スパースモデリングとモデル選択

Sparse Modeling and Model Selection

廣瀬　慧

Abstract

本稿は，スパースモデリングの代表的な手法であるLASSO（Least Absolute Shrinkage and Selection Operator）の理論研究に関するサーベイ記事である．まず，従来の変数選択法とLASSOとの関係性を明らかにした，LARSアルゴリズム（Least Angle Regression）を解説する．次に，変数の数が観測数よりも多い場合におけるLASSOの収束レートや変数選択の一致性に関する研究を幾つか紹介する．

キーワード：L_1正則化法，スパース推定，LASSO，LARSアルゴリズム

1. はじめに

スパースモデリングは，ここ10数年，情報学，機械学習，統計学など，様々な分野から注目を集めているが，統計学で最もよく用いられるスパースモデリングは，L_1正則化法(注1)であり，その代表が，Tibshiraniの提案したLASSO（Least Absolute Shrinkage and Selection Operator）[1]である．LASSOは，回帰モデルの損失関数にパラメータのL_1ノルムに基づく正則化項を加えた正則化損失関数を最小化することによってパラメータを推定する方法で，推定の安定化とともに変数選択も行うことができる．

LASSOはここ20年で様々な方向に発展してきたが，その発展の方向性をあえて三つにカテゴライズすると，筆者は以下のようになると考える．

① 推定値を効率的に計算するアルゴリズム
② 推定量の性質
③ モデルや罰則項の拡張

①に関して，最初にLASSOのアルゴリズムとして注目されたものが，LARSアルゴリズム（Least Angle Regression）[2]である．LARSは，単に効率的なアルゴリズムというだけでなく，LASSO推定値が従来の変数選択法と比べてどのような性質を持つのかを語りかけてくれるアルゴリズムでもある．LARSが知られるようになってから，LASSOが爆発的に人気が出たと言っても過言ではない．しかしながら，近年はLARSが用いられることは余りなく，座標降下法[3]やADMM（Alternating Direction Method of Multipliers）アルゴリズム[4]といった，高速かつ汎用性のあるアルゴリズムがよく用いられている．

②についてであるが，LASSOは元々「推定量にあえてバイアスを加えることによって，推定量の分散を小さくし，予測精度を上げる」というものであった．しかし，やはりサンプルサイズが大きいときには，一致性や漸近正規性(注2)といった性質があったほうがうれしい．そこで，Knightら[5]は，サンプルサイズが大きくなったときのLASSO推定量の漸近的性質を調べた．近年は，サンプルサイズが大きいだけでなく，変数の数も多い場合の収束レートや変数選択の一致性の研究が盛んに行われている[6]～[8]．しかしながら，変数の数が多い場

廣瀬　慧　九州大学マス・フォア・インダストリ研究所
E-mail hirose@imi.kyushu-u.ac.jp
Kei HIROSE, Nonmember (Institute of Mathematics for Industry, Kyushu University, Fukuoka-shi, 819-0395 Japan).
電子情報通信学会誌　Vol.99 No.5 pp.392-399 2016年5月

(注1)　正則化法とは，パラメータに何らかの制約を入れて損失関数を最小化する手法である．
(注2)　真のパラメータベクトルを$\boldsymbol{\beta}$，その推定量を$\hat{\boldsymbol{\beta}}$とする．このとき，任意の$\delta>0$に対し，$\lim_{n\to\infty} P(\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|>\delta)=0$を満たす性質を一致性と言い，$\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\to N(\mathbf{0},\Sigma)$（as $n\to\infty$）を満たす性質を漸近正規性と言う．

合，LASSO 推定量が良い性質を満たすためには，計画行列(注3)にかなり強い仮定を置く必要がある．そのため，現実の高次元データに LASSO が使える状況はかなり限られる．

③に関して，LASSO は回帰モデルのみならず，一般化線形回帰モデル[9),(10)]やグラフィカルモデル[11),(12)]，多変量解析[13),(14)]など，様々なモデルに適用されるようになった．また，罰則項に関しても，Group LASSO[15)]や Fused LASSO[16)]など，解析の目的に応じた様々な拡張がある．

①～③に述べたように，LASSO の研究は様々な方向に発展してきたが，余りに多岐にわたりすぎており，これまでの発展を概観することは難しい．そこで本稿では，主に②の LASSO 推定値の性質に焦点を当て，これまで提案されてきた LASSO のアルゴリズムや漸近理論を幾つか紹介する．まず，2. では，代表的な正則化法である Ridge と LASSO の定義とその性質について簡単に述べる．3. では，LASSO 推定値(注4)の性質を調べ，パラメータの推定アルゴリズムを紹介する．4. では，LASSO 推定量(注4)の漸近的性質に関する研究を紹介する．

2. Ridge と LASSO

目的変数 y と p 次元説明変数ベクトル $\boldsymbol{x}=(x_1,\cdots,x_p)^T$ に関する n 組のデータ $\{(y_i,\boldsymbol{x}_i)|i=1,\cdots,n\}$ が得られたとする．ただし，$\boldsymbol{x}_i=(x_{i1},\cdots,x_{ip})^T$ とする．$\boldsymbol{X}=(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)^T$，$\boldsymbol{y}=(y_1,\cdots,y_n)^T$ とする．$\boldsymbol{X}$，$\boldsymbol{y}$ は，それぞれ計画行列，目的変数ベクトルと呼ばれる．下記のように，計画行列 $\boldsymbol{X}$ の各列の長さが n となるように基準化し，$\boldsymbol{X}$ と $\boldsymbol{y}$ をそれぞれ中心化する[注5),(17)]．

$$\sum_{i=1}^{n} y_i=0,\quad \sum_{i=1}^{n} x_{ij}=0,\quad \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2=1,\quad j=1,\cdots,p.$$

