潜在変数モデルに基づく l_0 正則化付き最小二乗アルゴリズム
l_0-regularized least squares algorithm based on latent variable model

経営システム工学専攻　浅井 謙輔

1　はじめに

p 次元の説明変数 $\boldsymbol{x} := (x_1, \ldots, x_p)^\top$ を用いて，従属変数 y を以下の様に予測する線形モデルについて考える．

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon\ , \qquad \varepsilon \sim N(0, \sigma^2)$$

$\boldsymbol{y} := (y_1, \ldots, y_N)^\top$, $X := (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)^\top$ として，残差平方和 $\sum_{i=1}^N \varepsilon_i^2$ を最小にする $\boldsymbol{\beta} := (\beta_0, \beta_1, \ldots, \beta_p)^\top$ を求める問題である最小二乗法は以下の様に書ける．

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2$$

実際のデータにおいては，得られたデータに対して，従属変数を説明する特徴量が非常に少ない場合が多い．そこで，制約条件を付けた以下の最適化問題

$$\begin{aligned} \underset{\boldsymbol{\beta}}{\text{minimize}} \quad & \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 \\ \text{subject to} \quad & \|\boldsymbol{\beta}\|_0 \leq K \end{aligned} \tag{1}$$

を考える．ここで，$\|\boldsymbol{\beta}\|_0$ は $\boldsymbol{\beta}$ のゼロでない要素の数を表し，l_0 ノルムと呼ばれる．この問題は，NP 困難であることが知られている．そこで，貪欲法を用いて解を求める方法が一般的である．近年の研究では，(1) の制約条件である l_0 ノルムを 2 つの凸関数の差で表し，最適化を行う DC アプローチ [1] が存在する．l_0 ノルムを用いた関数は非凸であり，大域的最適解を得られるとは限らない．本論文では，(1) にラグランジュ乗数 λ を導入した

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_0 \tag{2}$$

を考える．l_0 正則化付き最小化問題である (2) の求解方法として，潜在変数モデルに基づく定式化を行い，EM アルゴリズムないし一般化 EM アルゴリズムに基づいたアルゴリズムによって，組み合わせ計算を行わずに良好な大域的収束性を得ることを目的とする．組み合わせ計算を行わないので，現実的に計算可能な時間で解が得られるという利点がある．

2　潜在変数モデルと EM アルゴリズム

パラメータ θ とする観測データ X の確率密度関数は，確率変数 X と観測されない確率変数 Z の同時確率密度関数の周辺分布であり，

$$f(x|\theta) = \sum_z f(x, z|\theta)$$

と表せる確率モデルについて考える．$f(x, z|\theta)$ は X と Z の同時確率密度関数である．この様な確率変数 Z を潜在変数と呼ぶ．EM アルゴリズムは潜在変数を持つ尤度を最大化するアルゴリズムであり，以下の E ステップと M ステップを繰り返す．

1. パラメータの初期値 θ^{old} を選ぶ．
2. 【E ステップ】
   潜在変数の事後確率 $p(z|x; \theta^{\text{old}})$ を計算する．
3. 【M ステップ】
   Q 関数を最大にする θ を計算し，
   $$\theta^{\text{new}} = \arg\max_\theta Q(\theta, \theta^{\text{old}})$$
   とする．ただし，Q 関数
   $$Q(\theta, \theta^{\text{old}}) := \sum_z p(z|x; \theta^{\text{old}})\ \log f(x, z|\theta)$$
   である．
4. 対数尤度関数またはパラメータの値が収束条件を満たしていれば終了．満たしていなければ，$\theta^{\text{old}} \leftarrow \theta^{\text{new}}$ として，2. に戻る．

また，Q 関数が増加するようにパラメータ θ を更新するアルゴリズムを一般化 EM アルゴリズムという．

3　モデルの定式化

3.1　潜在変数モデルの導入

データセット $d := ((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N))$ $(\boldsymbol{x}_i := (x_{i,1}, x_{i,2}, \ldots, x_{i,p})^\top)$ とする．第 1 章と同様，l_0 正則化付き最小化問題

