

混合整数二次錐計画法による情報量規準最小化手法の高速化

Improvement on minimizing information criteria
via mixed integer second-order cone programming

東京農工大学・大学院工学府 神谷 俊介

Shunsuke Kamiya

Graduate School of Engineering, Tokyo University of Agriculture and Technology

専修大学・ネットワーク情報学部 高野 祐一

Yuichi Takano

School of Network and Information, Senshu University

東京農工大学・大学院工学研究院 宮代 隆平

Ryuhei Miyashiro

Institute of Engineering, Tokyo University of Agriculture and Technology

1 はじめに

本研究では、線形回帰分析のモデル構築に用いる説明変数について、それらの最良部分集合を決定する変数選択問題を考える。変数選択問題はNP困難であることが知られており、これまで多くの凸緩和手法やヒューリスティクスが考案されてきた (lasso [16] やステップワイズ法 [5] など)。近年、計算機の高速化や数値計画ソルバーの性能向上を受けて、変数選択問題を数値計画法の枠組で厳密に解く研究が盛んに行われている。本研究ではこのうち、赤池情報量規準 (Akaike's Information Criterion, AIC) [1] やベイズ情報量規準 (Bayesian Information Criterion, BIC) [14] 等の情報量規準を指標とした最適化問題を、混合整数二次錐計画問題 (Mixed Integer Second-Order Cone Program, MISOCP) として定式化した研究 (宮代・高野 [13]) に着目する。彼らは同論文で、説明変数の個数が 30 個以下のインスタンスについては数分以内に最適解を発見できるが、より大きなインスタンスについては現実的な時間での求解が困難であると結論付けている。本研究では情報量規準最小化問題の MISOCP 定式化 [13] を改良し、より大規模なインスタンスの求解を目指す。

1.1 変数選択問題と情報量規準最小化

被説明変数と p 個の説明変数に対して、所与の n 個のサンプル $\mathbf{y} = [y_1, y_2, \dots, y_n]^\top \in \mathbb{R}^n$, $X = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]^\top \in \mathbb{R}^{n \times p}$ から、次式の線形回帰モデルを推定することを考える:

$$\hat{y}_i = f(\mathbf{a}, \mathbf{x}_i) = a_0 + a_1 x_{i1} + a_2 x_{i2} + \dots + a_p x_{ip} \quad (i = 1, 2, \dots, n).$$

ただし a_j ($j = 1, 2, \dots, p$) は j 番目の説明変数に対応する偏回帰係数であり, a_0 は切片項を示す. ここで残差二乗和を最小化する目的で全ての説明変数を用いてモデルを構築すると, 過適合を起こし未知のデータに対する予測力が低下したり, 人間にとって解釈し難いモデルになってしまう. したがって適切に説明変数を取捨選択し, 限られた説明変数だけを用いてモデルを構築することが重要となる.

次式で定義される AIC は期待平均対数尤度の不偏推定量であり, モデル選択に用いられる情報量規準の一つである. ここで S は選択された説明変数の集合, k は選択された説明変数の個数, $\text{RSS}(S)$ は S に含まれる説明変数だけを用いた時の残差二乗和の最小値を示す:

$$\text{AIC}(S) = n \log_e \frac{\text{RSS}(S)}{n} + 2k.$$

同様に BIC は, $\text{BIC}(S) = n \log_e \frac{\text{RSS}(S)}{n} + k \log_e n$ で定義される.

1.2 情報量規準最小化問題の定式化

情報量規準 AIC を最小化する変数選択問題は, 下記のとおり MISOCP としての定式化が与えられている [13]. BIC を最小化する場合は, 制約式 $g = \sum_{j=0}^p \left(w_j \cdot \exp \left(-\frac{2j}{n} \right) \right)$ を $g = \sum_{j=0}^p \left(w_j \cdot n^{-j/n} \right)$ に変更すればよい. なお以下では AIC の最小化について述べていくが, BIC の最小化についてもほぼ同様に議論を進められる.

