# スパース正則化に基づく変数のグルーピング

## 川崎　能典　モデリング研究系　准教授

**概要:** 統計モデルに冗長な変数が含まれているとすれば，分析者としてはそれらの係数は0になってほしいし，同様の意味で，変数の周辺効果が同じものは係数をまとめられればパラメータを節約でき，推定精度の向上に資する．本研究では，滑らかな閾値化を伴う推定方程式の枠組みで，変数のグループ化と選択(スパース解の生成)を自動的に実現する方法を提案する．この方法にはいわゆるoracle propertyがあり，実装にあたっては凸最適化は不要である．[本研究は，植木優夫氏（山形大学医学部）との共同研究である．]

### 1. 滑らかな閾値化に基づく変数選択とグルーピング

パラメータに関する罰則付き損失関数 $L(\theta)+\sum_{j=1}^{d}\rho_j(|\theta_j|)$ を考える．ここで$\rho_j$は$j$番目のパラメータに関する非負の罰則関数である．罰則は既知の重み$w_j \in [0,\infty]$を使って$\rho_j(\theta_j)=w_j\theta_j^2/2$と定める．もし$w_j=\infty$であれば罰則が評価関数の全てになるので，必然的に$\theta_j$はゼロに帰着する．この問題の解は次の$d$本の推定方程式から得られる．

$$\partial L(\theta)/\partial\theta_j + w_j\theta_j = 0 \ (j=1,\ldots,d)$$

$\delta_j \in [0,1]$によって$w_j=\delta_j/(1-\delta_j)$とパラメータを取り直すと，上の推定方程式に対して以下の等価な表現が得られる．

$$(1-\delta_j)\partial L(\theta)/\partial\theta_j + \delta_j\theta_j = 0 \ (j=1,\ldots,d)$$

$j$番目の推定方程式で$\delta_j=1(w_j=\infty$に対応)なら$\theta_j=0$と結論される．即ちスパース解である．

しかし，実際には$\delta_j$(あるいは$w_j$)はデータから何らかの形で決めなければならない．ここでは，adaptive LASSOの議論を借りて

$$\hat{\delta}_j = \min(1, \lambda/|\hat{\theta}_j^{(0)}|^{1+\gamma})$$

で与える．ここで$\hat{\theta}_j^{(0)}$はフルモデルの推定量などによる初期推定量であり，$\sqrt{n}$一致性をもつことを仮定する．このルールに含まれるチューニングパラメータ$(\lambda,\gamma)$はBIC型の規準で選ぶことができる．

以上の方法は文献[1]によるものであるが，罰則項が2次関数であるにもかかわらず，漸近的に閾値化を行う滑らかな関数が重み$w_j$の中に隠れている．従って，凸最適化を避けながらスパース解を手にすることが可能で，計算上有利である．

この枠組みは変数のグルーピングに直接適用できる．パラメータ$\theta_j$と$\theta_k$に対して重み$\hat{w}_{jk}=\hat{\delta}_{jk}/(1-\hat{\delta}_{jk})$をルール$\hat{\delta}_{jk}=\hat{\delta}_{jk}(\lambda,\delta)=\min(1,\lambda/|\hat{\theta}_j^{\rm ini}-\hat{\theta}_k^{\rm ini}|^{1+\gamma})$で与えることにして，罰則項を

$$h(\theta)=\sum_{j=1}^{d}\sum_{k>j}^{d}\hat{w}_{jk}(\theta_j-\theta_k)^2/2,$$

とすればよい．変数$j$と変数$k$がグループ化されて，なおかつ係数がゼロ(スパース解)となるようにするには，

$$h(\theta)=\sum_{j=1}^{d}\sum_{k>j}^{d}\hat{w}_{jk}(\theta_j-\theta_k)^2/2+\sum_{k>0}^{d}\hat{w}_{0k}\theta_k^2/2.$$

とすればよい．このグルーピング法は一致性を持ち([2]定理3.1)，パラメータの推定値は漸近的に正規分布に従うと同時に，その分散共分散行列はoracle propertyを持つ([2]定理3.2)．また，チューニングパラメータの選択にあたっては一致性を持つBIC規準を導出できる([3]定理3.3)．

### 2. 応用：信用スコアリング

ドイツ南部のある銀行の顧客1,000人の与信データに対して，交互作用も含めたスコアリングモデルをあてはめ，その中で変数のグルーピングと選択を行う．データはミュンヘン大学のウェブサイトから入手した．元々のデータは20種類の顧客属性を含んでいるが，主効果のみのロジットモデルによる初期解析から，以下の7つの変数に絞って定式化を探索する．

