# スパース推定の理論

北海道大学　経済学部　渡部元博

2022/12/23

## 目次

# 1　はじめに

川野秀一・松井秀俊・廣瀬慧 [2018]『統計学　One Point　6　スパース推定法による統計モデリング』共立出版を元に R の文書作成ツールである quarto を使用し、資料作成を行いました。

# 2　線形回帰モデルと lasso

## 2.1　最小二乗法・正則化法

連続値をとる目的変数 $Y(\in \mathbb{R})$ と p 次元説明変数 $\boldsymbol{x} = (x_1, ..., x_p)^T$ に関して、n 個の観測によりデータ $((y_i, \boldsymbol{x}_i) : i = 1, ..., n)$ が得られたとする。ただし、$\boldsymbol{x}_i = (x_{i1}, ..., x\_ip)^T$ とする。正則化法では、通常あらかじめ目的変数及び説明変数を中心化し、説明変数の各変数の長さを $\sqrt{n}$ とする。すなわち、

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

を満たすようにする。基本的には、変数について中心化されたものを用いることにする。正則化法で説明変数の長さを同じにする理由については後述する。

ここで線形回帰モデル

$$y_i = x_{i1}\beta_i + ... + x_{ip}\beta_p + \epsilon_i, \quad (i = 1, .., n) \tag{2}$$

回帰係数を推定するために最もよく用いられる方法は、**誤差二乗和**すなわち、回帰モデルの左辺から右辺の平均構造を引いたものの二乗和

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \tag{3}$$

を最小にする $\boldsymbol{\beta} = (\beta_1, ..., \beta p)^T$ を求める最小二乗法である。最小二乗法によって得られる推定量は最小二乗推定量と呼ばれる。上記の (2) 式及び (3) 式は、変数やパラメータをベクトルや行列を用いることで表現が簡潔になる。いま、$X = (\boldsymbol{x}_{(1)}, ..., \boldsymbol{x}_{(p)})$(これは計画行列と呼ばれる),$\boldsymbol{x}_{(j)} = (x_{1j}, ..., x_{nj})^T, \boldsymbol{y} = (y_1, ..., y_n)^T$ とおくと、回帰モデル式 (2) は、

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.a}$$

と表すことができ、誤差二乗和 (3) は

$$S(\boldsymbol{\beta}) = \|\boldsymbol{\epsilon}\|_2^2 = \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 \tag{3.a}$$

{eq-3.a}

と表すことができる。ただし、$\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^T$ である。回帰係数の最小二乗推定量は $S(\boldsymbol{\beta})$ をベクトル $\boldsymbol{\beta}$ について偏微分することで得られる

$$\frac{\partial(S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2X^T(\boldsymbol{y} - X\boldsymbol{\beta}) = \boldsymbol{0} \tag{4}$$

解くことによって得られる。この解、すなわち $\boldsymbol{\beta}$ の最小二乗推定量は、$X^TX$ が正則のとき、

$$\hat{\boldsymbol{\beta}}^{OLS} = (X^TX)^{-1}X^T\boldsymbol{y} \tag{5}$$

で与えられる。

ところが、データの構造や性質によっては、最小二乗法では適切な推定量が得られない場合がある。例えば、

- 説明変数間の相関係数が非常に高い
- 説明変数の数 (p) がサンプルサイズ (n) に近い、あるいは超えている

状況では、(5) 式中の逆行列 $(X^TX)^{-1}$ が計算できない、あるいは各要素の値が極端に大きくなるといった現象が起きてしまう。このような問題を解消するために、正則化法と呼ばれる手法がよく用いられる。まず、ベクトル $\boldsymbol{\beta}$ の実数値関数 $R(\boldsymbol{\beta})(\geq 0)$ を用意する。正則化法とはすなわち、$S(\boldsymbol{\beta})$に$R(\boldsymbol{\beta})$ を加えた式の最小化

$$\underset{\boldsymbol{\beta}}{min} S_\lambda(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}}{min}\left(\frac{1}{2n}S(\boldsymbol{\beta}) + \lambda R(\boldsymbol{\beta})\right)$$

により、パラメータの推定値を知る方法である。第一項に 1/2n が乗じてあるのは、後述するアルゴリズムの形を簡便にするためである。ここで、$R(\boldsymbol{\beta})$ は**正則化項**と呼ばれ、$\lambda(\geq 0)$ は**正則化パラメータ**と呼ばれる。$\lambda$ を大きくすると正則化項の影響は大きくなり、$\lambda$ を小さくすると正則化項の影響は小さくなる。$\lambda = 0$ の時は最小二乗法となり、この時得られる推定量は OLS 推定量となる。正則化項の関数形と正則化パラメータを適切に選ぶことで OLS 推定量よりも安定した推定量を得ることができるようになる。