ここで，次の線形回帰モデルを仮定する．

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}. \qquad (1)$$

ただし，$\boldsymbol{\beta}=(\beta_1\cdots,\beta_p)^T$ を回帰係数ベクトル，$\boldsymbol{\varepsilon}$ を誤差ベクトルとし，$\boldsymbol{\varepsilon}$ は多変量正規分布 $N(\boldsymbol{0},\sigma^2 I)$ に従うとする．なお，誤差分散 σ^2 は，本来はデータから推定する必要があるが，ここでは簡単のため，既知としておく．

1990 年代頃までは，正則化法と言えば，Ridge 推定[18)]

$$\widehat{\boldsymbol{\beta}}=\arg\min_{\beta}(\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_2^2) \qquad (2)$$

が主流であった．ただし，$\lambda\geq 0$ は正則化パラメータとし，m 次元ベクトル $\boldsymbol{a}=(a_1,\cdots,a_m)^T$ に対し，$\|\boldsymbol{a}\|_q=\left(\sum_{i=1}^{m}|a_i|^q\right)^{1/q}$ とする．$\lambda=0$ のとき，$\widehat{\boldsymbol{\beta}}$ は最小二乗推定値となり，λ が大きくなるにつれて $\widehat{\boldsymbol{\beta}}$ は $\boldsymbol{0}$ に近づく．

式(2)で $\lambda=0$ としたときの最小二乗推定量

$$\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \qquad (3)$$

は不偏推定量で（すなわち，$E[\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}]=\boldsymbol{\beta}$），かつ全ての線形の不偏推定量の中でその分散が最小になる[注6),(19)]．しかしながら，説明変数の次元が高かったり，変数間の相関が大きかったりすると，逆行列 $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ は大きな値を取り[注7)]，その結果最小二乗推定量は不安定になる．一方で，$\lambda>0$ のとき，Ridge 推定量は

$$\widehat{\boldsymbol{\beta}}=(\boldsymbol{X}^T\boldsymbol{X}+\lambda I)^{-1}\boldsymbol{X}^T\boldsymbol{y} \qquad (4)$$

で与えられる．そのため，λ がある程度大きければ，逆行列 $(\boldsymbol{X}^T\boldsymbol{X}+\lambda I)^{-1}$ の固有値が極端に大きな値を取ることがなくなり，推定量の分散が大幅に小さくなる．ただし，余り λ が大きくなりすぎると，推定量のバイアス $(\boldsymbol{X}^T\boldsymbol{X}/\lambda+I)^{-1}\boldsymbol{\beta}$ も大きくなるため，適切な λ を選択しなければならない[注8)]．

Ridge 推定を用いると，安定した推定はできるものの，変数選択ができないという問題がある[注9)]．そこで，Tibshirani[1)]は，Ridge 推定と同様に安定した推定ができ，かつ変数選択も同時に行うことのできる，次式の LASSO を提案した．

(注 3)　説明変数ベクトルを並べた行列．
(注 4)　推定値は，最適化問題を解いたときの解を意味し，推定量は，その最適化問題を解いた解が確率的に変動する，確率変数を意味する．
(注 5)　中心化・基準化をする理由は以下のとおりである．
- 中心化することにより，回帰係数の切片項を 0 として議論することができる．詳しくは小西[17)]3.4.4 を参照されたい．
- 一般に，説明変数の分散が大きくなれば，回帰係数の値は小さくなる．LASSO などの正則化法は，各係数に同じ大きさの制約を入れるため，もし変数を基準化しなければ，分散の大きい変数に対応する係数が 0 と推定されてしまう．それを防ぐために $\boldsymbol{X}$ の各列の二乗和を一定にする．

(注 6)　この性質はガウスマルコフの定理と呼ばれる．例えば Hastie ら[19)]の 3.2.2 を参照されたい．
(注 7)　$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ の最大固有値が最小固有値と比べて極端に大きいことを意味する．
(注 8)　実際は，平均二乗誤差の最小化によって選択することが多い（文献(19)，2.9 節参照）．
(注 9)　総当り法と併用すれば，一応変数選択はできるが，計算負荷が掛かる．

$$\hat{\boldsymbol{\beta}}=\arg\min_{\beta}(\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1). \quad (5)$$

式(2)と式(5)を比較すると，Ridgeでは，罰則項がパラメータの二乗和 $\lambda\|\boldsymbol{\beta}\|_2^2=\lambda\sum_{j=1}^{p}\beta_j^2$ で与えられるが，LASSOでは，罰則項がパラメータの絶対値の和 $\lambda\|\boldsymbol{\beta}\|_1=\lambda\sum_{j=1}^{p}|\beta_j|$ で与えられる．LASSOでは，この「絶対値の和」を用いることにより，パラメータの幾つかを正確に0と推定できる．その結果，パラメータ推定と同時に変数選択も行うことができる．

3. LASSO推定値の性質と計算アルゴリズム

LASSOで問題となるのが，推定値を陽に表すことが難しいということである．Ridge推定値は，式(2)を $\boldsymbol{\beta}$ に関して微分して推定方程式を解けばよく，式(4)のように陽に表される．しかし，LASSOでは，式(5)のペナルティ項 $\lambda\|\boldsymbol{\beta}\|_1$ が $\beta_j=0$ で微分できない．

