Ridge回帰およびLasso 回帰について

担当: 長倉大輔

Ridge回帰

ここでは Ridge 回帰 と Lasso 回帰について説明し、それらに関係した話題について紹介する。Ridge 回帰や Lasso 回帰は説明変数が高次元の時にも線形回帰モデルの係数を推定できることが大きな特徴である(説明変数の数が標本のサイズより大きい時でも推定可能)

以下の文献を参考にした。

Bruce, E. Hansen（2021）Econometrics,
Available at https://www.ssc.wisc.edu/~bhansen/econometrics/
(どうも今はもうダウンロードできなくなっている)

Ridge回帰

■ Ridge回帰推定量

以下の線形回帰モデルを考える。

$$Y_i = \mathbf{X}_i{}' \boldsymbol{\beta} + e_i\ , \quad i = 1,2,\ \ldots,\ n$$

ここで $\mathbf{X}_i = [X_{1i}, \ldots, X_{pi}]'$ は $p \times 1$ 説明変数ベクトル、$\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]'$ は $\mathbf{X}_i$ の係数ベクトル、e_i は誤差項である。$\mathbf{Y} = [Y_1, \ldots, Y_n]'$, $\mathbf{X} = [\ \mathbf{X}_1, \ldots., \mathbf{X}_n]'$,
$\mathbf{e} = [e_1, \ \ldots\ , e_n]'$ とすると、上記のモデルは

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

と書き直せる。ここで $\mathbf{X}$ は full column ランクであると仮定する(つまり $\mathbf{X}$ のp 個の列は1次独立)。

Ridge回帰

この時、$(\mathbf{X}'\mathbf{X})^{-1}$が存在し、$\boldsymbol{\beta}$ の最小二乗推定量は

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

で与えられる。最小二乗法の性質として、ここでは以下のものが重要である。

- 最小二乗推定量は $p \le n$ でないと一意に定まらない。

つまり説明変数の数（次元）が標本のサイズ以下でないと推定できない。これは、もし $p > n$ である場合、$\mathbf{X}'\mathbf{X}$ が特異になり、最小化の1階の条件を満たす $\boldsymbol{\beta}$ が無数に存在するからである。

Ridge回帰

Ridge 回帰を用いれば、$p > n$ の時でも $\boldsymbol{\beta}$ を推定できる。Ridge 回帰推定量は以下のように定義される。

$$\hat{\boldsymbol{\beta}}_{Ridge}(\lambda) = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{Y}$$

ここで $\lambda > 0$ はチューニングパラメーターである。

この推定量は $p > n$ の場合でも存在する。まずそれを確認し、その後この推定量がある**罰則付き残差二乗和の最小化**より導出されることを確認する。また λ の選択の仕方についても議論する。

Ridge回帰

この推定量が存在するということは、推定量の定義中の逆行列が存在するということを意味している。$\lambda > 0$ の時、これが保証されることを確認する($\mathbf{X}'\mathbf{X}$自体は特異行列でも)

まず、$\mathbf{X}'\mathbf{X}$ は常に半正値定符号行列 (この時、固有値は全て0以上)なので、固有値分解(スペクトラル分解)によって、その固有値を対角成分とする行列 $\mathbf{D} = \mathrm{diag}\{d_1,\ldots,d_p\}$を用いて、

$$\mathbf{X}'\mathbf{X} = \mathbf{H}'\mathbf{D}\mathbf{H}$$

と書ける。ここで $\mathbf{H}$ は $\mathbf{D}$ の i 番目の固有値 d_i に対応する固有ベクトル $\mathbf{h}_i$ を並べた行列 $\mathbf{H} = [\mathbf{h}_1,\ldots,\mathbf{h}_p]$ で直交行列である ($\mathbf{H}\mathbf{H}' = \mathbf{H}'\mathbf{H} = \mathbf{I}_p$ を満たす) 。ここで $\mathbf{I}_p$は $p \times p$ の単位行列)である。

Ridge回帰

これを代入すると

$$\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p = \mathbf{H}'\mathbf{D}\mathbf{H} + \lambda \mathbf{H}'\mathbf{H} = \mathbf{H}'(\mathbf{D} + \lambda \mathbf{I}_p)\mathbf{H}$$