- $H_1$: 当座預金の状態 (無/良/悪の3カテゴリー)
- $H_3$: 与信月数 (連続変量)
- $H_4$: 与信額 (連続変量)
- $H_5$: 過去の与信案件での支払い (良/悪の2カテゴリー)
- $H_6$: 利用目的 (私用/事業用の2カテゴリー)
- $H_7$, $H_8$: 性別 (男/女), 結婚状態(独居/それ以外)を表すダミー

ここではカテゴリカル変数$H_1,H_5,H_6,H_7,H_8$に関してだけ交互作用を考える．例えば$H_1$と$H_5$の間で考えるとするとしよう．$H_1$での無/良/悪の3カテゴリーを便宜上no/good/badで，$H_5$での良/悪の2カテゴリーを便宜上good/badで表し，以下のようにダミー変数を生成する．$X_1=I(H_1=\text{no},H_5=\text{bad}), X_2=I(H_1=\text{good},H_5=\text{good}), X_3=I(H_1=\text{good},H_5=\text{bad}), X_4=I(H_1=\text{bad},H_5=\text{good})$, and $X_5=I(H_1=\text{bad},H_5=\text{bad})$．ここでは，$I(H_1=\text{no},H_5=\text{good})$をベースライン効果として，計$(3\times2-1)=5$個のダミー変数を考える．

これを全ての組み合わせで考えると，結果的に38個のダミー変数が生成される．それらを$X_1,\ldots,X_{38}$と記すと，出発点となるモデルは

$$\text{logit}\{P(y=1|X)\}=\beta_0+\sum_{j=1}^{38}X_j\beta_j+H_3\beta_{39}+H_4\beta_{40}.$$

となる．($y=0$ or 1は，非事故 or 事故を表すダミー変数である．) 変数のグルーピングと選択の結果は以下の通り．

| 選択された変数 | グループ番号 | 係数 | $P$値 |
|---|---|---|---|
| $H_1=\text{good}, H_5=\text{good}$ | 1 | $-0.37$ | $<0.001$ |
| $H_1=\text{good}, H_6=\text{private}$ | 1 | | |
| $H_1=\text{good}, H_7=\text{man}$ | 1 | | |
| $H_1=\text{good}, H_7=\text{woman}$ | 1 | | |
| $H_1=\text{good}, H_8=\text{not live alone}$ | 1 | | |
| $H_1=\text{no}, H_5=\text{bad}$ | 2 | 0.76 | 0.003 |
| $H_5=\text{bad}, H_6=\text{professional}$ | 2 | | |
| $H_1=\text{no}, H_6=\text{private}$ | 3 | 0.33 | $<0.001$ |
| $H_1=\text{no}, H_6=\text{professional}$ | 3 | | |
| $H_1=\text{bad}, H_7=\text{woman}$ | 3 | | |
| $H_1=\text{no}, H_7=\text{woman}$ | 3 | | |
| $H_5=\text{bad}, H_8=\text{live alone}$ | 3 | | |
| $H_6=\text{professional}, H_7=\text{man}$ | 3 | | |
| $H_6=\text{professional}, H_8=\text{live alone}$ | 3 | | |
| Intercept | Not grouped | $-1.82$ | $<0.001$ |
| $H_3$ | Not grouped | 0.034 | $<0.001$ |
| $H_4$ | Not grouped | 0.000029 | 0.38 |

グループ1の変数は係数が負である．これは事故確率を下げることを意味するので，良い顧客の特徴付けになっている．全て当座預金の状態が何らかの意味で良いことと，他の条件の組み合わせである．性別は関係ない(変数$H_7$による条件付けは縮退している)ものの，与信履歴がクリーンであること，私用目的，独居でない等の条件は全て同じ効果を持っていることが結論される．

グループ2は高リスク因子，グループ3はややリスクを高める因子と解釈できる．生成したダミー変数のうち残りの24変数は言ってみればグループ4だが，それらは係数0のグループとしてまとめられ，変数選択で落とされている．

### 参考文献

[1] Ueki, M. (2009), A note on automatic variable selection using smooth-threshold estimating equations, *Biometrika*, 96, 1005–1011.

[2] Ueki, M. and Kawasaki, Y. (2011) Automatic grouping using smooth-thresholding estimating equations, *Electronic Journal of Statistics*, Vol. 5, 309-328.