統計数理研究所 75 周年記念事業（2019）
©2019 統計数理研究所

スパース正則化法によるリスク要因の探索

Identifying Risk Factors by Sparse Regularization

モデリング研究系 川崎 能典 （Yoshinori Kawasaki）

1. 円滑閾値型推定方程式による変数選択・グルーピング

情報通信技術やデータ計測技術等の発展に伴い，大規模なデータが蓄積されるに至って久しい．統計科学においては高次元データとしばしば言及される状況は，典型的には個体に付随して観測される属性が多数に及んでいる場合を指し，しばしば個体数に比べて属性数が遙かに上回る．こうした状況下で回帰分析を行う場合，「説明変数候補は多数得られているが，目的変数と関連性のある要因はごく少数である」という制約を置くのが現実的である．それを実現したのが LASSO (Least Absolute Shrinkage and Selection Operator) とその変種であり，一般にスパース正則化，L^1 正則化，スパース推定法などと言われる．

ペナルティつき損失関数を一般に $L(\theta)+\sum_{j=1}^{d}\rho_j(|\theta_j|)$ と記す．ここで ρ_j は j 番目のパラメータ θ_j に関する非負のペナルティ関数であり，$\rho_j(|\theta_j|)=\lambda_j|\theta_j|$ なら adapive LASSO である．ここで $\rho_j(|\theta_j|)=w_j\theta_j^2/2$ としつつも，$\delta_j\in[0,1]$ によって $w_j=\delta_j/(1-\delta_j)$ と取り直すことで，解くべき推定方程式は $(1-\delta_j)\partial L(\theta)/\partial\theta_j+\delta_j\theta_j=0$ $(j=1,\ldots,d)$ となる．$\delta_j=1$ は $w_j=\infty$ に対応し，このとき $\theta_j=0$ に帰着する．これを円滑閾値型推定方程式 (Smooth-Threshold Estimating Equation, Ueki (2009)) と呼び，以下 STEE と略す．

Ueki and Kawasaki (2011) は，STEE の方法論を変数のグルーピングにも拡張し，スパース変数選択とグルーピングを半自動的に行う方法を提案した．応用例としては，気管支疾患の判別や与信スコアリングの問題を取り上げている．Kawasaki and Ueki (2015) は，電話によるマーケティングデータを例に，STEE 法の性能を他のスパース正則化法と比較している．

2. 多重共線性と標準化更新度

STEE 法は，初期推定量の良さに依存している．初期推定量は飽和重回帰やリッジ回帰から構成するが，データ間に多重共線性が強ければ，有意な変数は偶発的に生き残っているに過ぎない可能性がある．そこで，ひとつのモデルを信頼するのではなく，たまたま選ばれなかったかもしれない真の因果変数も拾い上げる方法の構築が必要となる．

Ueki and Kawasaki (2013) は，線形回帰モデルの枠組みにおいて，飽和回帰モデルからの相対差で適合度基準を設定し，その基準を満たすモデルを前進選択法で探索することで，複数の「説明力同等」なモデルを手元に残す方法を提案した．手順は以下の通り．1) まず各変数を 1 つだけ含んだ p 個の単回帰モデルから出発し並列的に探索，2) 適合度基準を満たしたモデルにはお墨付きを与えて終了，3) 満たさないモデルについては，それ以外の変数を各ステップで 1 つずつ取り込み，適合度基準を満たすまで深掘り，4) 現在のモデルの変数添字集合を C とし，いま変数 k を加えたときのあてはまりの改善度を標準化更新度 (Standardized Update)

$$SU_{k,C}=\frac{||y-X_C\hat{\beta}_C||^2-||y-X_{C\cup\{k\}}\hat{\beta}_{C\cup\{k\}}||^2}{||y-\bar{y}1||^2-||y-X\hat{\beta}||^2},k\notin C$$

で測り，SU がある閾値を超えた時に変数 k を採用する．

数値実験の結果，完全多重共線性の下でも *SU* を使う提案手法は偽陽性・偽陰性ともに小さく，真の回帰関数に関連している変数の組合せを高精度で発見できた．一方 Elastic Net は全般的に偽陽性率が高く，完全多重共線性の下では偽陰性も高いことが示された．

3. 効果がマスクされた変数の探索

多重共線性が深刻な説明変数群を使って推定された線形飽和重回帰モデルでは，有意な説明変数は偶発的に有意になっているに過ぎない可能性がある．一方，飽和重回帰モデルで有意にならず，かつ目的変数との周辺相関がないと思われる説明変数でも，特定の説明変数集合を伴って推定されれば有意になることがある．

Ueki, Kawasaki and Tamiya (2017) では，効果が他の変数にマスクされている説明変数を効率的に探索する方法を提案している．飽和回帰モデルが推定可能な状況であっても，部分回帰的全探索は計算負荷が高い．提案手法は，説明変数・被説明変数を合わせた変数群内の全てのペアで計算した相関係数を基に変数間の接続性を双方向グラフで表現し，特定の説明変数から目的変数までの最短経路をダイクストラ法で求めることで，効果を浮かび上がらせる共変量集合を特定する．この方法は，ケースの次元より説明変数の次元が大きい状況下で，かつ説明変数群に多重共線性が潜んでいる状況でも適用可能である．(ただし上掲論文中の応用例では，対象に関する知見から，変数群を独立なブロックに分けることが妥当性を持つので，各ブロックごとに計算している．)

謝　辞

本稿で紹介した一連の研究は，植木優夫博士 (執筆時点では理化学研究所革新知能統合研究センター所属) との共同研究である．リスク解析戦略研究センター草創期の 2008 年に特任研究員として着任した同氏から，Ueki (2009) の草稿について討論させて頂いたのを契機に，その後は公募型共同利用 (25–共研–1018，26–共研–1014，27–共研–1013，28–共研–1011，29–共研–1009，30–共研–1013) が一連の研究成果につながっていった．ここに記して感謝申し上げたい．

参 考 文 献

Ueki, M. (2009). A note on automatic variable selection using smoothhreshold estimationg equations, *Biometirka*, **96**, 1005–1011.

Ueki, M. and Kawasaki, Y. (2011). Automatic grouping using smooth-threshold estimating equations, *Electronic Journal of Statistics*, **5**, 309–328.

Ueki, M. and Kawasaki, Y. (2013). Multiple choice from competing regression models under multicollinearity based on standardized update, *Computational Statistics and Data Analysis*, **63**, 31–41.

Kawasaki, Y. and Ueki, M. (2015). Sparse predictive modeling for bank telemarketing success using smooth-threshold estimating equations, *Journal of Japanese Society of Computational Statistics*, **28**, 53–66.

Ueki, M., Kawasaki, Y. and Tamiya, G. (2017). Detecting genetic association through shortest paths in a bi-directed graph, *Genetic Epidemiology*, **41**, 481–497.