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_0 \tag{2}$$

を解くことを考える．(2) の解を得るために，0 と 1 を取る変数を追加した以下の問題を解く．

$$\begin{aligned} \underset{\boldsymbol{\beta}, \boldsymbol{z}}{\text{minimize}} \quad & \|\boldsymbol{y} - Xb(\boldsymbol{z})\|_2^2 + \lambda\|\boldsymbol{z}\|_0 \\ \text{subject to} \quad & \boldsymbol{z} \in \{0, 1\}^p,\ \boldsymbol{\beta} \in \mathbb{R}^{p+1} \end{aligned} \tag{3}$$

ただし，

$$b(\boldsymbol{z}) := (\beta_0, z_1\beta_1, \cdots, z_p\beta_p)^\top$$

である．潜在変数 $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_p)^\top$，($\boldsymbol{Z} \in \{0,1\}^p$) の条件付確率として，データセット d をとる確率を

$$P(\mathcal{D} = d|\boldsymbol{Z} = \boldsymbol{z}) = \exp\left\{-\frac{1}{\|\boldsymbol{y}\|_2^2 + \lambda p}\left(\|\boldsymbol{y} - X\boldsymbol{\beta}(\boldsymbol{z})\|_2^2 + \lambda\|\boldsymbol{\beta}(\boldsymbol{z})\|_0\right)\right\}$$

とする．ただし，$\lambda(>0)$ は定数である．潜在変数の確率分布を

$$g(\boldsymbol{z}) := P(\boldsymbol{Z} = \boldsymbol{z}) = \prod_{j=1}^{p} \pi_j^{z_j}(1-\pi_j)^{1-z_j} \qquad (\pi_j := P(Z_j = 1))$$

とする．すなわち，$Z_1, Z_2, \ldots, Z_p$ は独立であり，それぞれベルヌーイ分布に従う．EM アルゴリズムによって，$\log P(\mathcal{D} = d)$ を最大にするパラメータ $\boldsymbol{\beta}, \boldsymbol{\pi}$ を求める．この問題を定式化すると，以下の通り．

$$\begin{aligned} \max_{\boldsymbol{\beta},\boldsymbol{\pi}} \quad & \log \sum_{\boldsymbol{z}} g(\boldsymbol{z}) \exp\left\{-\frac{1}{t}\left(\|\boldsymbol{y} - Xb(\boldsymbol{z})\|_2^2 + \lambda\|\boldsymbol{z}\|_0\right)\right\} \\ \text{s.t.} \quad & g(\boldsymbol{z}) = \prod_{j=1}^{p} \pi_j^{z_j}(1-\pi_j)^{1-z_j} \\ & \boldsymbol{\pi} \in [0,1]^p,\ \boldsymbol{\beta} \in \mathbb{R}^{p+1} \end{aligned} \tag{4}$$

求められた $\boldsymbol{\beta}$ と $\boldsymbol{\pi}$ を用いて，係数の推定値を以下の様に求める．

$$E[b(\boldsymbol{Z})] = (\beta_0, \pi_1\beta_1, \ldots, \pi_p\beta_p)^\top$$

ここで，以下の命題が成り立つ．

命題 3.1 (2) の最適解 $\boldsymbol{\beta}^\bullet$ の集合を $\boldsymbol{\mathcal{B}}^\bullet$，(3) の最適解 $(\boldsymbol{\beta}^*, \boldsymbol{z}^*)$ の集合を $(\boldsymbol{\mathcal{B}}^*, \boldsymbol{\mathcal{Z}}^*)$，(4) の最適解 $(\boldsymbol{\beta}^\star, \boldsymbol{\pi}^\star)$ の集合を $(\boldsymbol{\mathcal{B}}^\star, \boldsymbol{\Pi}^\star)$ とする．

$$\boldsymbol{\mathcal{B}}^\bullet \subset \boldsymbol{\mathcal{B}}^* = \boldsymbol{\mathcal{B}}^\star, \qquad \boldsymbol{\mathcal{Z}}^* \subset \boldsymbol{\Pi}^\star$$

である．また，(2) の最適解 $\boldsymbol{\beta}^\bullet$ がただ一つのとき，(3) の最適解 $\boldsymbol{z}^*$ と (4) の最適解 $\boldsymbol{\pi}^\star$ もただ一つであり，

$$\boldsymbol{z}^* = \boldsymbol{\pi}^\star$$

を満たし，

$$\boldsymbol{\beta}^\bullet = b^*(\boldsymbol{z}^*) = b^\star(\boldsymbol{\pi}^\star)$$

