■学生論文賞受賞論文　　要約■

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

小林　健

東京工業大学大学院社会理工学研究科経営工学専攻（現：株式会社富士通研究所）
指導教員：中田和秀 東京工業大学准教授

1. はじめに

本研究では線形回帰モデルの説明変数を選択する問題を扱う．説明変数間に強い相関や一次従属な変数関係がある場合，推定量の信頼度が低下する．この現象を多重共線性といい，回帰分析では多重共線性を回避するよう適切に説明変数を選択することが望まれる．

多重共線性を検出する指標として，相関係数行列の条件数と分散拡大要因がある．本研究ではこれらの指標に注目し，多重共線性の指標に関する制約のもと残差 2 乗和を最小化する変数選択問題を混合整数計画問題として定式化する．しかし相関係数行列の条件数に制約を与えた変数選択問題，分散拡大要因に制約を与えた変数選択問題はどちらも一般的な整数計画ソルバーで直接解くことは難しい．そこで本研究ではこれらの問題に対して適用可能な切除平面法に基づく厳密解法を提案する．また問題の特性を利用して切除平面法を効率化できることを示す．さらに本研究では，相関係数行列の条件数に制約を与えた変数選択問題について，この問題を混合整数半正定値計画問題 (MISDP) として定式化できることを示し，MISDP に対する切除平面法も併せて提案する．

2. 線形回帰モデルの最小 2 乗推定と多重共線性

本研究では n 個のサンプル $(y_i; x_{i1}, \ldots, x_{ip})$ $(i = 1, \ldots, n)$ を用いた，以下の線形回帰モデルを考える：

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{\varepsilon}. \tag{1}$$

ただし $\boldsymbol{y} := (y_1, \ldots, y_n)^\top$，$\boldsymbol{X} := (x_{ij})$，$\boldsymbol{a} := (a_1, \ldots, a_p)^\top$，$\boldsymbol{\varepsilon} := (\varepsilon_1, \ldots, \varepsilon_n)^\top$ とする．y_i $(i = 1, \ldots, n)$ は予測すべき被説明変数，x_{ij} $(i = 1, \ldots, n; j = 1, \ldots, p)$ は予測に用いる説明変数，a_j $(j = 1, \ldots, p)$ は推定すべき偏回帰係数，ε_i $(i = 1, \ldots, n)$ は予測残差である．被説明変数と説明変数は事前にすべて平均 0，分散 1 に正規化してあるものとする．

残差 2 乗和 $\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}$ を最小化する最小 2 乗推定量 $\hat{\boldsymbol{a}}$ は以下の正規方程式を解いて求められる：

$$\boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{a}} = \boldsymbol{X}^\top\boldsymbol{y}. \tag{2}$$

しかし正規方程式 (2) が悪条件の場合，数値計算の誤差により推定量 $\hat{\boldsymbol{a}}$ の信頼度が低下する．この現象を多重共線性という．多重共線性を検出する指標として以下の二つがある：

相関係数行列の条件数　説明変数の相関係数行列 $\boldsymbol{R}$ の条件数 $\mathrm{cond}(\boldsymbol{R})$ は，$\boldsymbol{R}$ の最大固有値と最小固有値の比 $\lambda_{\max}(\boldsymbol{R})/\lambda_{\min}(\boldsymbol{R})$ で定義される．多重共線性が存在する場合，相関係数行列の条件数は大きい値をとる．

分散拡大要因 (VIF)　j 番目の説明変数の分散拡大要因 VIF_j は，$\mathrm{VIF}_j = (\boldsymbol{R}^{-1})_{jj}$ と定義される．多重共線性が存在する場合，VIF の大きい説明変数が存在する．

3. 変数選択問題に対する切除平面法

3.1 多重共線性の指標に制約を与えた変数選択問題

本研究では多重共線性の指標に制約を与えた以下の変数選択問題に対する解法を考える：

$$\underset{\boldsymbol{a},\boldsymbol{z}}{\text{minimize}} \quad \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} a_j x_{ij}\right)^2 \tag{3}$$

$$\text{subject to} \quad z_j = 0 \Rightarrow a_j = 0 \ (j = 1, \ldots p), \tag{4}$$

$$z_j \in \{0, 1\} \qquad (j = 1, \ldots p), \tag{5}$$

$$\boldsymbol{z} = (z_1, \ldots, z_p)^\top \in \mathcal{F}. \tag{6}$$

ここで 0-1 決定変数 z_j は j 番目の説明変数を選択しない／するを表す変数であり，連続変数 a_j はその説明変数の偏回帰係数を表す．目的関数 (3) は残差 2 乗和の最小化を表し，制約式 (4) は $z_j = 0$ のときに j 番目の説明変数を回帰式から取り除く制約を表す．

制約 (6) における集合 $\mathcal{F}$ は，多重共線性の指標に関する制約を満たす変数選択の解 $\boldsymbol{z} \in \{0, 1\}^p$ の集合を表す．相関係数行列の条件数が上限 κ を超えないと

いう制約のもとで変数選択を行う場合，

$$\mathcal{F} = \mathcal{F}_{\mathrm{cond}} := \{\boldsymbol{z} \mid \mathrm{cond}(\boldsymbol{R}(\boldsymbol{z})) \leq \kappa\} \qquad (7)$$

とする．ここで $\boldsymbol{R}(\boldsymbol{z})$ は $\boldsymbol{z} \in \{0,1\}^p$ で選択された説明変数の相関係数行列（$\boldsymbol{R}$ の主小行列）とする．選択された変数の VIF が上限 α を超えないという制約のもとで変数選択を行う場合，

$$\mathcal{F} = \mathcal{F}_{\mathrm{VIF}} := \{\boldsymbol{z} \mid \max_j \ (\boldsymbol{R}(\boldsymbol{z}))^{-1}_{jj} \leq \alpha\} \qquad (8)$$

