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 １６ 日現在

機関番号：６２６０３
研究種目：基盤研究(C)（一般）
研究期間：2013～2015
課題番号：２５３３００４９

研究課題名（和文）スパース正則化による判別とグループ化に基づく意思決定システムの構築

研究課題名（英文）Implementation of decision support systems based on statistical classification/regression models with sparse regularization

研究代表者

川崎　能典（Kawasaki, Yoshinori）

統計数理研究所・モデリング研究系・教授

研究者番号：７０２４９９１０

交付決定額（研究期間全体）：（直接経費）　　3,800,000円

研究成果の概要（和文）：統計モデルにおいて目的変数を説明する候補変数が高次元の状況下で，適切な変数選択・グルーピングを行う円滑閾値型推定方程式法を援用し，高精度の予測を主眼に置いた意思決定システムの構築を行った．電話による直接顧客マーケティングデータの解析では，預金契約に至りやすい顧客の特徴の把握と予測性が，変数選択法ないしグルーピング法によってどう異なるかを統計的に検証した．また，ゲノムワイド関連研究において個人の疾患発症リスクを高精度に予測するモデルの構築を行った．また，高相関の説明変数が多い状況で，複数の競合モデルを残すことで誤って重要な変数を捨てることを防ぐ方法も提案した．

研究成果の概要（英文）：We promoted the smooth-threshold estimation equations (STEE) to develop a prediction model with high accuracy even in high dimensional regression problems. In the analysis of bank telemarketing data, STEE as well as other methods like lasso reveals the features of the customers who are likely to subscribe bank deposits. We also conducted comparative study of prediction accuracy among competing methods changing the variable selection methods, with or without grouping. We also worked on Genome Wide Association Study where we developed a variant of STEE, smooth-threshold multivariate genetic prediction model to forecast the personal risk related to a certain disease. Finally we proposed a method to retain multiple rival models where some of the explanatory variables are highly correlated.

研究分野：統計科学

キーワード：スパース正則化法　分類・パターン認識　変数選択　変数グルーピング　高次元交互作用　リスク解析　多重共線性　予測モデル

１．研究開始当初の背景

背景には正則化法による変数選択理論の進展がある．90 年代に入ると統計科学においては，モデルが含む未知母数に L1 型のペナルティ(ゼロ方向への絶対値型縮小制約)を課して推定する方法，すなわちスパース正則化法が変数選択の新手法として盛んに研究されるようになった．

年代順に記すと，Frank and Friedman (1993, Technometrics)による bridge regression, Tibshirani (1996, JRSS(B))による Least Absolute Shrinkage and Selection Operator (LASSO), Fan and Li (2001, JASA)による Smoothly Clipped Absolute Deviation penalty (SCAD), Zou and Hastie (2005, JRSS(B))による elastic net, Zou (2006, JASA)による Adaptive LASSO が挙げられる．

研究分担者の植木は，Ueki (2009, Biometrika)で計算負荷の少ないスパース正則化法を提案した．このアイディアは高次交互作用の検証を含む変数グルーピング問題に拡張可能であることがわかり，初期成果は Ueki and Kawasaki (2011, Electronic J. Statist.)として発表した．本研究課題ではこの方向性を推進し，判別・分類・パターン認識における予測変数や因子の探索に利用できる汎用的方法に関する研究を行うことを企図した．

２．研究の目的

上述の研究開始当初の背景に照らし，応答がカテゴリカルな変数である一方，説明変数候補が膨大で，それらの組み合わせで得られる交互作用項は爆発的に多い状況を考える．このようなデータセットに対し，高次元分割表解析による情報抽出法を経由して，有効な予測変数の探索法を構築することを目標に定めた．また，スパース正則化法を利用したリスク因子剪定法の確立により，これが効率的かつ実用的な変数減少法を与えることを明らかにする．凸最適化が不要な効率的方法に基づき，変数選択と同時に変数のグルーピングも可能となるところに特色がある．

更に，予測の観点から手法の有効性を示す実データ解析への取組と実装を通じて，特徴量の多いデータセットに対して円滑閾値型推定方程式による接近が有効であることを示すことが目標である．

３．研究の方法

本研究は円滑閾値型推定方程式(smooth-threshold estimating equation, STEE)に立脚し変数選択・自動グルーピングを適用あるいは拡張するので，まず STEE の基本的な着想を説明する．