が成り立つ．ただし，

$$b^*(\boldsymbol{z}^*) = (\beta_0^*, z_1^*\beta_1^*, \ldots, z_p^*\beta_p^*)^\top$$
$$b^\star(\boldsymbol{\pi}^\star) = (\beta_0^\star, \pi_1^\star\beta_1^\star, \ldots, \pi_p^\star\beta_p^\star)^\top$$

とする．

よって，(2) の最適解がただ一つのとき，EM アルゴリズムによって得られた (4) の最適解 $(\boldsymbol{\beta}^\star, \boldsymbol{\pi}^\star)$ を用いて，(2) の最適解が得られる．(2) の最適解がただ一つかどうかはデータセット d とハイパーパラメータ λ に依存する．

3.2 アルゴリズムの定式化

E ステップにおいて，

$$P(\mathcal{D} = d|\boldsymbol{Z} = \boldsymbol{z}) \fallingdotseq 1 - \frac{1}{t}\left(\|\boldsymbol{y} - X\boldsymbol{\beta}(\boldsymbol{z})\|_2^2 + \lambda\|\boldsymbol{\beta}(\boldsymbol{z})\|_0\right)$$

と近似し，M ステップにおいて，

$$P(\boldsymbol{Z} = \boldsymbol{z}|\mathcal{D} = d) \fallingdotseq \prod_{j=1}^{p} P(Z_j = z_j|\mathcal{D} = d)$$

と近似した値を用いた Q' 関数を最大化する．提案するアルゴリズムは以下の通り．

1. パラメータの初期値 $\boldsymbol{\beta}^{\text{old}} = \boldsymbol{0}, \boldsymbol{\pi}^{\text{old}} = (0.5, \ldots, 0.5)^\top$ とする．
2. 【E ステップ】 事後確率 ($\gamma_j := P(Z_j = 1|\mathcal{D} = d)$) の更新

$$\gamma_j = \pi_j \frac{(\|\boldsymbol{y}\|_2^2 + \lambda p) - \left(\|\boldsymbol{y} - X\boldsymbol{\beta}_j(\boldsymbol{\pi})\|_2^2 + R_j(\boldsymbol{\pi}) + \lambda\boldsymbol{\pi}_j^\top\boldsymbol{1}\right)}{(\|\boldsymbol{y}\|_2^2 + \lambda p) - \left(\|\boldsymbol{y} - X\boldsymbol{\beta}(\boldsymbol{\pi})\|_2^2 + R(\boldsymbol{\pi}) + \lambda\boldsymbol{\pi}^\top\boldsymbol{1}\right)} \qquad (j = 1, \ldots, m)$$

とする．ただし，

$$R(\boldsymbol{\pi}) := \sum_{k=1}^{p} g(\pi_k)\beta_k^2\ ,\ R_j(\boldsymbol{\pi}) := \sum_{k\neq j} g(\pi_k)\beta_k^2$$

$$g(\pi_k) := \pi_k(1-\pi_k)\sum_{i=1}^{N} x_{i,k}^2\ ,\ \boldsymbol{\beta}(\boldsymbol{\pi}) := (\beta_0, \pi_1\beta_1, \cdots, \pi_p\beta_p)^\top$$

$$\boldsymbol{\beta}_j(\boldsymbol{\pi}) := (\beta_0, \pi_1\beta_1, \cdots, \beta_j, \cdots, \pi_p\beta_p)^\top$$

$$\boldsymbol{\pi}_j := (\pi_1, \pi_2, \cdots, 1, \cdots, \pi_p)^\top$$

である．

3. $\gamma_j = 0$ となった変数 j があればそれを取り除く．
4. 【M ステップ】 パラメータ $\boldsymbol{\beta}, \boldsymbol{\pi}$ の更新
   Q' 関数

$$Q' = -\frac{1}{\|\boldsymbol{y}\|_2^2 + \lambda p}\left(\|\boldsymbol{y} - X^{(\boldsymbol{\gamma})}\boldsymbol{\beta}\|_2^2 + R(\boldsymbol{\gamma}) + \lambda\boldsymbol{\gamma}^\top\boldsymbol{1}\right) + \boldsymbol{\gamma}^\top\log\boldsymbol{\pi} + (\boldsymbol{1}-\boldsymbol{\gamma})^\top\log(\boldsymbol{1}-\boldsymbol{\pi})$$