そこで，パラメータの推定値を計算する効率的なアルゴリズムが必要となる．2010年より前にLASSOの効率的なアルゴリズムとして知られていたのが，LARSアルゴリズム[2]である．LARSアルゴリズムを少し修正したアルゴリズム（ここではmodified LARSと呼ぶ）はLASSO推定値を計算できる．modified LARSは高速であるだけでなく，これまで変数選択法として用いられてきた変数増加法と比べて，LASSOがどのような変数の組合せを選ぶのか，また，推定値がどのような性質があるのかを明らかとしている．

本章では，まず，計画行列 $\boldsymbol{X}$ が直交するとき，すなわち，$\boldsymbol{X}^T\boldsymbol{X}=nI$ のときのLASSOと従来の変数選択法を比較する．次に，$\boldsymbol{X}^T\boldsymbol{X}\neq nI$ のとき，modified LARSアルゴリズムと同等の，KKT条件から導かれるアルゴリズム[20),(21)]について述べる．なお，本章では $n>p$ かつ $\boldsymbol{X}$ がフルランクであると仮定し，議論を進める[注10]．

3.1 $\boldsymbol{X}^T\boldsymbol{X}=nI$ のとき

3.1.1 LASSO推定値について

$\boldsymbol{X}^T\boldsymbol{X}=nI$ のとき，式(5)の罰則付き損失関数は

$$\sum_{j=1}^{p}\{n(\beta_j^2-2\beta_j\hat{\beta}_j^{\mathrm{OLS}})+\lambda|\beta_j|\}+\mathrm{const.} \quad (6)$$

となる．ただし，$\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}=(\hat{\beta}_1^{\mathrm{OLS}},\cdots,\hat{\beta}_p^{\mathrm{OLS}})^T$ とする．なお，$\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ は式(3)で与えられる最小二乗推定量である．式(6)から，係数ベクトル $\boldsymbol{\beta}$ の推定値は，各要素 β_j $(j=1,\cdots,p)$ に対して独立に計算すればよく，下記のように解析的に求めることができる．

$$\hat{\beta}_j=\mathrm{sign}(\hat{\beta}_j^{\mathrm{OLS}})\left(|\hat{\beta}_j^{\mathrm{OLS}}|-\frac{\lambda}{2n}\right)_+. \quad (7)$$

ただし，

$$A_+=\begin{cases}A & A>0\text{のとき}\\ 0 & \text{その他}\end{cases}$$

とする．今，式(3)から，$\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}=\boldsymbol{X}^T\boldsymbol{y}/n$ となる．式(7)から，説明変数と目的変数の相関 $\boldsymbol{X}^T\boldsymbol{y}/n$ の絶対値が $\lambda/(2n)$ より大きいかどうかで変数が選ばれるかどうかが決まる．また，係数の値は，最小二乗推定値を $\lambda/(2n)$ だけ縮小した（0に近づけた）ものとなっている．

3.1.2 t 検定，変数増加法との比較

LASSOによって選ばれる変数と，t 検定，変数増加法によって選ばれる変数の組合せを比較する．まず，t 検定では，$|\hat{\beta}_j^{\mathrm{OLS}}|>s$ のときに，j 番目の変数が選択される．ただし，s は正数で，有意水準の設定に依存する．λ と s をうまく対応させると，t 検定とLASSOは同じ変数を選ぶことが分かる．

次に変数増加法とLASSOを比較する．変数増加法は，まず $\hat{\boldsymbol{\beta}}=\boldsymbol{0}$ とし，各ステップで一つずつ変数を加えていき，p 回のステップで最小二乗推定値を得る．今，k 番目のステップにおける係数ベクトルの推定値を $\hat{\boldsymbol{\beta}}^{(k)}$ とする．$k+1$ ステップでは，残差平方和 $RSS=\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(k+1)}\|^2$ が最小になるような変数を加える．ここで，$\boldsymbol{X}^T\boldsymbol{X}=nI$ なので，

$$RSS=\boldsymbol{y}^T\boldsymbol{y}-\frac{\|\boldsymbol{X}_{\mathcal{A}}^T\boldsymbol{y}\|_2^2}{n} \quad (8)$$

となる．ここで，$\mathcal{A}$ は，$k+1$ ステップでの変数の番号の集合で，$\boldsymbol{X}_{\mathcal{A}}$ は，$\boldsymbol{X}$ の $\mathcal{A}$ に対応する列を取り出した小行列を表す．式(8)から，RSS は，各ステップで最も相関の二乗が大きくなる変数を加えることにより最小となる．よって，変数増加法は，相関 $\boldsymbol{X}^T\boldsymbol{y}$ の絶対値の大きい変数を順に選ぶことと同等である．これは，LASSOで λ を少しずつ小さくして推定していったときに選ばれる変数の組合せと同じになる．

以上から，説明変数間に相関がない場合は，LASSO，t 検定，変数増加法のいずれを用いても，同じ変数の組合せが選ばれることとなる．ただし，LASSOはパラメータを $\lambda/(2n)$ だけ縮小推定するのに対し，t 検定と変数増加法は縮小推定しない．

(注10) $n\leqq p$ のときでもmodified LARSは実行可能である．

3.2 $\boldsymbol{X}^T\boldsymbol{X}\neq nI$ のとき

3.2.1 推定アルゴリズム

変数間に相関があるとき，LASSO 推定値はどのように表されるのだろうか．まず，劣微分の理論から，$\hat{\boldsymbol{\beta}}$ は

$$\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})=\frac{\lambda}{2}\boldsymbol{s} \tag{9}$$