$$\begin{aligned} & \text{minimize} && f \\ & \text{subject to} && \varepsilon_i = y_i - \left(a_0 + \sum_{j=1}^p a_j x_{ij} \right) \quad (i = 1, 2, \dots, n), \\ & && \varepsilon^\top \varepsilon \leq f \cdot g, \\ & && g = \sum_{j=0}^p \left(w_j \cdot \exp \left(-\frac{2j}{n} \right) \right), \\ & && \sum_{j=0}^p (j \cdot w_j) = \sum_{j=1}^p z_j, \\ & && \sum_{j=0}^p w_j = 1, \\ & && -Mz_j \leq a_j \leq +Mz_j \quad (j = 1, 2, \dots, p), \\ & && \mathbf{a} \in \mathbb{R}^{p+1}, \varepsilon \in \mathbb{R}^n, f \in \mathbb{R}_+, g \in \mathbb{R}_+, \mathbf{w} \in \{0, 1\}^{p+1}, \mathbf{z} \in \{0, 1\}^p. \end{aligned} \tag{1}$$

ただし, ここで M は十分大きな正の定数とする.

この定式化を用いて現実的な時間で求解を終えられるのは, 説明変数の個数 p が 30 程度までであることが実験的に知られている. これには, 以下の原因が考えられる.

- (i) MISOCP であるため, 子問題の SOCP を内点法で解く必要があり, 分枝限定法中にホットスタートが働かない.
- (ii) 本来表現したい非線形の等式制約である $g = \exp \left(-\frac{2k}{n} \right)$, $k = \sum_{j=1}^p z_j$ を, k が整数であ

ることを利用し SOS Type 1 [2] で $g = \sum_{j=0}^p (w_j \cdot \exp(-\frac{2j}{n}))$, $\sum_{j=0}^p (j \cdot w_j) = \sum_{j=1}^p z_j$ と線形化しているが、これは緩和問題の観点からは弱い定式化となる。

(iii) 双曲型制約 $\epsilon^T \epsilon \leq f \cdot g$ の左辺における決定変数 ϵ の次元数がサンプル数 n に依存するため、サンプル数が多い問題では計算時間の増大を招く。

(iv) 定式化に含まれる big- M 法が、連続緩和問題を弱くしている。

本研究では MISOCP の枠組を保ったまま、定式化を修正することで上記 (iii)(iv) の問題の改善を目指し、変数選択問題の計算の高速化を図る。

2 関連研究

前述の情報量規準最小化問題は、選択する説明変数の個数 k を固定すると混合整数凸二次計画問題 (Mixed Integer Quadratic Program, MIQP) に帰着される。さらに $k = 0, 1, \dots, p$ について得られた全ての最適値を用いて、元問題の大域的最適解を構築可能である。このように各 k についての残差二乗和の最適値をそれぞれ列挙する手法には、leaps and bound [7] や Gatu らの分枝限定法 [8] がある。これらの手法は、各ノードが説明変数のいくつかの部分集合を表す分枝木を構築したり各ノードで残差二乗和の最小化を行うアルゴリズムについて工夫をすることで、規模の大きいインスタンスに対する最適解の発見を目指している。ただし、これらは独自の分枝限定法を設計しているため、アルゴリズムに対し修正を加えることが困難である。その点で、木村・脇の手法 [11] は分枝限定法フレームワークの 1 種である SCIP [15] のユーザプラグインを用いて実装しているため、比較的拡張が容易である。同手法は、SCIP の分枝限定法における目的関数値の上界の算出や分枝順序に工夫を施したアルゴリズムを提案している。このうち、目的関数値の上界の算出については情報量規準最小化問題に対する実用的なヒューリスティクス的一种であるステップワイズ法 [5] を用いている。

一方で、分枝限定法のアルゴリズムは数理計画ソルバーに任せ、混合整数計画問題への帰着のさせ方に工夫をする手法が近年注目を集めている。特に宮代と高野のモデリング手法 [13] は、情報量規準最小化問題を単一の MISOCP に定式化する点で新しいと言える。以下では、[13] で与えられた定式化を改善する手法を第 3 節と第 4 節でそれぞれ述べる。

3 問題サイズの削減

この節では定式化 (1) に修正を施し、分枝限定法を高速化することを目指す。具体的には双曲型制約 $\epsilon^T \epsilon \leq f \cdot g$ を変形し SOCP 緩和問題の求解を高速化する手法、および決定変数 $w \in \{0, 1\}^{p+1}$ を修正し分枝木を小さくする手法について順に述べる。なお変数選択問題のインスタンスは $n > p$ のケースと $p \gg n$ の二つのケースに分類されることが多いが、ここでは $n > p$ のケースを仮定する。