を最大にする $\boldsymbol{\beta}, \boldsymbol{\pi}$ を求める．ただし，

$$X^{(\boldsymbol{\gamma})} := \begin{pmatrix} 1 & \gamma_1 x_{1,1} & \gamma_2 x_{1,2} & \cdots & \gamma_m x_{1,p} \\ 1 & \gamma_1 x_{2,1} & \gamma_2 x_{2,2} & \cdots & \gamma_m x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \cdots \\ 1 & \gamma_1 x_{N,1} & \gamma_2 x_{N,2} & \cdots & \gamma_m x_{N,p} \end{pmatrix}$$

とする．これは解析的に求まり，

$$\boldsymbol{\beta}^{\text{new}} = \left({X^{(\boldsymbol{\gamma})}}^\top X^{(\boldsymbol{\gamma})} + \text{diag}(g(\gamma_1), \ldots, g(\gamma_p))\right)^{-1} {X^{(\boldsymbol{\gamma})}}^\top \boldsymbol{y}$$

$$\boldsymbol{\pi}^{\text{new}} = \boldsymbol{\gamma}$$

である．逆行列の計算が困難であれば，

$$\boldsymbol{\beta}^{\text{new}} = \boldsymbol{\beta}^{\text{old}} + \alpha\frac{\partial Q'(\boldsymbol{\beta}^{\text{old}})}{\partial\boldsymbol{\beta}}$$

とする．ただし，α は学習係数と呼ばれる定数である．

5. $\|\boldsymbol{\pi}^{\text{new}} - \boldsymbol{\pi}^{\text{old}}\|_2^2 < \varepsilon$ を満たしていれば終了．満たしていなければ，

$$\boldsymbol{\beta}^{\text{old}} \leftarrow \boldsymbol{\beta}^{\text{new}}, \qquad \boldsymbol{\pi}^{\text{old}} \leftarrow \boldsymbol{\pi}^{\text{new}}$$

として，2. に戻る．

4　数値実験

4.1　前立腺癌データ

前立腺癌のデータ [2] は 97 個のデータであり，8 個の変数がある．変数の詳しい説明については，Stamey et al. [2] を参照して頂きたい．λ を $2^{-2}, 2^{-1}, 1, \ldots, 2^7$ と変化させた時の推定された係数は図 1 の通り．収束を判定する ε は 10^{-16} とした．

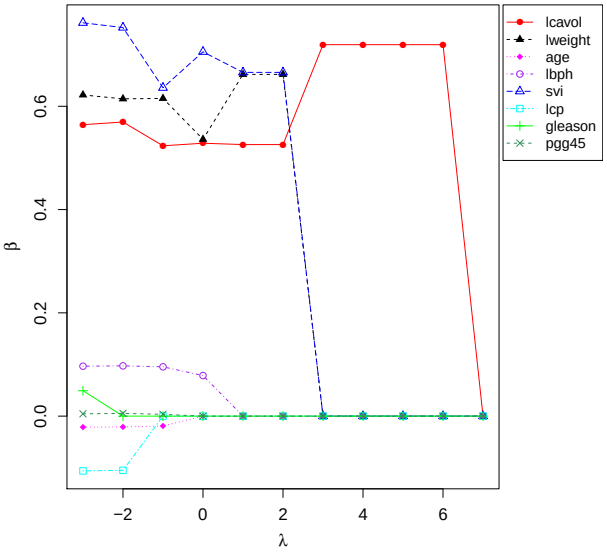

図 1　λ を変化させたときの係数の推定値

一番左 ($\log_2 \lambda = -3$ の位置に対応する) は全ての変数を用いた最小二乗法によって推定された推定値である．各 λ において，係数が 0 にならなかった変数のみを用いて最小二乗法を行った場合の推定値と図 1 の値は一致した．$\lambda = 2^3 \sim 2^6$ において，変数 lcavol は全て同じ推定値であり，この変数を除いて推定値が 0 になっている．LASSO と違い，λ が違っても選ばれている変数が同じであれば，推定値は同じ結果になる．これは，解いている問題が (3) である事を考えれば当然と言える．また，各 λ について，ステップの経過と潜在変数の確率をプロットしたものが図 2(凡例は図 1 と同じ) である．1 に収束する速さが遅い潜在変数の確率ほど，λ を大きくしていったとき，他の潜在変数よりも先に 0 に収束している．例えば，$\lambda = 2^{-1}$ において 1 に収束する速さは pgg45,age,lbph の順に遅く，pgg45 と age については $\lambda = 1$ において 0 に収束している．また，どの λ においても潜在変数の確率は 0 か 1 付近に早く近づくが，値が変化しなくなるまでには時間が掛かっている．