とする．

3.2 切除平面法に基づく汎用解法

ここでは 3.1 節で述べた条件数制約つき変数選択問題，VIF 制約つき変数選択問題どちらにも適用可能な切除平面法に基づく汎用解法について述べる．まず $\mathcal{F}_{\mathrm{cond}}, \mathcal{F}_{\mathrm{VIF}}$ および $\boldsymbol{z}, \bar{\boldsymbol{z}} \in \{0,1\}^p$ に対して，

$$\bar{\boldsymbol{z}} \notin \mathcal{F}_{\mathrm{cond}} \text{ かつ } \boldsymbol{z} \geq \bar{\boldsymbol{z}} \ \Rightarrow \boldsymbol{z} \notin \mathcal{F}_{\mathrm{cond}} \qquad (9)$$

$$\bar{\boldsymbol{z}} \notin \mathcal{F}_{\mathrm{VIF}} \text{ かつ } \boldsymbol{z} \geq \bar{\boldsymbol{z}} \ \Rightarrow \boldsymbol{z} \notin \mathcal{F}_{\mathrm{VIF}} \qquad (10)$$

が成り立つ．汎用解法では問題 (3)〜(6) に対して多重共線性の制約 (6) を取り除いた緩和問題 (3)〜(5) を考える．問題 (3)〜(5) は混合整数 2 次計画問題 (MIQP) であり整数計画ソルバーで解ける．続いて問題 (3)〜(5) の最適解 $\boldsymbol{z}_k$ に注目し，$\boldsymbol{z}_k$ が制約 (6) を満たすか判定する．$\boldsymbol{z}_k \in \mathcal{F}$ である場合，アルゴリズムを終了する．そうでない場合，式 (9)，(10) から $\boldsymbol{z} \geq \boldsymbol{z}_k$ である任意の $\boldsymbol{z} \in \{0,1\}^p$ に対して $\boldsymbol{z} \notin \mathcal{F}$ となる．そこでこのような $\boldsymbol{z}$ を取り除くため，$\bar{\boldsymbol{z}} \leftarrow \boldsymbol{z}_k$ とし，

$$\bar{\boldsymbol{z}}^\top \boldsymbol{z} \leq \mathbf{1}^\top \bar{\boldsymbol{z}} - 1 \qquad (11)$$

を問題 (3)〜(5) の制約に追加して問題を解きなおす．ここで $\mathbf{1}$ は要素がすべて 1 のベクトルを表す．このように多重共線性が存在する変数集合を含む解を除去する制約を逐次加えることで，高々2^p 回緩和問題を解けば元の問題 (3)〜(6) に対する最適解が得られる．

3.3 変数減少法を組み合わせた切除平面法

候補となる説明変数の数が多いデータでは，緩和問題を解きなおす回数が増大し切除平面法全体の計算時間も増大する．3.2 節で述べた汎用解法では $\bar{\boldsymbol{z}} \leftarrow \boldsymbol{z}_k$ として制約 (11) を追加し，$\boldsymbol{z} \geq \boldsymbol{z}_k$ である解 $\boldsymbol{z} \in \{0,1\}^p$ を取り除く．しかし $\bar{\boldsymbol{z}} \leq \boldsymbol{z}_k$ かつ $\bar{\boldsymbol{z}} \notin \mathcal{F}$ である $\bar{\boldsymbol{z}} \in \{0,1\}^p$ が存在する場合，その $\bar{\boldsymbol{z}}$ を用いて制約 (11) を追加すれば 1 回の反復でより多くの実行不能解を取り除くことができる．また多重共線性が検出される変数集合の解を多く取り除くためには，$\bar{\boldsymbol{z}}$ は選択する変数の数がなるべく少ない解であることが望ましい．そこで本研究では 3.2 節の汎用解法に変数減少法を組み合わせて各反復で所望の $\bar{\boldsymbol{z}}$ を構成し，その $\bar{\boldsymbol{z}}$ を用いることでより強い妥当不等式を制約に追加する切除平面法を提案した．

3.4 条件数制約つき変数選択問題に対する切除平面法

一般に条件数制約は半正定値制約として表現できることが知られている．そこで本研究では条件数制約つき変数選択問題を MISDP として定式化できることを示した．そして Konno et al. [1] が提案した半正定値計画問題を解くための切除平面法を拡張し，MISDP を解く切除平面法を提案した．さらに MISDP に対する切除平面法に関するいくつかの性質を示し，3.2 節の汎用解法と同様に変数減少法を組み合わせて強い妥当不等式を構成できることを示した．

4. 数値実験

UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/) で公開されているデータセットとランダムに生成した人工データを用いて数値実験を行った．その結果，提案手法の切除平面法に変数減少法を組み合わせると，求解に要する反復回数が削減され，計算時間も大幅に短縮されることが示された．ステップワイズ変数選択である変数増加法，変数減少法との比較実験では，提案手法はそれらの手法より決定係数の高い回帰式を得られた．説明変数の分類性能を検証する実験では，提案手法は Lasso より高い分類性能を示した．

参考文献

[1] H. Konno, J. Gotoh, T. Uno and A. Yuki, “A cutting plane algorithm for semidefinite programming problems with applications to failure discriminant analysis,” *Journal of Computational and Applied Mathematics*, **146**, pp. 141–154, 2002.