まず、定式化 (1) における双曲型制約 $\epsilon^T \epsilon \leq f \cdot g$ に着目する。SOCP においては双曲型制約

左辺の決定変数の次元数が求解速度に影響することが知られている．元々の制約式 $\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} \leq f \cdot g$ では左辺の決定変数は $\boldsymbol{\varepsilon}$ であり，その次元数は $O(n)$ であるが，これは以下の通り $O(p)$ に減らすことができる．次の $(p+1) \times (p+1)$ 対称行列 Q および $p+1$ 次元ベクトル $\hat{\mathbf{a}}$ を考える：

$$Q = \begin{pmatrix} X^\top X & -X^\top \mathbf{y} \\ (-X^\top \mathbf{y})^\top & \mathbf{y}^\top \mathbf{y} \end{pmatrix}, \quad \hat{\mathbf{a}} = [a_0, a_1, \dots, a_p, 1]^\top. \quad (2)$$

これらを用いると，残差二乗和の定義から $\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} = (\mathbf{y} - X\mathbf{a})^\top (\mathbf{y} - X\mathbf{a}) = \hat{\mathbf{a}}^\top Q \hat{\mathbf{a}}$ である．さらに行列 Q の半正定値性に着目すると， $\hat{\mathbf{a}}^\top Q \hat{\mathbf{a}} \leq f \cdot g$ は双曲型制約であることが分かり，また左辺の決定変数の次元数は $O(p)$ となる．

次に定式化 (1) において，制約式 $g = \sum_{j=0}^p (w_j \cdot \exp(-\frac{2j}{n}))$ に着目する．これは非線形の等式制約である $g = \exp(-\frac{2k}{n})$ を， k が整数であることを利用し SOS Type 1 [2] で線形化した制約式である．ここで元の定式化では $O(p)$ 個の 0-1 変数 $\mathbf{w}$ を用いて整数変数 k を表しているが，これは $O(\lceil \log_2(p+1) \rceil)$ 個の 0-1 変数でも表現可能である（例えば [6], [12]）．この事実を利用すると，0-1 変数 $\mathbf{b}$ で構成される以下の制約式で等価な制約を表現できる：

$$\begin{aligned} \sum_{j \in \mathbb{J}_0(\ell)} w_j &\leq 1 - b_\ell \quad (\ell = 0, 1, \dots, \lceil \log_2(p+1) \rceil - 1), \\ \sum_{j \in \mathbb{J}_1(\ell)} w_j &\leq b_\ell \quad (\ell = 0, 1, \dots, \lceil \log_2(p+1) \rceil - 1), \\ \mathbf{w} &\in \mathbb{R}_+^{p+1}, \quad \mathbf{b} \in \{0, 1\}^{\lceil \log_2(p+1) \rceil}. \end{aligned}$$

ただし， $\mathbb{J}_0(\ell) = \{j \in \{0, 1, \dots, p\} \mid j \text{ の第 } \ell \text{ ビットが } 0\}$ ， $\mathbb{J}_1(\ell) = \{j \in \{0, 1, \dots, p\} \mid j \text{ の第 } \ell \text{ ビットが } 1\}$ ($\ell = 0, 1, \dots, \lceil \log_2(p+1) \rceil - 1$) である．

しかしながら今回扱う情報量規準最小化問題については，上記の 0-1 変数 $\mathbf{b}$ を用いた工夫は一部のインスタンスを除き性能を悪化させてしまったため，以下では扱わないこととする．

4 Big-M 法に関する工夫

定式化 (1) の制約式 $-Mz_j \leq a_j \leq +Mz_j$ ($j = 1, 2, \dots, p$) は，説明変数を選択するか否かを規定する 0-1 変数 z と偏回帰係数を表す連続変数 $\mathbf{a}$ を関連付ける不等式である．従来手法 [13] では，この不等式は数値計画ソルバー CPLEX [9] の indicator という機能を用いて表現されていた．しかし，indicator で実現された制約式は緩和問題では取り除かれる．したがって a_j の値が 0 に固定されない，すなわち全説明変数を用いた線形回帰モデルが緩和解となるため，目的関数値の下界が上がらず分枝限定法における限定操作が実行されにくい．