## 2.2 リッジ回帰

線形回帰モデル (2) を正則化法によって推定する際、パラメータに関する $L_2$ ノルムを正則化項に用いたものは、リッジ回帰 (Hoerl and Kennard, 1970) とよばれ、最小化問題

$$\underset{\boldsymbol{\beta}}{min} S_\lambda(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}}{min}\left(\frac{1}{2n}\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{\beta}\|_2^2\right) \tag{6}$$

として定式化される。パラメータ $\boldsymbol{\beta}$ の推定量は、$S_\lambda(\boldsymbol{\beta})$ を $\boldsymbol{\beta}$ について偏微分して得られる推定方程式

$$\frac{\partial(S_\lambda(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = -\frac{1}{n}X^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \lambda\boldsymbol{\beta} = \boldsymbol{0}$$

を解くことによって得られる。実際の推定量は、

$$\hat{\boldsymbol{\beta}^{ridge}} = (X^TX + n\lambda I_p)^{-1}X^T\boldsymbol{y} \tag{7}$$

となる。これにより、最小二乗法において行列 $X^TX$ の逆行列が計算できないような状況においても、行列の対角成分に尾根 (リッジ) を作った $X^TX + n\lambda I_p$ は任意の $\lambda > 0$ に対して正則になる。これが「正則化」とよばれる所以である。(Appendix. 参照)

## 2.3 lasso

Tibshirani(1996) は線形回帰モデルの回帰係数の推定法として、**lasso**(least absolute shrinkage and selection operator) と呼ばれる次の制約付き最小化問題

$$\min_{\boldsymbol{\beta}} \frac{1}{2n}\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 \quad subject \quad to \quad \|\boldsymbol{\beta}\|_1 \leq s \tag{8}$$

による方法を提案した。ただし、$\|\cdot\|_1$ はベクトルの $L_1$ ノルム、すなわち $\|\boldsymbol{\beta}\|_1 = |\beta_1| + \cdots + |\beta_p|$ を表し、s は制約の強さを調節するパラメータである。lasso に基づいてパラメータの推定値を推定することで、いくつかのパラメータの推定値をちょうど 0 に縮小する性質を持つ。そして、0 と推定された係数に対応する説明変数は目的変数に寄与しないと解釈することでができ、疎なモデル (変数選択されたモデルの意味) を構築できる。このようにデータ発生の疎構造を表すモデルを推定する方法は、**スパース推定** (sparse estimation) と呼ばれている。一般に、s を大きくすると 0 と推定されるパラメータの数は少なくなり、s を小さくすると 0 と推定されるパラメータの数は多くなる。

問題 (8) により得られる解は、(8) に対して、ラグランジュの未定乗数法を適用することにより得られる関数

$$S_\lambda(\boldsymbol{\beta}) = \frac{1}{2n}\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \tag{9}$$

をパラメータ $\boldsymbol{\beta}$ に関して最小化することにより得られる解と同値になることが、カルーシュ-クーン-タッカー条件より示される。(9) をよく観察すると、正則化法の形を一致していることがわかる。つまり、lasso は正則化法として捉えることができる。特に、lasso はスパース推定を行うことが可能な正則化法であるため、スパース正則化法 (sparse regularization) とよばれる。正則化パラメータ $\lambda$ は調整パラメータ s と対応しており、$\lambda$ が大きくなることと s が小さくなることが対応している。つまり、$\lambda$ を大きくすると 0 と推定されるパラメータの数は多くなる (逆も然り)。リッジ推定では推定量を 0 の方向に縮小することはできるが、lasso は「ぴったり」0 に縮小することができる。

### 2.3.1 R による lasso の実装

R で lasso を利用する際にはパッケージ glmnet をよく用いる。ここで、アメリカ合衆国の都市における犯罪データに関するデータに対して lasso を適用した分析を実践する。このデータは 6 つの項目

| | |
|---|---|
| $Y$ | 人口 100 万人あたりの犯罪率 |
| $X_1$ | 警察への年間資金 |
| $X_2$ | 25 歳以上のうち高校を卒業した人口の割合 |
| $X_3$ | 16~19 歳のうち高校に通っていない (卒業していない) 人の割合 |
| $X_4$ | 18~24 歳のうち大学生の割合 |
| $X_5$ | 25 歳以上のうち 4 年制大学を卒業した人の割合 |