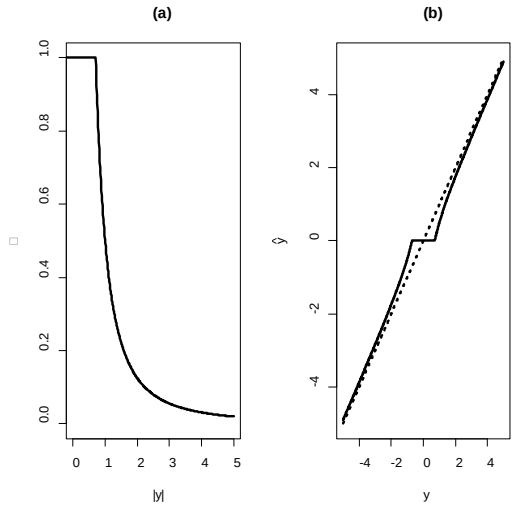
罰則付き損失関数として$\mathrm{L}(\theta)+\sum_{j=1}^{d}\rho_j(\theta_j)$を考える．ここで$\rho_j(\theta_j)$は j 番目のパラメータの非負罰則関数とする．従って設定としては Zou (2006)の adaptive lasso と同様である．

我々は罰則関数として，ある既知の重み$\mathrm{w_j}\in[0,\infty]$に基づき$\rho_j(\theta_j)=w_j\theta_j^2/2$という定式化を採用する．もし$\mathrm{w_j}=\infty$であれば，罰則項が損失関数に優越するので，$\theta_j$はゼロとなる．パラメータが d 次元であるとすると，d 本の推定方程式を解くことで解は得られる．

一見すると罰則関数は 2 次制約であり，スパース正則化というよりリッジタイプの制約と思われるかもしれない．しかし以下に述べるように，STEE では閾値則が$\mathrm{w_j}$の中に含まれている．

新たにパラメータ$\delta_j\in[0,1]$を導入し，$\mathrm{w_j}=\delta_j/(1-\delta_j)$という形でパラメータを取り直す．推定方程式はおのずと$(1-\delta_j)\partial L(\theta)/\partial\theta_j+\delta_j\theta_j=0,(j=1,\dots,d)$と書き換えられる．j 番目の推定方程式で$\delta_j=1$($\mathrm{w_j}=\infty$に対応)であれば$\theta_j=0$となり，スパース解に帰着する．Adaptive lasso の議論を借用することで，δ_jはθ_jの推定に先立ってデータから決めておくことができる．我々の問題においては例えば$\hat{\theta}_j^{(0)}$を$\sqrt{\mathrm{n}}$一致性を持つような何らかの初期推定量として$\hat{\delta}_j=\min\left(1,\lambda/\left|\hat{\theta}_j^{(0)}\right|\right)^{1+\gamma}$と決めることが自然である．チューニングパラメータ$(\lambda,\gamma)$は BIC 型の規準で選択すればよい．

STEE の閾値関数の挙動とそのときの$\hat{\delta}_j$の推移を以下の図に示す．右パネルで横軸が大きいところから徐々に左に向かって(0 に向かって)動いたとき，左パネルで徐々に$\hat{\delta}_j$の値が 1 に近づき，1 に達したところでスパース解$(\hat{y}=0$生成している様子がわかる．この図と Fan and Li (2001)の Figure 2 (c)を見比べれば，STEE が本質的に SCAD と同じような効能を持つことが理解できる．



４．研究成果

(1) マーケティングにおける予測モデル構築法

電話による直接顧客マーケティングを利用した定期預金の販売に関するデータを分析し，

スパース正則化を利用した自動変数グルーピング法の数値的側面の研究を完成させた．結果は雑誌論文②として公刊された（学会発表④⑧⑨⑩⑫も関連）．

預金契約に至りやすい顧客の特徴の把握と予測性が，変数選択法ないしグルーピング法によってどう異なるかを統計的に検証した．比較の対象として，LASSO，Elastic-Net，SCAD，MCP に加えて，我々の提案する Smooth-Threshold Estimating Equation 法(STEE 法)を取り上げた．推定用データと予測検証用データに分ける分割をランダムに 10 回実行して，受信者操作特性(ROC)曲線下の面積(AUC)で予測精度を比較した．

実験結果から観察されるところでは，STEE でグルーピングを考慮すると，平均的には AUC の値はやや下がるが，複数回の実験における最大 AUC は，STEE でグルーピングを考えた場合において散見された．非ゼロの係数が推定された回数を計測してみると，当初 STEE と他の方法では変数の絞り込みの傾向が大きく異なっているように見えたが，チューニングパラメータの選択を交差検証型から BIC 型に切り替えると，LASSO や SCAD 等も STEE の変数選択版とほぼ同様の傾向を示すに至った．