を得る。$\mathbf{D} + \lambda \mathbf{I}_p$ の対角成分は $d_i + \lambda > 0,\ i = 1,\ldots, p$ となる。

固有値、固有ベクトルの定義より $d_i\mathbf{h}_i = \mathbf{X}'\mathbf{X}\mathbf{h}_i$ であるので $(d_i + \lambda)\mathbf{h}_i = \mathbf{X}'\mathbf{X}\mathbf{h}_i + \lambda\mathbf{h}_i = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p)\mathbf{h}_i$ となる。よって $d_i + \lambda$ は $(\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p)$ の i 番目の固有値 に対応することがわかる。固有値が全て正であるので$(\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p)$は正則行列であることがわかる。

Ridge回帰

■ Ridge 回帰推定量の罰則付き 残差平方和最小化による導出

次に、Ridge回帰推定量は、ある**罰則付き残差平方和を最小化**することによって得ることができることを確認する。

具体的には、次の罰則付き残差平方和を最小化することによって得られる。

$$\mathrm{SSE}_2(\boldsymbol{\beta}, \lambda) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\, \boldsymbol{\beta}'\boldsymbol{\beta}$$

これは、括弧を開いて展開すると

$$\mathrm{SSE}_2(\boldsymbol{\beta}, \lambda) = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta}'\boldsymbol{\beta}$$

であるので、その最小化のための 1階 の条件は $\partial \mathrm{SSE}_2(\boldsymbol{\beta}, \lambda) / \partial \boldsymbol{\beta} = \mathbf{0}$ である。

Ridge回帰

$\partial \mathrm{SSE}_2(\boldsymbol{\beta}, \lambda) / \partial \boldsymbol{\beta}$ を計算すると

$$\frac{\partial \mathrm{SSE}_2(\boldsymbol{\beta}, \lambda)}{\partial \boldsymbol{\beta}} = \frac{\partial (\mathbf{Y}'\mathbf{Y})}{\partial \boldsymbol{\beta}} - \frac{\partial (\mathbf{Y}'\mathbf{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} - \frac{\partial (\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y})}{\partial \boldsymbol{\beta}} + \frac{\partial (\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} + \lambda \frac{\partial (\boldsymbol{\beta}'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$$

$$= -2\mathbf{Y}'\mathbf{X} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + 2\lambda\boldsymbol{\beta}$$

となるので、これより $\partial \mathrm{SSE}_2(\boldsymbol{\beta}, \lambda) / \partial \boldsymbol{\beta} = \mathbf{0}$ を解いて、

$$\hat{\boldsymbol{\beta}}_{Ridge}(\lambda) = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}'\mathbf{Y}$$

を得る。

Ridge回帰

■ チューニングパラメーター λ の選択

ここではチューニングパラメーターλの選択について説明する。

最もよく用いられる方法は**クロスバリデーション**(cross validation) と呼ばれる方法で、以下の目的関数を最小化するように λ を選ぶ。

$$CV(\lambda) = \sum_{i=1}^{n} \tilde{e}_i(\lambda)^2,$$

ここで

$$\tilde{e}_i(\lambda) = Y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}}_{-i}(\lambda) \quad \text{および} \quad \hat{\boldsymbol{\beta}}_{-i}(\lambda) = \left[\sum_{j\neq i} \mathbf{X}_j\mathbf{X}_j' + \lambda \mathbf{I}_p\right]^{-1}\left[\sum_{j\neq i} \mathbf{X}_j Y_j\right]$$

である。$\hat{\boldsymbol{\beta}}_{-i}(\lambda)$ は ***i* 番目の観測値を除いて推定した Ridge回帰推定量**である。

Ridge回帰

クロスバリデーションでは $CV(\lambda)$を最小化する λ を選択するが、以下の点で注意が必要である。

(1) 最小化は $\lambda = 0$ や $\lambda = \infty$ で起こったり、複数の局所的最小値があるかもしれない。
(2) λ として非常に小さい値も探索することが重要である。

次の公式は $\tilde{e}_i(\lambda)$ の計算を簡単化するのに非常に有用である。

$$\tilde{e}_i(\lambda) = \left[1 - \mathbf{X}_i'(\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p)\mathbf{X}_i\right]^{-1} \hat{e}_i(\lambda)$$

ここで $\hat{e}_i(\lambda) = Y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}}_{Ridge}(\lambda)$ である。