(縦軸:潜在変数の確率，　横軸:ステップ数)

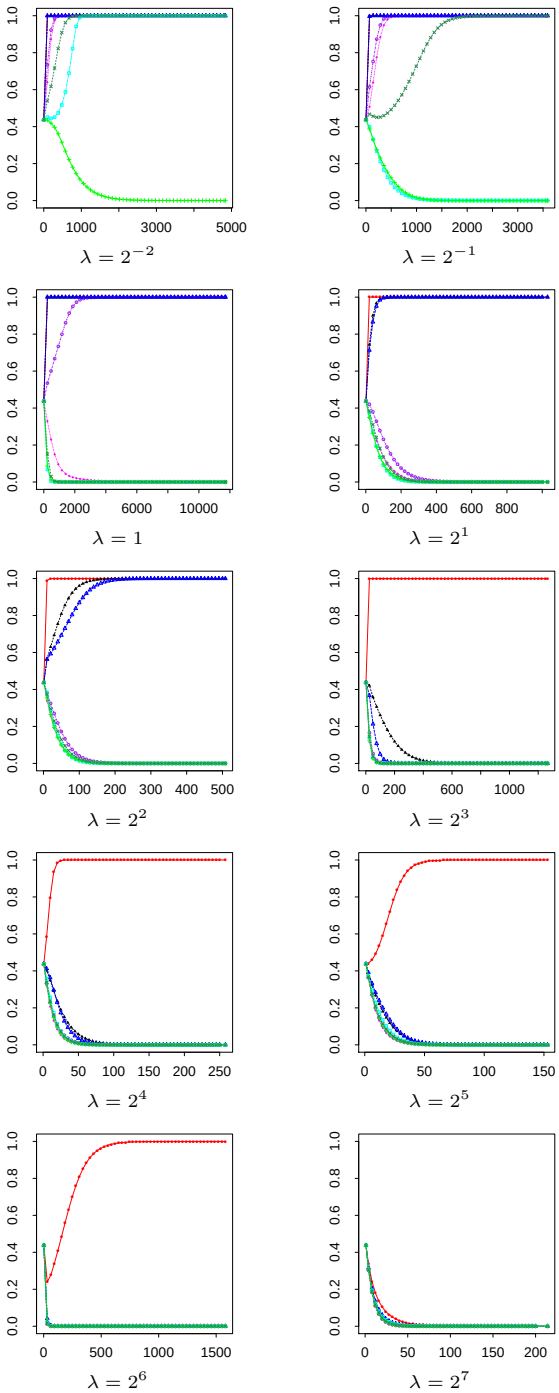

図 2　ステップ数と潜在変数の確率の変化

4.2　シミュレーション

人工データによるシミュレーションを行った．$X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ ，$\boldsymbol{x}_i \sim N(\mathbf{0}, \Sigma)$，$\Sigma = (\sigma_{i,j}) = 0.5^{|i-j|}$

とする説明変数と

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}\ , \qquad \varepsilon_i \sim N(0,1) \qquad (i = 1, \ldots, n)$$

とする従属変数の標本を生成する．0 でない係数 β_j は p 個の内 5 個であり，その 5 つの係数 $\beta = 1, 1.5, 2, 2.5, 3$ とする．$j = 1, \ldots, p$ の内どの 5 つかはランダムに決定されるものとする．上記のモデルに従う標本を訓練・検証・テスト用にそれぞれ 50 個ずつ生成する．訓練データで係数を推定し，検証データを用いてハイパーパラメータ λ を決定する．その後，テスト用データを用いて MSE(平均二乗誤差) を計算しモデルの汎化誤差を評価する．

また，潜在変数の確率は比較的早い繰り返し回数で 0,1 に近づき，$\boldsymbol{\beta}$ は 0 でない係数のみを用いた最小二乗法の解となった 4.1 の結果を踏まえて，アルゴリズムの手順 5 を以下の様に変更する．

