多重共線性を考慮した変数選択手法の提案

田村　隆太

東京農工大学大学院工学府情報工学専攻（現：株式会社オクトーバー・スカイ）
指導教員：宮代隆平 東京農工大学准教授

1. はじめに

本研究では，回帰分析におけるモデル作成手法の一つである多重共線性を考慮した変数選択に対して，数理計画法を用いた厳密解法の提案を行う．

回帰分析を行う際に，説明変数間に線形な従属関係が存在すると，多重共線性 (multicollinearity) と呼ばれる現象が現れる．多重共線性の影響下では，回帰式の信頼性低下や得られた結果が真の値に反するなどの問題が起きる．そのため，多重共線性を引き起こす説明変数集合を用いて回帰分析を行うことは望ましくない．多重共線性を捕捉するための指標には，説明変数の相関係数や相関係数行列の条件数，分散拡大要因 (Variance Inflation Factors, VIF) が広く用いられている．このうち相関係数行列の条件数またはVIF を制約とした変数選択を，それぞれ条件数制約付き変数選択問題，VIF 制約付き変数選択問題と呼ぶ．本研究では特に，残差二乗和 (Residual Sum of Squares, RSS) の最小化を目的とする条件数制約付き変数選択問題とVIF 制約付き変数選択問題を扱う．

変数選択問題の解法としてはステップワイズ法が用いられることが多い．ステップワイズ法は出力する解の厳密性を保証しないが，質のよい解が得られることが経験的に知られており，動作も高速である．そのためR などの統計ソフトにも実装されている．また，厳密解法としては，近年数理計画法による手法が注目を集めている．

しかし，条件数制約やVIF 制約は固有値問題や逆凸型の制約を含むため，これらを扱いやすい形の数理計画問題として定式化することは難しいとされていた．そこで多重共線性を考慮した変数選択においては，厳密解法として数理計画法と切除平面法を組み合わせた汎用解法が提案されている [1–3]．しかし切除平面法では，指数回オーダーの混合整数二次計画問題 (Mixed Integer Quadratic Program, MIQP) を解く必要があり，データによっては膨大な計算時間がかかる点に問題があった．

2. 条件数制約付き変数選択問題

入力データとして p 個の説明変数と n 個のサンプルデータが与えられたとき，i 番目のサンプルデータを $(y_i; x_{i1}, x_{i2}, \ldots, x_{ip})$ とし，P を全説明変数の集合とする．また，z_j, a_j $(j \in P)$ をそれぞれ説明変数 j を選ぶとき 1 となる 0-1 変数と，説明変数 j についての偏回帰係数とする．このとき，選択された説明変数の相関係数行列の条件数を κ 以下とする制約下での変数選択問題は次の形で表現される．

$$\begin{aligned} \underset{\boldsymbol{a},\boldsymbol{z}}{\text{minimize}} \quad & ||\boldsymbol{y} - X\boldsymbol{a}||_2^2 \\ \text{subject to} \quad & z_j = 0 \quad \Rightarrow \quad a_j = 0 \quad (j \in P), \quad (1) \\ & \mathrm{Cond}(S(\boldsymbol{z})) \leq \kappa, \quad (2) \\ & \boldsymbol{z} \in \{0,1\}^p. \quad (3) \end{aligned}$$

ただし $X, \boldsymbol{y}, \boldsymbol{a}$ はそれぞれ x_{ij}, y_i, a_j $(i = 1, 2, \ldots, n, j \in P)$ で構成される行列とベクトル，$S(\boldsymbol{z})$ は $\boldsymbol{z}$ によって選択された説明変数集合，$\mathrm{Cond}(S(\boldsymbol{z}))$ は $S(\boldsymbol{z})$ に対応する相関係数行列の条件数とする．

上記の問題については，目的関数が二次の混合整数半正定値計画問題 (Mixed Integer Semidefinite Program, MISDP) としての定式化が提案されている [2]．しかし，現在利用可能な MISDP ソルバーで扱うことができるのは線形な目的関数のみである．ここで，任意の凸二次制約は半正定値制約により表現できることが知られている．これを用いて，上記の問題に対し目的関数が線形な MISDP としての定式化を提案する．

また，MISDP の求解の際には，膨大な数の半正定値計画問題 (Semidefinite Program, SDP) を解く必要がある．SDP の計算コストは高いため，その高速化により MISDP 全体の求解の効率化が期待される．選択される変数集合が定まったとき，最小の RSS は正規方程式と呼ばれる連立一次方程式を解くことで解析的に求めることができる．そこで本研究では，選択された変数集合についての正規方程式に対応する制約式の提案を行い，定式化へと追加することで高速化を実現する．提案した制約式を正規方程式制約と呼ぶこととする．

ベクトル $\boldsymbol{s} = (s_1, s_2, \ldots, s_p)^\top$ を新たな連続変数として定義する．ここで正規方程式制約 (4)，(5) を考える．

$$X^\top X\boldsymbol{a} + \boldsymbol{s} = X^\top \boldsymbol{y}, \tag{4}$$
$$z_j = 1 \quad \Rightarrow \quad s_j = 0 \quad (j \in P). \tag{5}$$