を 50 のアメリカの都市から取得したものである。ここでは、$X_1, ..., X_5$ を説明変数、$Y$ を目的変数として回帰分析を行うことを考える。データは (1) に従い、中心化と標準化を行った。以下 R プログラムである。

```
library(glmnet)
```

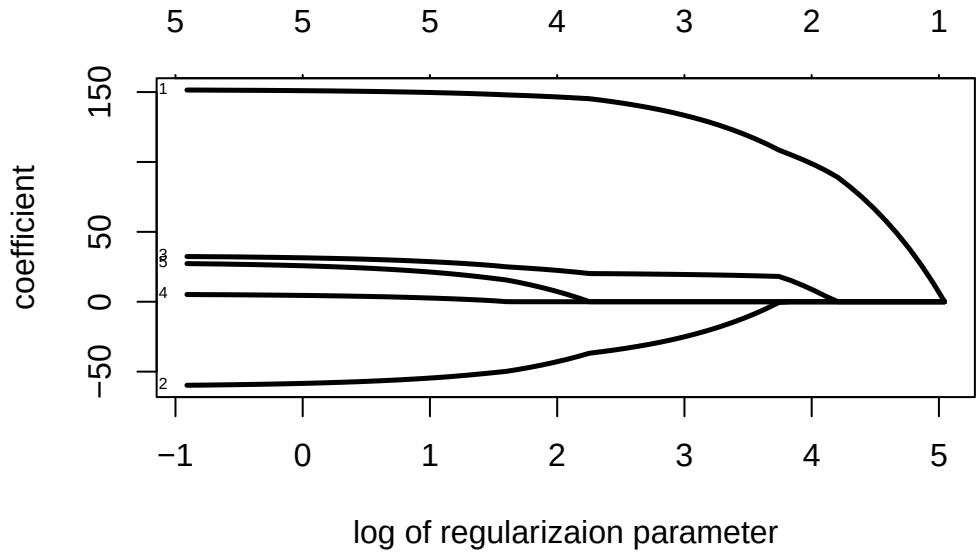
```
Warning: package 'glmnet' was built under R version 4.1.1
```

```
Loading required package: Matrix
```

```
Loaded glmnet 4.1-4
```

```
crime <- read.table("crime.txt")
crime <- as.matrix(crime)
X <- crime[, 3:7]  # 説明変数
y <- crime[, 1]    # 目的変数
X <- scale(X)      # 説明変数を標準化
y <- y - mean(y)   # 目的変数を中心化

# Lasso 推定
res <- glmnet(x=X, y=y)
# 解パス図描画
plot(res, xvar="lambda", label=TRUE, xlab="log of regularizaion parameter",
     ylab="coefficient", col="black", lwd=2.5)
```




```
# 正則化パラメータの値を 20 と固定
res1 <- glmnet(x=X, y=y, lambda=20)
res1$beta  # 係数の推定値
```

```
5 x 1 sparse Matrix of class "dgCMatrix"
          s0
V3 133.50551
V4 -25.22804
V5  19.45576
V6   .
V7   .
```

R の解析の結果、説明変数 $X_4, X_5$ の係数が 0 と推定されていることがわかる。これを予測式で表すと、

$$Y = 133.5X_1 - 25.2X_2 + 19.5X_3 + 0 * X_4 + 0 * X_5$$

となる。この予測式より、大学生の割合と大学を卒業した人の割合は犯罪率に影響しないことがわかる。

上図は、正則化パラメータ $\lambda$ の自然体数値を横軸に、それに応じた回帰係数の推定値 $\hat{\beta}_j (j = 1, ..., 5)$ の推移を縦軸に表したもので、**解パス** (solution path) 図とよばれている。また、各解パスの左側についている数字はそれぞれ説明変数の番号を表している。全ての変数が $\lambda$ の値が大きくなるにつれて $\hat{\beta}_j$ は縮小され、ある位置において 0 に縮小されていることがわかる。

## 2.4 正則化パラメータ $\lambda$ の決定方法

lasso 推定値は正則化パラメータ $\lambda$ の値に依存している。特に、lasso では推定値の違いは変数選択の結果の違いに直接影響するため、正則化パラメータの値の選択は非常に重要な問題である。ここでは、交差検証法と拡張 BIC によるパラメータ選択について紹介する。

