via sparse regularization techniques

研究代表者

片山　翔太（Katayama, Shota）

東京工業大学・工学院・助教

研究者番号：５０７４２４５９

交付決定額（研究期間全体）：（直接経費）　　2,600,000円

研究成果の概要（和文）：スパース正則化法のアイデアを駆使し，高次元かつ複雑な構造を持つデータの解析法について研究を行った．本研究では特に，外れ値構造とグループ構造に着目し研究を進めた．前者については，結果変数に外れ値が混入する場合の，線形回帰分析について研究を行い，回帰係数のロバストかつスパースな推定量を導出した．また，セルワイズな外れ値を持つ大規模データ行列から，変数間の条件付き独立性を推定する手法も提案した．後者については，層別線形回帰モデルを考え，結果変数全体に影響を与える共変量と，層ごとに影響を与える共変量を同時に特定する方法を構築した．

研究成果の概要（英文）：Some researches on complex and high dimensional data have been conducted via sparse regularization techniques. This research particularly focuses on dealing with outliers and exploiting a group structure. On the former case, robust and sparse linear regression analyses have been proposed when responses may be corrupted. An estimation technique of conditional independences among large dimensional variables also has been proposed under cell-wise corruption of data matrix. On the latter case, a simultaneous detection method of both covariates that entirely and partially affect responses has been proposed in the context of stratified linear regression.

研究分野：　多変量解析

キーワード：　高次元データ　スパース正則化　ロバスト推測　グラフィカルモデル

１．研究開始当初の背景

高次元データ（変数の次元が標本サイズに比べて大きい）に対する統計解析の研究は，その需要の高さから近年盛んに行なわれている．特に最近では，外れ値の混入やグループ構造などの，より複雑なデータに対する解析法が求められている．そのようなデータを扱うためには通常，想定する統計モデルの複雑化を行う．ところが高次元データにおいては，モデルの複雑化はパラメータ数の膨大化を招いてしまう．

２．研究の目的

本研究では，膨大なパラメータの中から重要なもののみ抽出できるスパース正則化法のアイデアを駆使し，複雑高次元データに対する統計解析法の確立を目指す．そのために，「統計理論」と「効率的計算」のふたつの視点から研究を行う．つまり，ただ単に解析ができるというだけではなく，高次元データに適した形での理論的妥当性を保証するとともに，効率よく計算できることまで目指す．本研究では特に，外れ値構造やグループ構造を持つ高次元データを扱う．遺伝子情報やコンピュータビジョンなど挙げられるように，多くのデータがこのような構造を持っており，応用上で特に重要であると考えられる．具体的な統計モデルとしては，線形回帰モデルやガウシアングラフィカルモデルを考える．

３．研究の方法

（１）外れ値構造．本研究ではふたつの種類の外れ値混入パターンを考える．ひとつは何らかの余分が真の状態に混入してしまったケースである．この場合，もし余分な部分をデータから取り除くことができれば，クリーンなデータが得られる．このようなケースでは，She and Owen(2011)や Nguyen and Tran(2013)を始めとして，外れ値への対処をスパースモデリングに帰着する研究が行なわれている．つまり，外れ値を表すパラメータを新たに導入し，それをスパース正則化法を用いて推定している．このモデリングを土台として，結果変数に外れ値が混入し得る場合の高次元線形回帰分析の研究を進める．高次元データの状況，すなわち，共変量の個数が大きい場合は，回帰係数にもスパース性(いくつかの係数が完全に 0)を想定することは自然である． そのため Lasso，SCAD，MCP などのスパース正則化法を用いて回帰係数と外れ値を推定する．計算アルゴリズムは交互最適化法や近接勾配法を利用する．

もう一方の外れ値混入パターンは，計測機器の誤作動や記入ミスなどで生じるような，外れ値が何の情報も持っていないケースである．この場合は，外れ値と思われるデータの影響を出来るだけ取り除かなければならないため，Fujisawa and Eguchi(2008)で提案されたγ-divergence を用いて推定を行う．

（２）グループ構造．例えば遺伝情報解析においては，生物の進化の過程によって標本はいくつかのグループに分かれている．このような状況でスパース線形回帰を行うと，グループに関する重要な情報を取りこぼしてしまう．枝分かれのように生物が進化してきたと考えると，結果変数全体に影響を与える共変量と，グループ毎に影響を与える共変量がそれぞれ存在すると考えるのは妥当であり，これらを特定したい．そこで，層別線形回帰モデルを考え，グループ毎の回帰係数パラメータを，スパース正則化法を用いて近づけることでこれを達成する．差がないと推定された箇所が全体に影響を与える共変量に対応する．

４．研究成果

（１）外れ値構造．最初の外れ値混入パターンに関しては，She and Owen(2011)と同様に，以下のモデルを考える：

$$y = X\beta + w + \varepsilon$$

ここで，wが外れ値を表すパラメータであり，これを取り除けば従来の回帰モデルである．本研究では先ず，回帰係数に対して Lasso 制約，外れ値に対して Lasso 制約，L0 制約，SCAD 制約などを課して，パラメータのスパース推定を行った．実際にはβとwを交互に最適化する計算アルゴリズムになっている．回帰係数の更新には座標降下法などの効率的なアルゴリズムが利用可能であり，外れ値パラメータの更新は解析的に書くことができる．この意味で効率的計算は達成できた．一方で統計理論に関しては，実際にアルゴリズムで得られる解(出力)に関する誤差評価とサポートの復元性を与えた．得られた誤差評価は，アルゴリズムから生じる計算機的誤差と，モデルから生じる確率的誤差の和と解釈することができる．特に，アルゴリズムの繰り返し数を増やすと計算機的誤差は 0 に収束し，また，確率的誤差は外れ値がない場合の Lasso の誤差評価と定数倍を除いて一致した．つまり，外れ値wは正しく取り除かれることが理論的に示されている．以上の研究は論文誌 Statistica Sinica に発表している．