Ridge回帰

クロスバリデーションの他にも、以下の **Mallows 基準を最小化**して λ を選ぶ方法もある。

$$C(\lambda)=\sum_{i=1}^{n}\hat{e}_i(\lambda)^2+2\hat{\sigma}^2\,\mathrm{tr}\left(\left(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}_p\right)^{-1}\mathbf{X}'\mathbf{X}\right)$$

ここで $\hat{\sigma}^2$ は最小二乗法による誤差項の分散の推定値である。

この方法によって選んだ λ を使ったRidge回帰推定量は、誤差項が正規分布に従っている場合は、(モデルフィットの意味で)**最適であるが実行不可能な Ridge 推定量**と漸近的に一致する。

しかしながら、最小二乗法による誤差項の分散の推定値を用いているので $p > n$ の時にはこれは計算できないことに注意が必要である。

Ridge回帰

- その他の注意点

Ridge回帰推定量の重要な注意点として、Ridge回帰推定量は、変数の線形変換に関して不変ではないことに注意が必要である。

Ridge回帰

■ Ridge 回帰推定量の統計的性質

線形回帰モデル: $\mathbf{Y} = \mathbf{X}'\boldsymbol{\beta} + \mathbf{e}$, $\mathbf{Y}=[y_1,\ldots, y_n]^{\mathrm{T}}$, $\mathbf{X}=[\mathbf{x}_1, \ldots, \mathbf{x}_n]$, $\mathbf{x}_i=[x_{1i}, \ldots, x_{pi}]^{\mathrm{T}}$, において $E[\mathbf{e} \mid \mathbf{X}] = \mathbf{0}$ さらに $\mathbf{S} = E[\mathbf{ee}' \mid \mathbf{X}]$ とする($\mathbf{X}_i$ の条件付き不均一分散を許容していることに注意)。Ridge回帰推定量のバイアスと分散共分散行列は以下で与えられる。

$$\mathrm{bias}\left[\hat{\boldsymbol{\beta}}_{Ridge}(\lambda) \mid \mathbf{X}\right] = E\left[\hat{\boldsymbol{\beta}}_{Ridge}(\lambda) \mid \mathbf{X}\right] - \boldsymbol{\beta} = -\lambda(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\boldsymbol{\beta} \ ,$$

$$\mathrm{var}\left[\hat{\boldsymbol{\beta}}_{Ridge}(\lambda) \mid \mathbf{X}\right] = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p)^{-1}(\mathbf{X}'\mathbf{S}\mathbf{X})(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p)^{-1},$$

Ridge回帰

Bias は$E\left[\hat{\boldsymbol{\beta}}_{Ridge}(\lambda) \mid \mathbf{X}\right] = E\left[(\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{Y} \mid \mathbf{X}\right]$

$$
\begin{aligned}
&= E\left[(\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p)^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) \mid \mathbf{X}\right] \\
&= E\left[(\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \mid \mathbf{X}\right] + E\left[(\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{e} \mid \mathbf{X}\right] \\
&= (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}
\end{aligned}
$$

より

$$
\begin{aligned}
E\left[\hat{\boldsymbol{\beta}}_{Ridge}(\lambda) \mid \mathbf{X}\right] - \boldsymbol{\beta} &= (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta} \\
&= (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p)^{-1}[\mathbf{X}'\mathbf{X} - (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p)]\boldsymbol{\beta} \\
&= -\lambda(\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p)^{-1}\boldsymbol{\beta}
\end{aligned}
$$

となる。

Ridge回帰

分散 は

$$\hat{\boldsymbol{\beta}}_{Ridge}(\lambda) - E\left[\hat{\boldsymbol{\beta}}_{Ridge}(\lambda) \mid \mathbf{X}\right] = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{e}$$

より

$$\begin{aligned}\operatorname{var}\left[\hat{\boldsymbol{\beta}}_{Ridge}(\lambda) \mid \mathbf{X}\right] &= E\left[(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{e}\mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p)^{-1} \mid \mathbf{X}\right] \\ &= (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}'E[\mathbf{e}\mathbf{e}' \mid \mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p)^{-1} \\ &= (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{S}\,\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\end{aligned}$$