### 2.4.1 交差検証法

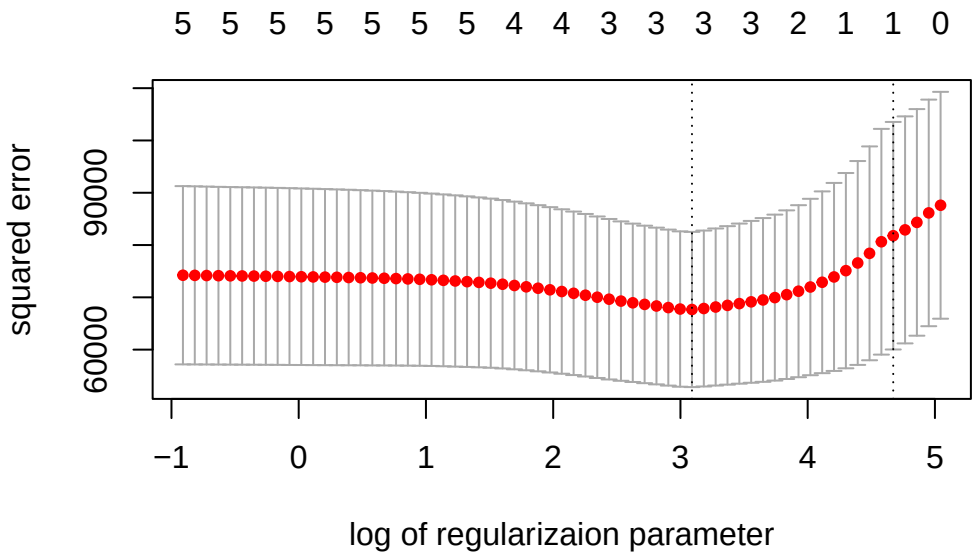
推定されたモデルの良さの観点の一つに、将来新たに観測されるデータに対する予測精度がある。これは、現在得られているデータではなく、将来観測されるデータへの当てはまりの良さを評価すべきであるという考えに基づいている。しかし、実際には将来のデータを得ることはできない。そこで、将来のデータを現在観測されているデータに置き換えてモデルの良さを評価したものが、誤差二乗和 (3) であった。しかし、誤差二乗和は観測されたデータの当てはまりのみに注目した基準であるため、観測されたデータに火適合したモデルをもっとも良いモデルと判断してしまう。

交差検証法 (cross-validation) は、観測されたデータの中から擬似的に将来のデータを分解することにより、将来観測されるデータの予測という観点に基づき構成される評価基準である (Stone, 1974)。いま、n 個のデータを K 個のデータ集合 $D_1, ..., D_k$ に分割する。続いて、k 番目のデータ集合 $D_k$ のみを除いた (K-1) 個のデータ集合を学習データとして用いて係数パラメータを推定し、得られた推定量を $\hat{\boldsymbol{\beta}}^{(-k)}$ とおく。そして、パラメータ推定に用いなかった残りのデータ集合 $D_k$ を検証データとして用いて、検証誤差 $CV_k = \sum_{i \in D_k} \left( y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}^{(-k)} \right)^2 / n_k$ を計算する。ただし、$n_k$ はデータ集合 $D_k$ の大きさとする。この計算を繰り返すことで得られる K 個の検証誤差の平均

$$CV = \frac{1}{K} \sum_{k=1}^{K} CV_k$$

を、モデル選択の基準値として用いる。lasso においても交差検証法はよく用いられており、パッケージ glmnet にも実装されている。

```
# CV の計算
res.cv <- cv.glmnet(x=X, y=y)
# CV 値の推移をプロット
plot(res.cv, xlab="log of regularizaion parameter", ylab="squared error")
```




```
# CV 値が最小となる正則化パラメータ値を出力
res.cv$lambda.min
```

```
[1] 21.99244
```

```
# 1 標準誤差ルールにより選択された正則化パラメータの値を出力
res.cv$lambda.1se
```

```
[1] 106.9405
```

R の解析により、CV が最小になる正則化パラメータが $log\lambda = 3$ であるモデルを最適なモデルであるとわかった。また、交差検証法で $\lambda$ の値を選択する基準として、次を考えることもできる。いま、CV がある $\lambda$ で最小値 $CV_0$ をとるとし、$CV_0$ における K 個の検証誤差の標準誤差を $se_0$ とおく。この時、CV の値が $CV_0 + se_0$ よりも小さくなる最大の $\lambda$ を最適な正則化パラメータとして選択する。これは、CV の値が小さい中でも、できるだけ単純なモデル ($\lambda$ が大きいモデル) を選択しようという考え方に基づいている。このデータは **1 標準誤差ルール**とよばれている。右の破線はこのルールによって選択された $\lambda$ の値を示している。Leng et al.(2006) によると、真に目的変数に関与している変数を選択することが分析の目的の場合、交差検証法のような予測誤差に