ここで，以下の定理を考える*1．

定理. $X^\top X \in \mathcal{S}_{++}^p$ ならば， $|a_j^{(k)*} - a_j^{\text{OLS}}| \leq \sqrt{(\text{RSS}^{(k)} - \text{RSS}^{\text{OLS}}) \{(X^\top X)^{-1}\}_{jj}}$ を満たす．

1 なお，定理中の不等式左辺 $|a_j^{(k)} - a_j^{\text{OLS}}|$ の上界が計算可能であること自体は Bertsimas ら [3] の論文（付録 A.1）で述べられているが，具体的に右辺が $\sqrt{(\text{RSS}^{(k)} - \text{RSS}^{\text{OLS}}) \{(X^\top X)^{-1}\}_{jj}}$ で抑えられることの証明については神谷 [10] による．本稿では証明は省略する．

ここで S_{++}^p は $p \times p$ の正定値行列全体であり、 $\mathbf{a}^{\text{OLS}}$ および RSS^{OLS} は最小二乗法の最適解となる偏回帰係数および残差二乗和である。また $\mathbf{a}^{(k)*}$ および $\text{RSS}^{(k)}$ は、 k を固定した際の残差二乗和最小化問題の最適解および目的関数値の上界である。本手法では上記の定理を用いるため、 $X^\top X \in S_{++}^p$ を仮定する。この仮定を満たさないインスタンスは、天気や曜日などの複数カテゴリを表現するダミー変数を説明変数に持つものなどが考えられる。

次に、以下の補題を考える。

補題. S と T を説明変数の部分集合とする。 $S \subseteq T$ ならば、 $\text{RSS}(S) \geq \text{RSS}(T)$ が成り立つ。

したがって、 $u_j^{k\pm} = a_j^{\text{OLS}} \pm \sqrt{(\text{RSS}^{(k)} - \text{RSS}^{\text{OLS}}) \{(X^\top X)^{-1}\}_{jj}}$ と定義すると、制約式 $u_j^{1-} z_j \leq a_j \leq u_j^{1+} z_j$ ($j = 1, 2, \dots, p$) を定式化 (1) に追加しても最適解が取り除かれることはない。

さらに、定式化 (1) で SOS Type 1 に用いた 0-1 変数 w を利用すると、

$$u_j^{\ell-} w_\ell + u_j^{1-} (1 - w_\ell) \leq a_j \leq u_j^{\ell+} w_\ell + u_j^{1+} (1 - w_\ell) \quad (j = 1, 2, \dots, p; \ell = 0, 1, \dots, p) \quad (3)$$

を妥当不等式として利用することができる。この制約式は、

$$\sum_{\ell=0}^p (u_j^{\ell-} w_\ell) \leq a_j \leq \sum_{\ell=0}^p (u_j^{\ell+} w_\ell) \quad (j = 1, 2, \dots, p) \quad (4)$$

とすると制約式の本数を減らすことができる。制約式 (4) は制約式 (3) に比べ、緩和問題において良い下界を与えることが示せる。

なお、0-1 変数 z を含めた制約式として

$$\begin{aligned} a_j &\leq u_j^{\ell+} z_j + (1 - w_\ell) (u_j^{1+} - u_j^{\ell+}) \quad (j = 1, 2, \dots, p; \ell = 0, 1, \dots, p), \\ a_j &\geq u_j^{\ell-} z_j + (1 - w_\ell) (u_j^{1-} - u_j^{\ell-}) \quad (j = 1, 2, \dots, p; \ell = 0, 1, \dots, p) \end{aligned}$$

という表現も可能だが、予備的な計算機実験で良い性能を示さなかったため割愛する。

既存研究 [3][4] において、 $u_j^{k\pm}$ の算出に用いる $\text{RSS}^{(k)}$ は勾配射影法をベースとした離散一次法で見積もられている。 $\text{RSS}^{(k)}$ を用いると AIC の上界を見積もれることに注意すると、所与の $\text{RSS}_0^{(k)}$ ($k = 0, 1, \dots, p$) に対し

$$\text{RSS}^{(k)} := \exp\left(-\frac{2(k - \bar{k})}{n}\right) \cdot \text{RSS}_0^{(\bar{k})} \quad (k = 0, 1, \dots, p)$$