5. 以下の (a),(b) どちらかの条件を満たしていれば，終了とする．
   (a) $\|\boldsymbol{\pi}^{\mathrm{new}} - \boldsymbol{\pi}^{\mathrm{old}}\|_2^2 < \varepsilon$
   (b) E ステップと M ステップを I 回以上繰り返している．
   また，(b) で終了した場合，$\boldsymbol{\pi}$ の値を小数点以下を四捨五入した 0,1 の値とし，$\boldsymbol{\beta}$ の予測値を $\pi = 1$ となった変数のみで最小二乗法を行った予測値

$$\boldsymbol{\beta} = \left(X^\top X\right)^{-1} X^\top \boldsymbol{y}$$

   とする．($\pi = 1$ となった変数の数 $> n$ の場合，終了せずに 2. に戻る．)
   満たしていなければ，

$$\boldsymbol{\beta}^{\mathrm{old}} \leftarrow \boldsymbol{\beta}^{\mathrm{new}}\ , \qquad \boldsymbol{\pi}^{\mathrm{old}} \leftarrow \boldsymbol{\pi}^{\mathrm{new}}$$

   として，2. に戻る．

150 個の標本を 200 回発生させて，真の説明変数 5 つを過不足なく選んだ割合は表 1 の通りとなった．

表 1　正答率 (%)

p	(l_0)				(l_1)	(l_2)
	EM:500	EM:1000	EM:2000	GEM:5000	LASSO	Ridge
10	80.5	80.5	81.5	76.0	3.5	0.0
50	87.0	93.0	95.0	92.5	0.0	0.0
100	87.0	89.5	91.0	92.5	0.0	0.0

EM:500,1000,2000 はそれぞれ EM アルゴリズムに基づいたモデル (手順 4 において，逆行列を計算する) において，終了条件の繰り返し回数 $I = 500, 1000, 2000$ とした場合である．GEM:5000 は一般化 EM アルゴリズムに基づいたモデル (手順 4 において，逆行列を計算せず，$\boldsymbol{\beta}$ を勾配方向に更新する) において，終了条件の繰り返し回数 $I = 5000$ とした場合である．$\boldsymbol{\beta}$ の更新の際の学習係数 $\alpha = 1$ とした．また，$p = 10, 50, 100$ は真の説明変数 5 つとダミー変数を含めた説明変数の総数である．Tibshirani [3] のシミュレーション結果からも分かる通り，LASSO は 0 でない説明変数を含む解を選ぶが過不足なく選ぶことは難しい．Ridge 回帰については罰則化項の性質上，完全に係数を 0 にする訳では無い．また，どの説明変数を選ぶかは訓練データと検証データに依存しているため，アルゴリズムによって大域的最適解が得られている場合でも，必ずしも真の説明変数 5 つを選ぶわけでは無いことに注意する．しかし，訓練データにおいて (2) の大域的最適解が得られ，検証データにおいて適切な λ を選択出来ていれば，真の説明変数を選ぶことが期待される．表 1 より，EM アルゴリズムに基づいたモデルはどの繰り返し回数，説明変数の数においても 80(%) を超える正答率を得られた．

5　おわりに

本論文では，線形モデルに関する l_0 正則化付き最小化問題に対して，潜在変数モデルに基づいた定式化を行った．また，線形モデルに関する l_0 正則化付き最小化問題の最適解が潜在変数モデルに基づいた定式化によって得られる最適解に含まれることを示し，最適解を得るアルゴリズムを提案した．数値実験の結果，従来の LASSO と比べて高い割合で真の変数を過不足なく選び，良好な結果を確認できた．

参考文献

[1] Hoai An Le Thi, T Pham Dinh, Hoai Minh Le, and Xuan Thanh Vo. Dc approximation approaches for sparse optimization. *European Journal of Operational Research*, Vol. 244, No. 1, pp. 26–46, 2015.

[2] Thomas A Stamey, John N Kabalin, John E McNeal, Iain M Johnstone, Fuad Freiha, Elise A Redwine, and Norman Yang. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. ii. radical prostatectomy treated patients. *The Journal of urology*, Vol. 141, No. 5, pp. 1076–1083, 1989.

[3] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 58, No. 1, pp. 267–288, 1996.