基づき導出された評価基準は、変数選択に関する一致生が保証されないといった指摘がある。ここで、サンプルサイズ n を大きくするにつれ、正しい変数の組み合わせを選択する確率が 1 に近づくとき、変数選択に関する一致生を持つという。変数選択に関する一致生を有するモデル評価基準に関しては、次節を参照されたい。

### 2.4.2 拡張 BIC

Chen and Chen(2008) によって提案された**拡張 BIC** は、n<p の状況でも変数選択に関する一致生を持つ。なお、拡張 BIC は lasso などの正則化法によって推定されたモデルを評価する基準でなく、最尤法によって推定されたモデルを評価する基準である。それゆえ、拡張 BIC は、全ての変数の組み合わせを探索する**部分集合選択**に適用することが望ましい。以後、部分集合選択における拡張 BIC を解説する。　いま、変数番号の部分集合 $m \subset \mathcal{J} = (1, ..., p)$ が与えられたとし、$|m| = j$ とする。また、変数の数が j であるようなモデルの集合族を $\mathcal{M}_j$ とする。このとき、Chen and Chen(2008) は次の拡張 BIC を提案した。

$$BIC_\gamma(m) = n \log \hat{\sigma}^2 + j \log n + 2\gamma \log \tau(\mathcal{M}_j) \tag{10}$$

$\tau(\mathcal{M}_j)(= \binom{p}{j})$ はモデル集合 $\mathcal{M}_j$ の大きさである。また、$\gamma(0 \leq \gamma \leq 1)$ は調整パラメータとする。(10) 式の右辺第 3 項は、各 j に対して $\mathcal{M}_j$ 内のモデルに異なる事前確率を与えていることによる。$\gamma = 0$ の値が増加するにつれてより少ない変数を選択する。

ここで、p が十分に大きく、j が十分に小さい時は、

$$\log \tau(\mathcal{M}_j) = \sum_{k=p-j-1}^{p} \log k - \sum_{k=1}^{j} \log k \approx j \log p$$

と近似できる。そのため、(10) 式の拡張 BIC は、

$$BIC_\gamma(m) \approx n \log \hat{\sigma}^2 + j \log n + 2\gamma j \log p \tag{11}$$

$BIC_\gamma(m)$ の近似値を得ることができる。(11) 式の右辺を拡張 BIC と呼ぶこともある。

それでは、実際に $\gamma$ の値をどのように選択すれば良いのだろうか。一般に、適切な $\gamma$ をデータから決めることは容易ではない。拡張 BIC が実装されいている R のパッケージでは、$\gamma$ のデフォルトの値をあらかじめ定めている (たとえば、$\gamma = 1/2, \quad \gamma = 1/4$。遺伝子データのような $n \ll p$ を満たすような高次元データに対しては,$\gamma = 1$ がうまく機能すると考えられている。なお、際ほど述べたように拡張 BIC は本来最尤法によって推定されたモデルを評価する基準であるため、lasso などの正則化法に対しては、必ずしも変数選択に関する一致性を有するとは限らない。しかしながら、多くの研究者が拡張 BIC をそのまま lasso に適用している。

# 3 lasso 正則化項の拡張

## 3.1 エラスティックネット

前章で紹介した lasso は次の 2 つの問題を有していることが知られている。

1. 相関の高い 2 つの説明変数が目的変数に関係している場合、lasso で推定するとどちらか一方の変数しかモデルに含まない。
2. サンプルサイズより説明変数の数の方が大きい場合、高々サンプルサイズ分の変数までしか選択されない。

これらの問題を同時に解決するために、Zou and Hastie(2005) は**エラスティックネット** (elastic net) とよばれるスパース推定法を提案した。エラスティックネットは、lasso にリッジ項を加えた

$$\frac{1}{2n}\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} \left( \alpha|\beta_j| + \frac{(1-\alpha)\beta_j^2}{2} \right) \tag{12}$$

$\boldsymbol{\beta}$ に関して最小化する。ここで、$\lambda(>0)$ は正則化パラメータ,$\alpha(0 \leq \alpha \leq 1)$ は調整パラメータである。正則化法に注目すると、これは $L_1$ ノルムと $L_2$ ノルムを合わせた形になっているため、エラスティックネットはリッジと lasso の両方の性質を有していることがわかる。