となる。

Ridge回帰

■ Ridge 回帰推定量の効率性

線形回帰モデル: $\mathbf{Y} = \mathbf{X}'\boldsymbol{\beta} + \mathbf{e}$ は $E[\mathbf{e} \mid \mathbf{X}] = \mathbf{0}$ を満たすとし、さらに $E[\mathbf{e}\mathbf{e}' \mid \mathbf{X}] = \mathbf{D}$ とする。ここで $\mathbf{D} = \mathrm{diag}\{\sigma^2(\mathbf{x}_1), \sigma^2(\mathbf{x}_2), \ldots., \sigma^2(\mathbf{x}_n)\}$, であり $\sigma^2(\mathbf{x})$ は $p \times 1$ ベクトル $\mathbf{x}$ の関数とする。このとき。$\boldsymbol{\beta}$ の任意の推定量 $\hat{\boldsymbol{\beta}}$ の $\mathbf{X}$という条件付きの Mean square error (平均二乗誤差)を

$$MSE\left[\hat{\boldsymbol{\beta}} \mid \mathbf{X}\right] = E\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mid \mathbf{X}\right]$$

とすると、Ridge回帰推定量 $\hat{\boldsymbol{\beta}}_{Ridge}(\lambda)$ と OLS 推定量 $\hat{\boldsymbol{\beta}}_{OLS}$ について、もし $0 < \lambda < 2\underline{\sigma}^2 / (\boldsymbol{\beta}'\boldsymbol{\beta})$ が成り立てば

$$MSE\left[\hat{\boldsymbol{\beta}}_{Ridge}(\lambda) \mid \mathbf{X}\right] < MSE\left[\hat{\boldsymbol{\beta}}_{OLS} \mid \mathbf{X}\right]$$

が成り立つ。ここで $\underline{\sigma}^2 = \min_{x \in A} \sigma^2(x)$ であり、Aは $\mathbf{X}_i$ に共通の取りうる値の範囲(サポート)である。

Ridge回帰

- この結果は $\sigma^2(\mathbf{x}) = \sigma^2$ でも成り立つ。この時、誤差項がガウスマルコフの定理の条件を満たすので OLS推定量は**最良線形不偏推定量**であるが、この結果は Ridge回帰推定量は、そのOLSよりもさらに小さいMSEを持っていることを意味しており非常に興味深い(Ridge 回帰推定量は**不偏推定量ではない**ことに注意)。

- OLS推定量はあくまで不偏推定量の中で最良(分散が一番小さいという意味で)であるということなので、不偏でない推定量まで考えると、OLS推定量よりMSEが小さいという意味でよい推定量が存在しうるが、Ridge推定量はその一つであるということである。

- ただし、$0 < \lambda < 2\underline{\sigma}^2 / (\boldsymbol{\beta}'\boldsymbol{\beta})$ という条件において、この上限は真の値に依存しているため、実際にはわからない。

Lasso回帰

■ Lasso 回帰推定量

Lasso回帰推定量は Least Absolute Shrinkage and Selection Operator の略で、p が大きく、スパース推定の文脈で Ridge 回帰推定量よりよく使われる。

Lasso回帰推定量も罰則付きSSEの最小化の解として得られる。Ridge回帰推定量が

$$\mathrm{SSE}_2(\boldsymbol{\beta}, \lambda) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\, \boldsymbol{\beta}'\boldsymbol{\beta} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2$$

を $\boldsymbol{\beta}$ について最小化したのに対して、Lasso回帰推定量は

$$\mathrm{SSE}_1(\boldsymbol{\beta}, \lambda) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|$$

を $\boldsymbol{\beta}$ について最小化することにより得られる。罰則項が $\boldsymbol{\beta}$ のL_2ノルムから L_1ノルムになっている。$\lambda > 0$ であれば、Lasso回帰推定量は $p > n$ の時でも計算可能である。ただし、Lasso回帰推定量は、特別な場合を除き、Ridge回帰推定量のように**明確な解をもたず**、計算は数値計算の手法を使う必要がある。また、Ridge回帰推定量同様、Lasso推定量も**変数の線形変換について不変ではない**。

Lasso回帰

■ λの選択

Lasso回帰推定量における λ の選択において、Ridge回帰推定量の時に紹介した Leave-one-out CV を使用すると、計算量が膨大になってしまうため、通常あまり使われなず、代わりに K - fold CV がよく用いられる。ここでは K-fold CV を紹介する。