を満たさなければならない(注11),(注12)．ただし，$\boldsymbol{s}$ は劣微分 $\boldsymbol{s}\in\partial\|\hat{\boldsymbol{\beta}}\|_1$ で，次で与えられる．

$$s_j\in\begin{cases}\{\mathrm{sign}(\hat{\beta}_j)\} & \hat{\beta}_j\neq 0,\\ [-1,1] & \hat{\beta}_j=0.\end{cases} \tag{10}$$

それゆえ，λ を与えたとき，$\mathcal{A}=\{j|\hat{\beta}_j\neq 0\}$ と定義すると，$\hat{\boldsymbol{\beta}}_{\mathcal{A}}$ と $\hat{\boldsymbol{\beta}}_{\mathcal{A}^C}$ は次のように表される．

$$\hat{\boldsymbol{\beta}}_{\mathcal{A}}=(\boldsymbol{X}_{\mathcal{A}}^T\boldsymbol{X}_{\mathcal{A}})^{-1}\left\{\boldsymbol{X}_{\mathcal{A}}^T\boldsymbol{y}-\frac{\lambda}{2}\boldsymbol{s}_{\mathcal{A}}\right\},\quad \hat{\boldsymbol{\beta}}_{\mathcal{A}^C}=\boldsymbol{0}. \tag{11}$$

ただし，$\hat{\boldsymbol{\beta}}_{\mathcal{A}}, \boldsymbol{s}_{\mathcal{A}}$ はそれぞれ，$\hat{\boldsymbol{\beta}}, \boldsymbol{s}$ のうち $\mathcal{A}$ に対応する要素のみを取り出したベクトルである．

LASSO アルゴリズムは式(11)を使って構築できる．まず，十分大きな λ に対しては，$\hat{\boldsymbol{\beta}}=\boldsymbol{0}$ である．$\hat{\boldsymbol{\beta}}=\boldsymbol{0}$ となる最小の λ を λ_0 とすると，式(11)から，$\lambda_0=\max_j|\boldsymbol{x}_j^{*T}\boldsymbol{y}|/n$ で与えられる．ただし，$\boldsymbol{x}_j^*$ は $\boldsymbol{X}$ の j 列目を取り出した n 次元ベクトルとする．$j_1=\arg\max_j|\boldsymbol{x}_j^{*T}\boldsymbol{y}|/n$ とする．このとき，λ を λ_0 より少しだけ小さいときの推定値は，$\mathcal{A}=\{j_1\}$ としたときの式(11)で与えられる．更に λ を小さくすると，$\lambda=\lambda_1$ のとき，$|s_j|=1$ となる $j\neq j_1$ が出現する．このときの j を j_2 とし，$\mathcal{A}=\{j_1, j_2\}$ とする．λ が λ_1 より少しだけ小さいときの推定値は，$\mathcal{A}=\{j_1, j_2\}$ としたときの式(11)で与えられる．

以下，同様にして変数を追加または削除していく．変数が追加または削除される瞬間の λ を λ^* とすると，$\lambda=\lambda^*$ となるのは，以下の2パターンある．

① λ が λ^* より少しだけ大きいときに s_j が -1 から1の間の値をとっていたにもかかわらず，$\lambda=\lambda^*$ になった瞬間に $s_j\in\{-1,1\}$ となる．このとき，$\mathcal{A}\leftarrow\mathcal{A}\cup\{j\}$ とする．

② λ が λ^* 以上のときに $s_j\in\{-1,1\}$ であったにもかかわらず，$\lambda=\lambda^*$ より小さくなった瞬間に s_j が -1 から1の間の値をとる．このとき，$\mathcal{A}$ を $\mathcal{A}\leftarrow\mathcal{A}\backslash\{j\}$ とする．

なお，λ^* は陽に解くことができる（文献(21)3.1）．係数ベクトルの推定値 $\hat{\boldsymbol{\beta}}$ は，上記の①または②により $\mathcal{A}$ が変化したら，その変化した $\mathcal{A}$ に対して，式(11)を適用すればよい．

3.2.2 変数増加法との比較

計画行列が直交であるとき，LASSO と変数増加法は同じ変数が選ばれることを示したが，非直交であるときは，LASSO と変数増加法は同じ変数が選ばれるとは限らない．また，推定値に関して，LASSO より変数増加法の方が「貪欲な」アルゴリズムとなる(2)．ここで言う「貪欲」というのは，変数増加法は縮小推定しないために，パラメータを大きめに推定するという意味である．実際，LASSO では，式(11)から，非ゼロの要素に対しては，最小二乗推定値 $(\boldsymbol{X}_{\mathcal{A}}^T\boldsymbol{X}_{\mathcal{A}})^{-1}\boldsymbol{X}_{\mathcal{A}}^T\boldsymbol{y}$ にバイアス項 $(\boldsymbol{X}_{\mathcal{A}}^T\boldsymbol{X}_{\mathcal{A}})^{-1}\frac{\lambda}{2}\boldsymbol{s}_{\mathcal{A}}$ を加えた形になっており，このバイアス項がパラメータを0に縮小する役割を果たしている．なお，$\hat{\boldsymbol{\beta}}_{\mathcal{A}}$ には，逆行列 $(\boldsymbol{X}_{\mathcal{A}}^T\boldsymbol{X}_{\mathcal{A}})^{-1}$ が存在するため，それが特異に近い場合は，LASSO 推定値は不安定になってしまう．