これらの制約について次の定理が成り立つ．

定理 1. 制約式 (4)，(5) は，$\boldsymbol{z}$ により選択された変数集合に対する正規方程式と等価である．

数値実験として，UCI Machine Learning Repository で公開されているデータセットに対して変数選択を行った．結果として，提案手法は説明変数が 25 個の小規模なデータまで最適解を得ることができたものの，数値的不安定性の面で切除平面法に劣っていることが示された．今後の MISDP 求解に関する研究の進展により，提案手法の実用性向上が見込まれる．

3. VIF 制約付き変数選択問題

本節では，VIF 制約付きの変数選択問題について，単一の MIQP としての定式化を二種類提案する．選択された説明変数集合 $S(\boldsymbol{z})$ に含まれる変数 ℓ についての VIF 値は，次の 2 通りに定義される．

$$\mathrm{VIF}(\ell, S(\boldsymbol{z})) = \frac{1}{1 - R^2(\ell, S(\boldsymbol{z}))} \tag{6}$$
$$= (R^{-1}_{S(\boldsymbol{z})})_{\ell\ell} \tag{7}$$

ただし，$R^2(\ell, S(\boldsymbol{z}))$ を $S(\boldsymbol{z}) \setminus \{\ell\}$ で変数 ℓ を回帰したときの決定係数，$R_{S(\boldsymbol{z})}, R^{-1}_{S(\boldsymbol{z})}$ を $S(\boldsymbol{z})$ からなる相関係数行列とその逆行列とする．ここで選択された変数に対する VIF 値を α 以下とする制約のもとでの変数選択問題は，次の形で表現することができる．

$$\begin{aligned}
\underset{\boldsymbol{a},\boldsymbol{z}}{\text{minimize}} \quad & ||\boldsymbol{y} - X\boldsymbol{a}||_2^2 \\
\text{subject to} \quad & z_j = 0 \quad \Rightarrow \quad a_j = 0 \quad (j \in P), && (8)\\
& z_\ell = 1 \quad \Rightarrow \\
& \qquad \mathrm{VIF}(\ell, S(\boldsymbol{z})) \le \alpha \quad (\ell \in P), && (9)\\
& \boldsymbol{z} \in \{0,1\}^p. && (10)
\end{aligned}$$

VIF 制約 (9) について，定義 (6) を用いて制約を構築すると，RSS が二次の関数であることから逆凸制約となる．しかし，今回提案した正規方程式制約を用いて RSS を線形に変形することで逆凸性を回避し，MIQP として定式化することができる．定義 (7) では，選択された説明変数の相関係数行列を求める必要がある．このとき $R_{S(\boldsymbol{z})}$ の逆行列 $R^{-1}_{S(\boldsymbol{z})}$ が存在するならば，$R^{-1}_{S(\boldsymbol{z})} R_{S(\boldsymbol{z})} = I$ を満たし，また $R^{-1}_{S(\boldsymbol{z})}$ の対角成分がそれぞれモデルに含まれる説明変数についての VIF 値に対応する．このような $R^{-1}_{S(\boldsymbol{z})}$ を線形制約によって表現できればよい．

ここで，連続変数 $q_{j\ell}, u_{j\ell}$ $(j, \ell \in P)$ からなる $p \times p$ 行列 $Q = (q_{j\ell}), U = (u_{j\ell})$ を定義する．Q, U を用いて以下の制約式 (11)–(13) を考える．

$$QR_P + U = I, \tag{11}$$
$$z_j = 0 \quad \Rightarrow \quad q_{j\ell} = q_{\ell j} = 0 \quad (j, \ell \in P), \tag{12}$$
$$z_j = 1 \quad \Rightarrow \quad u_{\ell j} = 0 \quad (j, \ell \in P). \tag{13}$$

このとき次の定理が成り立つ．

定理 2. Q, U, $\boldsymbol{z}$ が制約式 (11)–(13) を満たすとき，Q の部分行列 $(q_{j\ell})_{(j,\ell)\in S(\boldsymbol{z})\times S(\boldsymbol{z})}$ は $R^{-1}_{S(\boldsymbol{z})}$ に等しい．

数値実験を行い，提案する定式化とステップワイズ法，切除平面法の性能を比較した．今回提案した手法によって，説明変数が 32 個のデータまで 10000 秒以内に最適解を得ることができた．提案手法では小規模なデータにおいてステップワイズ法とほぼ同等の時間で最適解が得られ，またステップワイズ法では最適解が得られないデータがあることを示した．また，65 変数以上のデータにおいて求解を途中で打ち切って得られた暫定解は，ステップワイズ法で得られたものより高い精度であった．切除平面法との比較では，実行時間・求解を打ち切った時に得られた暫定解の精度は，いずれも提案手法が優れている結果となった．

参考文献

[1] D. Bertsimas and A. King, “OR forum—an algorithmic approach to linear regression,” *Operations Research*, **64**, pp. 2–16, 2016.

[2] 小林健，“多重共線性を考慮した回帰式の変数選択問題に対する混合整数計画法を用いた厳密解法，”東京工業大学修士論文，2014.

[3] R. Tamura, K. Kobayashi, Y. Takano, R. Miyashiro, K. Nakata and T. Matsui, “Best subset selection for eliminating multicollinearity,” *Journal of the Operations Research Society of Japan*, **60**, pp. 321–336, 2017（第 2 節の内容に対応）.

[4] R. Tamura, K. Kobayashi, Y. Takano, R. Miyashiro, K. Nakata and T. Matsui, “Mixed integer quadratic optimization formulations for eliminating multicollinearity based on variance inflation factor,” Optimization Online, 2016（第 3 節の内容に対応）.