で残差二乗和の上界を強化できることが分かる (図 1)。ここで

$$\bar{k} := \underset{k}{\operatorname{argmin}} \left\{ n \log_e \left(\frac{1}{n} \text{RSS}_0^{(k)} \right) + 2k \mid k \in \{0, 1, \dots, p\} \right\}$$

である。さらに残差二乗和の下限が RSS^{OLS} である事実を利用すると、 $\text{RSS}^{(k)} < \text{RSS}^{\text{OLS}}$ を満たす最小の $k' \in \{0, 1, \dots, p\}$ について不等式制約 $k \leq k'$ を加えても最適性を失わない (図 1 の例では $k \leq 18$)。また本手法は $\bar{k}$ および $\text{RSS}^{(\bar{k})}$ が得られれば十分であるため、離散一次法をステップワイズ法などのヒューリスティクスに代替可能である。

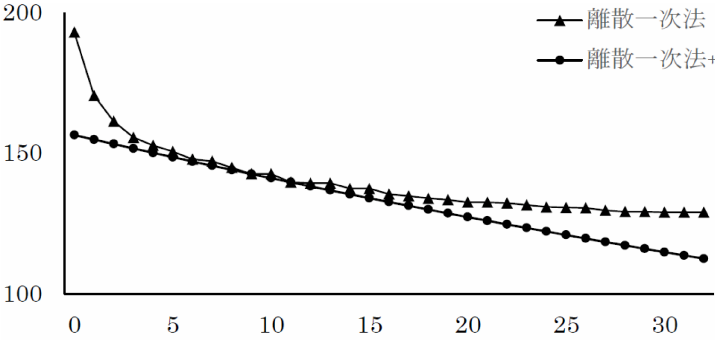


図 1: 異なる前処理で得られた $RSS^{(k)}$ の比較
 (離散一次法 + : 離散一次法の最良解に提案手法を適用;
 インスタンス : BreastCancer ($p = 32, n = 194$))

なお本節で述べた仮定 $X^T X \in S_{++}^p$ は, 前述したものとは別のアプローチ (詳細は [4]) で偏回帰係数の上下界を解析的に見積もれば外すことができる. ただし, [4] で指摘されているように上下界の質は劣る.

5 計算機実験

提案した AIC 最小化問題の定式化の改良について, それぞれの計算時間を比較した (表 1). 計算機環境は, PC: Dell Optiplex 9020, OS: Windows 7 Professional SP1 (64bit), CPU: Intel Core i7-4790 3.60 GHz, RAM: 8 GB RAM, ソルバー : CPLEX 12.6.3, 分枝限定法のスレッド数 : 1 である. それぞれのインスタンスは UCI Machine Learning Repository [17] で公開されており, Crime i ($i = 1, 2, 3$) はインスタンス Crime の説明変数を削減したものを示す. 表中の定式化は, 定式化 Original が論文 [13] のもの, 定式化 P が式 (2) を用いて双曲型制約のサイズを小さくしたもの, 定式化 PS が P に制約式 (4) を追加したもの, 定式化 PS+ が PS に第 4 節で述べた不等式制約 $k \leq k'$ を追加したものである. また, 実行時間の >10000 は 1 万秒で求解を打ち切ったインスタンスを示し, AIC* および k^* は得られた最適解または最良の実行可能解における AIC 値と k の値である.

実験の結果, 定式化 PS+ が最も優れていることが確認できた. 特に BreastCancer では, 従来の定式化に比べ 100 倍以上高速な求解が実現された. また, 従来の定式化では 10000 秒以内に解ききれなかった Crime 1, 2, 3 についても定式化 PS+ では現実的な計算時間で計算が終了していることがわかる.