4. LASSO 推定量の性質

LASSO はあえてバイアスを持たせて推定を安定させる方法であるため，サンプルサイズが小さい場合には最小二乗法より予測精度が高いと考えられる．それでは，サンプルサイズが十分大きいときに，LASSO 推定値は真の値に近づくのだろうか．本章では，LASSO 推定量の性質を見ていく．

4.1 p が固定されているとき

今，真のパラメータ $\boldsymbol{\beta}$ が $\boldsymbol{\beta}=(\boldsymbol{\beta}_{S_0}^T, \boldsymbol{\beta}_{S_0^C}^T)^T$ と書けるとする．ただし，$S_0=\{j|\beta_j\neq 0\}$ とする（すなわち，真の非ゼロ要素を表す）．また，$S_0^C=\{1,\cdots,p\}\backslash\{S_0\}$ とする．S_0 の要素の数を s_0 とする．すなわち，s_0 は真の非ゼロ要素の数である．

一般に，スパース推定では，$\hat{\boldsymbol{\beta}}_{S_0^C}$ を正確に $\boldsymbol{0}$ と推定することが重要である．また，非ゼロパラメータ $\hat{\boldsymbol{\beta}}_{S_0}$ に関しても漸近正規性が言えると望ましい．そこで，Fan ら(22)は，良い推定量というのを次のように定義している．

(注11) 詳しくは，R. Tibshirani の講義資料 http://statweb.stanford.edu/~tibs/stat315a/LECTURES/convex-notes.pdf を参照されたい．

(注12) 一見すると $\hat{\beta}_j=0$ のときに $s_j=\pm 1$ となることはなさそうであるが，実際はあり得る．実際，$\hat{\beta}_j=0$ となる最大の λ を λ^* と置くと，$\lambda=\lambda^*$ のときは $s_j=\pm 1$ となる．計画行列 $\boldsymbol{X}$ が直交しているとき，式(7)で，$\lambda=2n|\hat{\beta}_j^{\mathrm{OLS}}|$ のときは $\hat{\beta}_j=0$ かつ $s_j=\pm 1$ となることを確かめることができる．

① $\hat{\boldsymbol{\beta}}_{S_0^C}=\mathbf{0}$

② $\sqrt{n}(\hat{\boldsymbol{\beta}}_{S_0}-\boldsymbol{\beta}_{S_0})\to_d N\left(\mathbf{0}, \sigma^2\Sigma_{S_0}^{-1}\right)$

ただし，$\Sigma_{S_0}=\lim_{n\to\infty}X_{S_0}^TX_{S_0}/n$ とする(注13)．推定量 $\hat{\boldsymbol{\beta}}$ が①，②を満たし，正則化パラメータに関して連続なら，その推定量はオラクルプロパティを持つという[22]．これは，我々はどのパラメータが0かを知らないのだが，それを知ってて最ゆう推定したときと同じ推定量の性質を持っている，ということを意味する．

一般に，LASSO はオラクルプロパティを持たない(文献(23)，2.)．その理由は，LASSO 推定量がバイアスを持つためである．もちろん，罰則を弱くする（具体的には，λ のオーダを $\sqrt{n}$ より小さくする（文献(5)，Theorem 2)）ことによって②の漸近正規性は成り立つ．しかし，そうすると，罰則が弱すぎるため，①の一致性を示すことができない．オラクルプロパティを持たせるには，正則化項として非凸ペナルティ[22]を使うか，Adaptive LASSO[23]などの二段階推定を行う必要がある．

4.2 $n\ll p$ のとき

4.1 では，変数の次元 p が固定された下での理論を述べた．しかしながら，実際に得られるデータは，サンプルサイズ n に比べて変数の次元数 p の方がはるかに大きい場合がある．このような場合，p と n 両方が十分に大きいとき，推定量がどのような性質を持つかを調べる必要がある．以後，$s_0\ll n\ll p$ を想定する．

4.2.1 収束レート

次の二乗誤差リスク

$$R=\frac{1}{n}\|\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 \tag{12}$$

を評価することを考える．まず，最小二乗推定量に対する二乗誤差リスク R の期待値は，

$$E[R]=\frac{1}{n}E\|\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon}\|_2^2=\frac{p}{n}\sigma^2 \tag{13}$$

で与えられる(注14)．これより，p が大きいときは，$E[R]$ は明らかに大きな値を取る．

仮に真の非ゼロ集合 S_0 を知っているとして，最小二乗法によって推定したとすると，

$$E[R]=\frac{s_0}{n}\sigma^2 \tag{14}$$

となる．よって，$s_0\ll p$ のとき（すなわち十分にスパースなとき）リスクの期待値 $E[R]$ は式(13)と比べて十分に小さくなる．もちろん，我々は集合 S_0 を知らないので，式(14)は，「理想的な状況下で得られた推定量に対するリスク」と解釈できる．

我々は集合 S_0 を知らないので，データから推定しなければならない．その推定コストは，実は収束レートに影響する．それを理解するために，L_1 正則化法について議論する前に，次の L_0 正則化法で推定することを考える．

$$\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_0. \tag{15}$$

ただし，$\|\boldsymbol{\beta}\|_0=\sum_{j=1}^{p}1\{\beta_j\neq0\}$ とする．L_0 正則化法では，説明変数の全ての組合せに対して最小二乗推定値を求めて，式(15)の値を計算する．これは，従来の総当り法と同じである．$\lambda=2\sigma^2$ のときは，AIC 若しくは Mallows の C_p 基準に対応し，リスク $E[R]$ の不偏推定量となっている．