また，STEE 変数選択版では多くの変数が落とされる一方，STEE グルーピング版のほうが非ゼロで生き残る変数は多い．単独では残らない変数でも，グループ化されると生き残る場合が多いと解釈できる．グループ化された変数の多くは，職業や学歴等被験者の社会的地位に関するもので自然な解釈が与えられる．

データ自体は既に入念な変数選択の後にリポジトリに提供されたデータと思われ，グループ化が大きな予測精度の向上をもたらす事例とはならなかったが，カテゴリカルな応答変数に対して説明変数候補が膨大で，交互作用項が組み合わせ爆発的に多い状況で，スパース正則化法を利用したリスク因子剪定法が効率的かつ実用的な変数減少法を与えることを明らかにできた．

(2) 疾患発症リスク予測モデリング

ゲノムワイド関連研究（Genome Wide Association Study；GWAS）での重要な目標のひとつに，個人の疾患発症リスク（あるいは血糖値や血圧値などの連続的臨床値）の予測がある．そのためには複数のリスクバリアントと共変量を入力とする数学的予測モデルの構築が求められる．これまで様々な接近法による試行錯誤が行われたが，現在に至るまで GWAS データを用いた予測モデルは，実用に耐える性能（一般的には AUC で計測される）を示さなかった．

雑誌論文①として公刊した研究（学会発表③⑤⑦も関連）では，全ゲノムシークエンス(WGS)データにおいても高速で，かつ高精度な予測モデリングを実現する統計手法として，新しいスパースモデリング手法（STMGP; smooth-threshold multivariate genetic prediction）を提案した．この STMGP について，様々な遺伝的モデルを仮定したシミュレーションを行った結果，最新の遺伝子スコア法やポリジーン予測法といったこれまで知られるどの手法よりも迅速かつ高い予測精度を示すことが明らかとなった．

さらに，米国 Alzheimer's Disease Neuroimaging Initiative (ADNI)が提供するアルツハイマー病の WGS データに適用した結果，シミュレーション研究で示唆された通り，既存手法よりも高い性能のリスク予測モデルを構築することができた．

(3) 多重共線下における複数の回帰モデル選択

スパース正則化法は近年活発に研究され，ゲノムデータ等の高次元データ解析に適用されているが，背後のメカニズムからかけ離れた結論が導かれ，再現性にも欠けるという現象も多々報告されている．

高次元データになるほど多重共線性は避けられない問題であるが，類似した説明変数が混在している状況では，データのゆらぎに依存して特定の変数がたまたま有意に出ているに過ぎない（従って再現性がない）ことは，回帰モデルの構造の幾何的解釈から自明である．

雑誌論文④（学会発表⑪も関連）では，多重共線条件下での回帰分析を念頭に，モデル選択のように一つのモデルに結論を絞るのではなく，結果の解釈可能性を担保することを目的に，複数の競合的モデルを手元に残す方法を提案した．

そのため独自に提案した基準が標準化更新度(Standardized Update, SU)である．SU と適合度尺度(GOF)を使い，以下の手続きでモデルを探索する．

- 各変数をひとつだけ含んだ p 個の一変量モデルからスタート(並列的探索)
- GOF 基準を満たしたモデルにはお墨付きを与えて終了
- 満たさないモデルについて，それ以外の変数を各ステップでひとつずつ取り込み，GOF 基準を満たすまで深掘り
- 取り込みの可否は SU により判断(取り込める変数がなければ良いモデルはなかったとして終了)

詳細は雑誌論文④を参照されたい．

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 Ueki, M., Tamiya, G., Smooth-threshold multivariate genetic prediction with unbiased model selection, Genetic Epidemiology, 40,

2016, 233-243, [査読有] DOI: 10.1002/gepi.21958

② Kawasaki, Y., Ueki, M., Sparse predictive modeling for bank telemarketing success using smooth-threshold estimating equations, Journal of Japanese Society of Computational Statistics, 28, 2015, 53-66. [査読有] DOI: 10.5183/jjscs.1502003_217

③ Okamura K, Ohe R, Abe Y, Ueki M, Hozumi Y, Tamiya G, Matsunaga K, Yamakawa M, Suzuki T, Immunohistopathological analysis of frizzled-4-positive immature melanocytes from hair follicles of patients with Rhododenol-induced leukoderma, Journal of Dermatological Science, 80, 2015, 156-158. [査読有] DOI: 10.1016/j.jdermsci.2015.07.015

④ Ueki, M. and Kawasaki, Y., Multiple choice from competing regression models under multicollinearity based on standardized update, Computational Statistics and Data Analysis, 63, 2013, 31-41. [査読有] DOI: 10.1016/j.csda.2013.01.019