エラスティックネットがどのようにして、lasso が抱える問題を解決しているのか見ていく。1. についてはリッジの**グループ効果**とよばれる性質により解決される。証明は省略する。興味のある読者は Zou and Hastie(2005) を参照されたい。次に、2. について見ていく。まず、次の擬似データを考える。

$$X^* = \begin{pmatrix} X \\ \sqrt{n(1-\alpha)\lambda} I_p \end{pmatrix}, \quad y^* = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0}_p \end{pmatrix}$$

このとき、$\gamma = \alpha\lambda$ と定め、

$$\frac{1}{2n}\|\boldsymbol{y}^* - X^*\boldsymbol{\beta}\|_2^2 + \gamma \sum_{j=1}^{p} |\boldsymbol{\beta}|_j \tag{13}$$

$\boldsymbol{\beta}$ に関して最小化する問題を考える。簡単な計算により (12) 式と (13) 式は同等であることがわかるので、実際は (13) 式を求めれば十分である。擬似データはサンプルサイズが n+p、次元が p の lasso を解いてることになる。したがって、(13) 式は最大 p この変数を選択することが可能となっている。したがって、2. は解決された。

# 4 Appendix.

## 4.1 正則化法で各説明変数の長さを同じにする理由

lasso を例として考える。(9) の式からわかるように右辺の第 2 項は各説明変数に同じ大きさの罰則を課していることがわかる。それゆえ、最小二乗推定量の絶対値 $|\hat{\beta}_j^{LS}|$ が小さいほど $\hat{\beta}_j^{lasso}$ はゼロになりやすい。したがって、各説明変数の長さを調整せずに lasso を実行すると、説明変数の長さが大きい変数に対応する係数が 0 になりやすくなってしまう。この問題に対処するため、あらかじめ説明変数の長さをそろえておく必要がある。

## 4.2 リッジ推定量が正則となる理由

リッジ推定量 (7) にある逆行列が正則になることを示す。計画行列 $X$ に対して実対称行列 $X^TX$ は非負値定符号行列であるので、$X^TX = P\Gamma P^T$ と分解可能である。ここで、$P$ は直行行列であり、$\Gamma = diag(\gamma_1, ..., \gamma_p)$

は $X^TX$ の固有値 $(\gamma_1 \geq \cdots \geq \gamma_p \geq 0)$ を対角成分にもつ対角行列である。もし、$\gamma_p = 0$ ならば $X^TX$ の逆行列は存在せず、$\gamma_p > 0$ ならば、逆行列は存在し、

$$(X^TX)^{-1} = (P\Gamma P^T)^{-1} = P\Gamma^{-1}P^T = Pdiag(1/\gamma_1, ..., 1/\gamma_p)P^T \tag{A.1}$$

と変形できる。

いま、$X^TX$ の逆行列が計算できない状況を考える。たとえば、$X^TX$ の最小固有値が $\gamma_p \to 0$ の場合を考えてみると、(A.1) 式の最後の式より $1/\gamma_p \to \infty$ となり、逆行列が計算できないことがわかる。一方、$X^TX + n\lambda I_p$ の逆行列は

$$\begin{aligned}(X^TX + n\lambda I_p)^{-1} &= (P\Gamma P^T + n\lambda I_p)^{-1} \\ &= (P(\Gamma + n\lambda I_p)P^T)^{-1} \\ &= P(\Gamma + n\lambda I_p)^{-1}P^T \\ &= Pdiag(1/(\lambda_1 + n\lambda), ..., 1/(\lambda_p + n\lambda))P^T\end{aligned}$$

と変形できる。したがって、$\gamma_p \to 0$ の場合でも $1/(\lambda_p + n\lambda)$ は発散しないので、行列 $X^TX + n\lambda I_p$ は正則となる。

# 5 参考文献


- 川野秀一・松井秀俊・廣瀬慧 [2018]『統計学　One Point　6　スパース推定法による統計モデリング』共立出版
- Chen, J., & Chen, Z. [2008]. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*(3), pp.759-771.
- Hoerl, A. E., & Kennard, R. W. [1970]. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), pp.55-67.
- LeBlanc, M., & Tibshirani, R. [1996]. Combining estimates in regression and classification. *Journal of the American Statistical Association*, *91*(436), pp.1641-1650.
- Zou, H., & Hastie, T. [2005]. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), pp.301-320.