一方で，Foster ら[24]は，p が十分大きな場合における式(12)のリスクを評価した．Foster らは，$\lambda=2\sigma^2(\log p+\log\log p)$ としたとき，

$$E[R]\leqq4s_0\sigma^2\left(\frac{\log p}{n}+\frac{\log\log p}{n}\right)+O(n^{-1}) \tag{16}$$

であることを示した（文献(24)，Corollary 5.2)．式(16)は，式(14)に大体 $\log p$ を乗じた形になっている．式(16)は，式(14)よりリスクは大きくなるかもしれないが，式(13)の，全ての変数を使った最小二乗推定値のリスクよりはるかに小さい．

式(16)から，サンプルサイズ n が $\log p$ よりも速いスピードで ∞ になるとすると，$E[R]$ は0に近づく．また，s_0 と σ^2 が小さいほど0に近づきやすい．これは，非ゼロ要素の数 s_0 が少なく，SN 比が高ければリスク $E[R]$ が小さくなることを意味する．

L_0 正則化法は，全ての変数の組合せに対して推定量を計算する必要があり，p が大きいときに計算するのが困難である(注15)，[25]．そこで，L_1 正則化法を使うとどのような収束レートが得られるかを考える．ここで，計画行列 $\boldsymbol{X}$ に対して制約条件を課す．具体的には，ある ϕ_0 が存在して，任意の $\|\boldsymbol{\beta}_{S_0^C}\|_1\leqq3\|\boldsymbol{\beta}_{S_0}\|_1$ を満たす $\boldsymbol{\beta}$ に対して，

(注13) ここでは Σ_{S_0} が退化しないことを仮定している．

(注14) 厳密には $n<p$ のときに最小二乗推定値は存在しないため，ここでは $n>p$ で p が n に十分近いときを考えている．

(注15) 最近は，L_0 正則化に対する効率的なアルゴリズムが提案されている[25]．

$$\|\boldsymbol{\beta}_{S_0}\|_1^2 \leqq (\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta}) s_0 / \phi_0^2 \tag{17}$$

を満たすとする．この条件は compatibility condition と呼ばれる（文献(26)，式(6.4)）．式(17)の条件は，$\|\boldsymbol{\beta}_{S_0}\|_1^2 \leqq s_0 \|\boldsymbol{\beta}_{S_0}\|_2^2$ であること（例えば文献(27)，Appendix 式(A. 3)）に注意すると，$\boldsymbol{X}^T\boldsymbol{X}$ の最小固有値に対する条件のように見える．一般に，変数間に相関があれば $\boldsymbol{X}^T\boldsymbol{X}$ の最小固有値は小さくなるため，変数間の相関が小さければ compatibility condition を満たしやすいと考えられる．しかし，$\|\boldsymbol{\beta}_{S_0^C}\|_1 \leqq 3\|\boldsymbol{\beta}_{S_0}\|_1$ という条件があるため，実際は，最小固有値に対する条件よりも弱い条件となっている．

式(17)を満たすとき，任意の $0<\delta<1$ に対して，$\lambda = 4\sigma\sqrt{2n\log(2p/\delta)}$ と置くと，確率 $1-\delta$ で

$$R \leqq C s_0 \sigma^2 \frac{\log p}{n} \frac{1}{\phi_0^2} + O(n^{-1}) \tag{18}$$

が成り立つ．証明は文献(26)の Corollary 6.2，あるいは文献(28)の定理 6 を参照されたい．

ここで，式(18)の結果と，L_0 ペナルティに対する式(16)の結果は，漸近的に似た結果になっている．実際，n が $\log p$ よりも速いスピードで ∞ になるとすると，（収束の仕方は異なるものの）R は 0 に収束する．しかし，大きな違いが一つある．それは，L_0 正則化法に対しては，$\boldsymbol{X}$ に対して何も条件を課さなかったが，L_1 正則化法に対しては，$\boldsymbol{X}$ に対して compatibility condition が必要になるということである．

なお，compatibility condition のほかにも restricted eigenvalue condition[6]等，様々な条件に対する収束レートがある．詳しくは文献(26)を参照されたい．

4.2.2　モデル選択の一致性

L_1 正則化法のモデル選択の一致性について議論する．ここで，

$$\mathrm{sign}(\widehat{\boldsymbol{\beta}}) = \mathrm{sign}(\boldsymbol{\beta})$$

を満たすとき，「モデル選択の一致性を持つ」と定義する．一般に，モデル選択の一致性というと，パラメータがゼロであるか否かを正しく判定できるかということを議論するが，ここではそれに加え，符号の一致性まで考えているため，一般のモデル選択の一致性よりも強い．モデル選択の一致性については文献(7)，(8)で詳しく述べられており，文献(26)に分かりやすくまとめられている．ここでは，文献(7)の結果を紹介する．