⑤ Notsu, A., Kawasaki, Y., Eguchi, S., Detection of heterogeneous structures on the Gaussian copula model using predictive power entropy, ISRN Probability and Statistics, 2013, 1-10. [査読有] DOI: 10.1155/2013/787141

〔学会発表〕(計１２件)

① 植木優夫, 嶋村海人, 川野秀一, 小西貞則, 田宮元,「超高次元スパース回帰法によるゲノムデータ解析」, 岡山大学 IPSR × 九州大学 IMI × 理研 CSRS シンポジウム 生命データ科学による新たな社会的価値の創造～医療,農業,環境分野における役割と作物設計への応用, 2016 年 2 月 23 日, 理化学研究所 横浜キャンパス(神奈川).

② 植木優夫,「ヒトゲノムデータから見た統計科学」, 遺伝学と統計学における数理とモデリング, 2016 年 1 月 25 日, 政策研究大学院大学 (東京).

③ Ueki, M., Tamiya, G., Multivariate smooth-threshold genetic prediction with unbiased model selection, East Asia Regional Biometric Conference 2015, 2015 年 12 月 22 日, Conference Hall, Kyushu University Station-I for Collaborative Research (福岡).

④ Kawasaki, Y. and Ueki, M., Sparse predictive modeling for bank telemarketing success using smooth-threshold estimating equations, 8th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2015), 2015 年 12 月 12 日, London (U. K.).

⑤ Ueki, M., Tamiya, G., Smooth-threshold multivariate genetic prediction with unbiased model selection, 広島統計談話会, 2015 年 12 月 4 日, 放射線影響研究所 (広島).

⑥ 植木優夫, 田宮元,「円滑閾値型推定方程式による遺伝的予測」, 生命科学データ解析の方法論と健康科学への応用, 2015 年 10 月 16 日,東京大学医科学研究所(東京).

⑦ 植木優夫, 田宮元,「円滑閾値型推定方程式による遺伝的予測」,2015 年度統計関連学会連合大会, 2015 年 9 月 9 日, 岡山大学津島キャンパス(岡山).

⑧ Kawasaki, Y., Variable selection and grouping by smooth-threshold estimating equations, Statistical Computing Asia 2015, 2015 年 7 月 1 日, Taipei (The Republic of China).

⑨ 川崎能典, 「円滑閾値型推定方程式によるリスク因子の探索法とその応用」, 応用統計学会 2015 年度年会, 2015 年 3 月 14 日, 京都大学芝蘭会館稲盛ホール (京都市).

⑩ Kawasaki, Y., Predictive modeling in socio-economic data using smooth-thresholding, International Conference on Statistical Analysis of Large Scale High Dimensional Socio-Economic Data, 2014 年 11 月 6 日, 東北大学 (仙台市).

⑪ Kawasaki, Y. and Ueki, M.: Keeping competitive regression models on hand - Standardized update and its application, IMS-ASC 2014, 2014 年 7 月 7 日, Sydney (Australia).

⑫ 川崎能典, 植木優夫, 「スパース正則化に基づく変数選択・グルーピングとその応用」, 第 8 回日本統計学会春季集会(招待講演), 2014 年 3 月 8 日, 同志社大学今出川キャンパス(京都市).

〔図書〕(計１件)

① 北川源四郎, 田中勝人, 川崎能典〔監訳〕(T. Subbarao, S. Subarao, C. R. Rao 編)「時系列解析ハンドブック」,2016 年, 朝倉書店, 全 788 ページ.

〔その他〕
プレスリリース
雑誌論文①の出版に伴い, 東北大学メディカルメガバンク機構 (研究分担者・植木優夫氏の前所属先, 当該論文共著者田宮元教授所属先) より「全ゲノムシークエンシングデータ

を用いた発症リスク予測モデリングのための
手法を開発」と題したプレスリリースが,
2016 年 4 月 15 日に行われた.
本報告書執筆時点での参考ウェブサイト：
https://www.tohoku.ac.jp/japanese/newimg
/pressimg/tohokuuniv-
press20160415_02web.pdf

６．研究組織
(1)研究代表者
川崎　能典（Kawasaki, Yoshinori）
統計数理研究所・モデリング研究系・教授
研究者番号：70249910

(2)研究分担者
植木　優夫（Ueki, Masao）
久留米大学・バイオ統計センター・講師
研究者番号： 10515860

赤司　健太郎（Akashi, Kentaro）
学習院大学・経済学部・教授
研究者番号： 50610747