表 1: AIC 最小化問題に関する異なる定式化の比較実験

インスタンス	n	p	定式化	AIC*	k^*	実行時間
BreastCancer	194	32	Original	508.40	10	24296.61
			P	508.40	10	4075.04
			PS	508.40	10	3577.77
			PS+	508.40	10	235.70
Crime 1	1993	40	Original	3621.61	25	>10000
			P	3620.29	25	1860.72
			PS	3620.29	25	500.92
			PS+	3620.29	25	89.19
Crime 2	1993	50	Original	3543.49	30	>10000
			P	3543.16	30	>10000
			PS	3543.16	30	5190.01
			PS+	3543.16	30	2334.88
Crime 3	1993	50	Original	3596.18	24	>10000
			P	3596.18	23	3400.23
			PS	3596.18	23	1355.29
			PS+	3596.18	23	248.18

6 おわりに

本研究では情報量規準最小化問題について、従来研究に比べてより大きなインスタンスを扱える定式化を提案した。具体的には二次錐制約における決定変数の次元の修正や連続変数 $\boldsymbol{a}$ の許容領域の制限を施した結果、 $n < p$ および $X^\top X \in S^p_{++}$ の仮定を満たすインスタンスで約 100 倍の高速化に成功した。また緩和問題の質の向上に伴い、従来の定式化に比べインスタンスの数値データに依存せずに安定した実行時間で最適解が求められることも確認されている。ただし、 $p = 50$ を超えるインスタンスについては未だに現実的な実行時間での求解が難しいため、今後の課題である。

参考文献

[1] H. Akaike, “A New Look at the Statistical Model Identification,” IEEE Transactions on Automatic Control, Volume 19, Issue 6, pp. 716–723, 1974.

[2] E. M. L. Beale, “Two Transportation Problems,” Proceedings of the Third International Conference on Operational Research, pp. 780–788, 1963.

[3] D. Bertsimas, A. King, R. Mazumder, “Best Subset Selection via a Modern Optimization Lens,” preprint on arXiv, <https://arxiv.org/pdf/1507.03133.pdf> (2017/11/1 アクセス), 2015.

- [4] D. Bertsimas, A. King, R. Mazumder, “Best Subset Selection via a Modern Optimization Lens,” *The Annals of Statistics*, Volume 44, Number 2, pp. 813–852, 2016.
- [5] M. A. Efroymson, “Multiple Regression Analysis,” *Mathematical Methods for Digital Computers*, Wiley, pp. 191–203, 1960.
- [6] 藤江哲也, 整数計画法による定式化入門, オペレーションズ・リサーチ, Volume 57, Number 4, pp. 190–197, 2012.
- [7] G. M. Furnival, R. W. Wilson, “Regressions by Leaps and Bounds,” *Technometrics*, Volume 16, Number 4, pp. 499–511, 1974.
- [8] C. Gatun, E. J. Kontoghiorghes, “Branch-and-Bound Algorithms for Computing the Best-Subset Regression Models,” *Journal of Computational and Graphical Statistics*, Volume 15, Issue 1, pp. 139–156, 2006.
- [9] IBM ILOG CPLEX Optimizer 12.6.3, 2015.
- [10] 神谷俊介, 整数計画法による高速な変数選択手法の提案, 東京農工大学工学部情報工学科卒業論文, 2017.
- [11] K. Kimura, H. Waki, “Minimization of Akaike’s Information Criterion in Linear Regression Analysis via Mixed Integer Nonlinear Program,” *Optimization Methods and Software*, to appear.
- [12] 久保幹夫, J. P. ペドロソ, 村松正和, A. レイス, 新しい数理最適化: Python 言語と Gurobi で解く, 近代科学社, 2012.
- [13] R. Miyashiro, Y. Takano, “Mixed Integer Second-Order Cone Programming Formulations for Variable Selection in Linear Regression,” *European Journal of Operational Research*, Volume 247, Issue 3, pp. 721–731, 2015.
- [14] G. Schwarz, “Estimating the Dimension of a Model,” *Annals of Statistics*, Volume 6, Number 2, pp. 461–464, 1978.
- [15] SCIP: Solving Constraint Integer Programs,
<http://scip.zib.de/> (2016/12/21 アクセス).
- [16] R. Tibshirani, “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society, Series B*, Volume 58, Issue 1, pp. 267–288, 1996.
- [17] UCI Machine Learning Repository,
<http://archive.ics.uci.edu/ml/> (2016/12/21 アクセス).