まず，計画行列 $\boldsymbol{X}$ と係数ベクトル $\boldsymbol{\beta}$ に対して条件を課す．ある定数 γ，C_0 が存在し，

$$\left\| \left| (\boldsymbol{X}_{S_0^C}^T \boldsymbol{X}_{S_0}) (\boldsymbol{X}_{S_0}^T \boldsymbol{X}_{S_0})^{-1} \right| \right\|_\infty \leqq 1-\gamma \tag{19}$$

$$\min_{j \in S_0^C} |\beta_j^0| > g(\lambda)/(2n) \tag{20}$$

を満たすとする．ただし，$g(\lambda) = \lambda[\|(\boldsymbol{X}_{S_0}^T \boldsymbol{X}_{S_0}/n)^{-1}\|_\infty + 4\sigma/\sqrt{C_0}]$ とする．また，m 次元ベクトル $\boldsymbol{a}$ に対し，$\|\boldsymbol{a}\|_\infty = \max_j |a_j|$ とし，$m \times m$ 行列 $\boldsymbol{A} = (\boldsymbol{A}_{ij})$ に対し，$|||\boldsymbol{A}|||_\infty = \max_i \sum_{j=1}^m |\boldsymbol{A}_{ij}|$ とする．また，λ は

$$\lambda > 4\sigma/\gamma \cdot \sqrt{2n \log p} \tag{21}$$

を満たすとする．

式(19)の条件は，incoherence condition と呼ばれ，計画行列に対する条件である．なお，式(19)と似た条件として，irrepresentable condition[8]があり，文献(26)の 7.5.4 では，irrepresentable condition は式(17)の compatibility condition より強い条件であることを示している．また，式(20)は，最小の非ゼロの係数が小さ過ぎないという条件である．

式(19)，式(20)が成り立ち，更に，$\Lambda_{\min}(\boldsymbol{X}_{S_0}^T \boldsymbol{X}_{S_0}/n) > C_1$（ただし，$\Lambda_{\min}(\cdot)$ は最小固有値，C_1 は n に依存しない定数）を仮定すると，次が成立する（文献(23)，Theorem 1）．

$$P[\mathrm{sign}(\widehat{\boldsymbol{\beta}}) = \mathrm{sign}(\boldsymbol{\beta})] \geqq 1 - 4\exp(-c_1\lambda^2/n) \tag{22}$$

ただし，c_1 は定数とする．式(22)から，$p>n\to\infty$ のとき，$\mathrm{sign}(\widehat{\boldsymbol{\beta}}) = \mathrm{sign}(\boldsymbol{\beta})$ となる確率が 1 に収束し，モデル選択の一致性を持つ．

4.3　考察

4. では，LASSO の理論について述べたが，LASSO を使おうとする工学エンジニアにとって，各解析結果をどのように解釈すればいいのであろうか．4.2 の結果から，高次元データに対し，LASSO 推定値が良い性質を満たすためには，計画行列 $\boldsymbol{X}$ に対して compatibility condition や incoherence condition を満たす必要がある．実際問題，真の $\boldsymbol{\beta}$ が分からないため，これらの条件を満たすかどうかをチェックすることは難しい．しかしながら，compatibility condition や incoherence condition を見ると，変数間の相関が大きい場合に，LASSO がうまく機能しない傾向にあるということは言える．そのため，$\boldsymbol{X}^T\boldsymbol{X}$ の最大固有値と最小固有値の比が大きければ大きいほど，LASSO はうまく機能しにくくなると考えられる．

そのような場合は，縮小ランク回帰[29),(30)]が有効に機能する．これは，相関の高い説明変数を一つの変数としてまとめ，そのまとめた変数を使って回帰分析を行う手法である．

5. おわりに

本稿では，LASSOと従来の変数選択法との関係性を調べ，推定値を求めるアルゴリズムについて述べ，推定量の漸近的性質について解説した．こうして見ると，最初にLASSOが提案されたときは，Ridge回帰のようにパラメータを0に縮小推定することにより，推定量に少しだけバイアスを加えて分散を大幅に小さくするというコンセプトであったが，近年は，高次元での漸近理論へとシフトしている印象がある．本稿では述べなかったが，最近はLASSOの検定も提案されており[31]，$n<p$の場合，変数選択の一致性で使われるirrepresentable conditionと似た条件が使われている．

4.2で述べたように，$n \ll p$のときのLASSOの漸近的性質を示すには，かなり強い条件が必要である．そのため，実際に解析する際は，得られたデータがその条件を満たすかどうかを慎重に判断する必要がある．具体的には，irrepresentable conditionまたはincoherence conditionを満たし，非ゼロパラメータの最小値がある程度大きな値を持ち，SN比が高く，非ゼロパラメータ数s_0が少ない，という条件を全て満たすかどうかを確認しなければならない．しかしながら，そのような条件を満たす高次元データは余りない，と言ってもいいのではないだろうか．例えば，LASSO回帰がよく使われるのが遺伝子データであるが，遺伝子データは変数間の相関が大きいため，irrepresentable conditionのような条件を満たすとは思えない．

以上から，計画行列$\boldsymbol{X}$があらかじめ与えられているLASSO回帰そのものは，高次元データに余り使えないかもしれない．しかしながら，計画行列$\boldsymbol{X}$がデータとして与えられているわけでなく，自分でデザインすることができるとし，$\boldsymbol{X}$としてランダム行列を用いてL_1ノルムを使った制約付き最適化問題を解くと，高い確率で係数を正しく推定できる．この性質を画像復元にうまく適用した方法が，圧縮センシング[32]である．また，冒頭に述べたように，LASSOには様々な拡張したモデルがあり，その中には実用的な手法が幾つかある．例えば，Fused LASSOを使えば，画像の雑音除去ができる[16]．

このように，LASSOを何も考えずに使うと，全くうまく機能しないことが多々あるが，LASSOの性質を理解し，その性質をうまく利用すれば，意味のあるデータ解析ができる．