一方で，通常の線形回帰では，Lasso 制約よりも SCAD や MCP などの非凸制約の方が経験的に良い結果を返すことが知られている．そこで，回帰係数に対するスパース制約関数の一般化も行った．上記の研究は，統計理論の導出において，回帰係数の Lasso 制約が本質的であった．その凸性から，各更新において大域的最適解が保証でき，これを用いて統計理論を導出していた．非凸制約の場合は大域性が保たれず，そのため一般化は自明ではない．近接勾配法を用いて回帰係数βを更新し，最終的に得られる解(出力)に対して誤差評価を与えた．交互最適化の内部でさらに計算アルゴリズムを用いているため，上述の計

算機的誤差の部分に，近接勾配法による局所解への収束率が新たに生じる結果となった．本研究は国内・国際学会で報告し，論文誌に投稿中である．

もうひとつの外れ値混入パターンに関しては，ガウシアングラフィカルモデルのロバスト推定を目標に研究を進めた．従来的には，データ行列の行(サンプル)にひとつでも外れ値が混入してしまうと，その行全ての影響を取り除くように推定していた．しかし高次元データでこれを行ってしまうと，情報の損失が著しい．そこで本研究では，セルワイズな外れ値の影響を取り除けるようなロバスト推定法を与えた．データに正規性を仮定すると，グラフィカルモデルは共分散行列から推定できることに着目し，各変数のペアに対してγ-divergence を用いて共分散をロバスト推定し，共分散行列を構成した．その上で Graphical Lasso(Friedman et al.,2007)，Neighborhood selection(Meinshausen and Buhlmann,2006)，CLIME(Cai et al.,2011)を適用している．以上の結果は論文誌 Stat に発表している．

（2）グループ構造．当初，複数の回帰パラメータに対して，Fused Lasso(Tibshirani et al.,2005)のアイデアを基に，それぞれを近づける制約を課した推定法を考えていた．各ペアに対して差分 $\beta_i - \beta_k$ に Lasso 制約を課した推定法である．しかし予備実験を行ってみると，ほとんどの差分が 0 にならず，実用には耐えられなかった．これを解決するため，Adaptive Lasso(Zou,2006)の考え方をもとに，データ依存の重みを制約に付与した推定法を新たに構築した．当初の推定量を初期推定量とし，それが十分小さい場合は対応する重みを 0 に近づけ，大きな場合は対応する箇所を適切に重み付けている．Alternating Direction Method of Multiplier を用いて効率的計算アルゴリズムを導出し，推定量の誤差評価やサポートの復元性を示した．理論結果から，標本サイズが変数の次元よりも小さな場合でも良い誤差レートを達成しており，また，サポートの復元も可能である．この意味で高次元データでも適切に機能することが示されている．以上の研究は，より一般的な形で，国内学会で発表しており，論文誌にも投稿準備中である．

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 2 件）

① Shota Katayama, Hironori Fujisawa, Mathias Drton. Robust and sparse Gaussian graphical modelling under cell-wise contamination. Stat, volume 7, e181, 2018, 査読有.
② Shota Katayama, Hironori Fujisawa. Sparse and robust linear regression: an optimization algorithm and its statistical properties. Statistica Sinica, volume 27, pages 1243-1264, 2017, 査読有.

〔学会発表〕（計 7 件）

① 片山翔太. Support recovery of adaptive generalized lasso under high dimensionality. 平成 29 年度科学研究費シンポジウム「大規模複雑データの理論と方法論，及び，関連分野への応用」, 2017 年 12 月.
② 片山翔太. Adaptive generalized lasso for high dimensional linear regression model. 統計関連学会連合大会，2017 年 9 月.
③ 片山翔太. 外れ値にロバストな非凸スパース正則化回帰. 統計関連学会連合大会, 2016 年 9 月.
④ 片山翔太. 非凸スパース制約に基づくロバスト線形回帰アルゴリズムとその性質. 統計サマーセミナー, 2016 年 8 月.
⑤ Shota Katayama. Robust non-convex penalized linear regression with algorithmic and statistical convergence. IMS-APRM, June 2016 Hong Kong.
⑥ 片山翔太，藤澤洋徳. Robust high dimensional regression with algorithmic convergence and support recovery. 統計関連学会連合大会 Wakimoto Memorial Session (招待講演)，2015 年 9 月.
⑦ 片山翔太, 藤澤洋徳. スパース正則化法によるロバスト高次元回帰. 大規模統計モデリングと計算統計 II (招待講演), 2015 年 9 月.

〔図書〕（計 0 件）

〔産業財産権〕

○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：

国内外の別：

〔その他〕
ホームページ等
http://www.me.titech.ac.jp/~miyalab/katayama/index.html

６．研究組織
(1)研究代表者
片山 翔太（Katayama, Shota）
東京工業大学・工学院・助教
研究者番号：50742459

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：

(4)研究協力者
（　　）