Leave – one –out CV では標本の大きさが n の時はそれぞれ少しずつ違う、大きさ $n\ -1$ の標本（元の標本から i 番目の変数を除いた標本）を用いて $n-1$ 回推定を行う必要があったが、これは **n が大きくなるにつれて、計算量が膨大になってしまうという欠点がある**（特に1回の推定の計算量が多いとき）。これに対して、K – fold CV は 標本を K 個のグループに分けて、それぞれのグループを除いた標本で推定を行い、CVを行う。具体的には次ページのような手順で行う。

Lasso回帰

■ $K-\text{fold}\ CV$ の手順

K- fold CV は以下の手順で行われる。

(1) 観測値をランダムに並べ替える。

(2) 観測値をK個に区切り K 個の標本を作る。この時だいたい $n_k \approx n/K$ となるようにする。ここで n_k は k 番目$(k=1, \ldots, K)$ の標本の大きさである。

(3) $k=1,\ldots, K$ に対して

(a) k 番目の標本を除いた標本を作る(このとき標本の大きさは $n-n_k$)

(b)この標本から $\hat{\boldsymbol{\beta}}_{(-k)}(\lambda)$ を計算する

(c) 推定した $\hat{\boldsymbol{\beta}}_{(-k)}(\lambda)$ を用いて、$i \in I_k$ に対して、 $\tilde{e}_i^{(k)}(\lambda) = \mathbf{Y}_i - \mathbf{X}_i' \boldsymbol{\beta}_{(-k)}$を計算する。ここで I_k は k番目の標本に属する観測値の番号(よって n_k 個の $\tilde{e}_i^{(k)}(\lambda)$ が得られる)。

(d) $CV_k(\lambda) = \sum_{i \in I_k} \tilde{e}_i^{(k)2}(\lambda)$を計算する。

(4) $CV(\lambda) = K^{-1} \sum_{k=1}^{K} CV_k(\lambda)$ 　を計算する。

(5) $CV(\lambda)$を最小化するλを見つける。

Lasso回帰

- *K*–fold CV について

(1) *K* – fold CV において *K*=10 や 20 がよく使われる。

(2) *K* – fold CV の欠点として、最初の ランダムな並べ替えの影響を強く受けることにある。

(3) 異なった並び替えに対して異なった CVが計算される。

(4) これを和らげる方法として *K*- fold CV をM 回行い（M回並び替える）、そこから計算したM個の CV を平均するというやり方も提案されている。

Lasso回帰

■ Lasso回帰推定量の統計的性質

Lasso回帰推定量の統計的性質は、まだ十分にわかっている訳ではないが、いくつかわかっていることを述べる。

Lasso回帰推定量の統計的性質を導く際に、よく用いられる仮定として、Sparsity 条件や近似 sparsity 条件がある。これは、大雑把に言うと 真の $\boldsymbol{\beta}$ の要素の値の多くが 0 であるという仮定である。

いくつかの追加的な仮定の下で以下の結果が得られる。

Lasso回帰

それぞれの説明変数は $n^{-1}\mathbf{X}'\mathbf{X}=\mathbf{1}_p$ となるように基準化されているとする, ここで$\mathbf{1}_p$は要素が全て 1 の $p\times 1$ベクトル)。$e_i \mid \mathbf{X} \sim N(0, \sigma^2(\mathbf{x}))$, であり(ここで $\sigma^2(\mathbf{x})$は前のページに出てきたような関数)、$\sigma^2(\mathbf{x}) \le \bar{\sigma}^2 < \infty$ とする。また$\|\boldsymbol{\beta}\|_0$ を $\boldsymbol{\beta}$ の要素の中で 0 でない係数の数とする。さらに 適当な十分大きな C について

$$\lambda = C\sqrt{n \log p}$$

であるとする。このとき、

$$\left\|\hat{\boldsymbol{\beta}}_{Lasso} - \boldsymbol{\beta}\right\|_1 \le D\|\boldsymbol{\beta}\|_0 \sqrt{\frac{\log p}{n}}$$

が成り立つ。ここで $\left\|\mathbf{X}\right\|_1 = |x_1| + |x_2| + \ldots + |x_p|$で、$D$ は $D < \infty$ の定数である。

よって $\hat{\boldsymbol{\beta}}_{Lasso}$ は一致推定量であることがわかる。