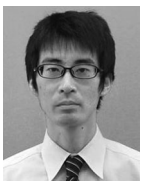
謝辞　査読者には原稿を読んで頂き，有益なコメントを頂きました．また，大阪大学の伊森晋平氏には，LASSOの理論に関するコメントを頂きました．ここに深謝致します．

文　献

(1) R. Tibshirani, "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society : Series B (Statistical Methodology), vol. 58, no. 1, pp. 267-288, 1996.
(2) B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression (with discussion)," The Annals of Statistics, vol. 32, pp. 407-499, 2004.
(3) J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," Biostatistics, vol. 9, no. 3, pp. 432-441, 2008.
(4) S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," Foundations and Trends® in Machine Learning, vol. 3, no. 1, pp. 1-122, 2011.
(5) K. Knight and W. Fu, "Asymptotics for lasso-type estimators," The Annals of Statistics, vol. 28, no. 5, pp. 1356-1378, 2000.
(6) P.J. Bickel, Y. Ritov, and A.B. Tsybakov, "Simultaneous analysis of lasso and dantzig selector," The Annals of Statistics, vol. 37, no. 4, pp. 1705-1732, 2009.
(7) M.J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso)," IEEE Trans. Inf. Theory, vol. 55, no. 5, pp. 2183-2202, 2009.
(8) P. Zhao and B. Yu, "On model selection consistency of lasso," J. Mach. Learn. Res., vol. 7, no. 2, p. 2541, 2007.
(9) J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," Journal of Statistical Software, vol. 33, no. 1, pp. 1-22, 2010.
(10) M.Y. Park and T. Hastie, "L1-regularization path algorithm for generalized linear models," Journal of the Royal Statistical Society : Series B (Statistical Methodology), vol. 69, no. 4, pp. 659-677, 2007.
(11) N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," The Annals of Statistics, vol. 34, no. 3, pp. 1436-1462, 2006.
(12) M. Yuan and Y. Lin, "Model selection and estimation in the gaussian graphical model," Biometrika, vol. 94, no. 1, pp. 19-35, 2007.
(13) D.M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," Biostatistics, vol. 10, no. 3, p. kxp008, 2009.
(14) H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," Journal of Computational and Graphical Statistics, vol. 15, no. 2, pp. 265-286, 2006.
(15) M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," Journal of the Royal Statistical Society : Series B (Statistical Methodology), vol. 68, no. 1, pp. 49-67, 2006.
(16) R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," Journal of the Royal Statistical Society : Series B (Statistical Methodology), vol. 67, no. 1, pp. 91-108, 2005.
(17) 小西貞則，多変量解析入門—線形から非線形へ，岩波書店，東京，2010.
(18) A.E. Hoerl and R.W. Kennard, "Ridge regression : Biased estimation for nonorthogonal problems," Technometrics, vol. 12, no. 1, pp. 55-67, 1970.
(19) T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed., Springer, New York, 2008.
(20) R. Tibshirani, "Sparsity and the lasso," 2015, http://www.stat.cmu.edu/~larry/=sml/sparsity. pdf
(21) R.J. Tibshirani, "The lasso problem and uniqueness," Electronic Journal of Statistics, vol. 7, pp. 1456-1490, 2013.
(22) J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," J. Am. Stat. Assoc., vol. 96, no. 456, pp. 1348-1360, 2001.
(23) H. Zou, "The adaptive lasso and its oracle properties," J. Am. Stat. Assoc., vol. 101, no. 476, pp. 1418-1429, 2006.
(24) D.P. Foster and E.I. George, "The risk inflation criterion for multiple regression," The Annals of Statistics, vol. 22, no. 4, pp. 1947-1975, 1994.
(25) S. Xiong, "Better subset regression," Biometrika, vol. 101, no. 1, pp.

71-84, 2014.
(26) P. Bhlmann and S. van de Geer, Statistics for High-Dimensional Data : Methods, Theory and Applications, 1st ed., Springer Publishing Company, Incorporated, 2011.
(27) S. Foucart and H. Rauhut, A mathematical introduction to compressive sensing, Springer, 2013.
(28) 鈴木大慈，“スパース推定における確率集中不等式(高次元量子トモグラフィにおける統計理論的なアプローチ)，”数理解析研究所講究録，vol. 1908, pp. 39-48, 2014.
(29) L. Chen and J.Z. Huang, “Sparse reduced-rank regression for simultaneous dimension reduction and variable selection,” J. Am. Stat. Assoc., vol. 107, no. 500, pp. 1533-1545, 2012.
(30) M. Vounou, T.E. Nichols, G. Montana, and A.D.N. Initiative, “Discovering genetic associations with high-dimensional neuroimaging phenotypes : A sparse reduced-rank regression approach,” Neuro-image, vol. 53, no. 3, pp. 1147-1159, 2010.
(31) R. Lockhart, J. Taylor, R.J. Tibshirani, and R. Tibshirani, “A significance test for the lasso,” The Annals of Statistics, vol. 42, no. 2, pp. 413-468, 2014.
(32) E.J. Candès, J. Romberg, and T. Tao, “Robust uncertainty principles : Exact signal reconstruction from highly incomplete frequency information,” IEEE Trans. Inf. Theory, vol. 52, no. 2, pp. 489-509, 2006.

(平成 27 年 11 月 2 日受付　平成 27 年 12 月 9 日最終受付)



廣瀬　慧（ひろせ　けい）

平 19 九大・理・数学卒．平 23 同大学院博士課程了．現在，九大マス・フォア・インダストリ研究所准教授．統計科学，機械学習，スパースモデリングの研究に従事．機能数理学博．