LARS 最適化の方法

池田 思朗

統数研

4 August 2009

- 二次計画法と Lasso
- Lasso と Elastic net
- LARS-Lasso
- 線形計画法と Compressed Sensing

二次計画法

$\boldsymbol{x} = (x_1, \cdots, x_n)^t$ を実数ベクトルとして，この $\boldsymbol{x}$ に関する次のような最適化問題を考えるとき，これを二次計画問題 (Quadratic Programming) と呼ぶ.

基本形式

$$\min_{\boldsymbol{x}} \frac{1}{2}\boldsymbol{x}^t H \boldsymbol{x} + \boldsymbol{f}^t \boldsymbol{x}$$

subject to

$$A\boldsymbol{x} \leq \boldsymbol{b}.$$

ただし，H は n 次の半正定値対称行列，$\boldsymbol{f}$ は n 次の実ベクトルである.

二次計画法

二次計画法を解く数値処理パッケージは沢山存在する．問題や計算機環境に応じて適切なものが選択できる．
これを用いて Lasso の問題を解く．

Lasso の問題 (再掲)

$$\min_{\beta_1,\cdots,\beta_p} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_j |\beta_j|$$

$$\min_{\beta_1,\cdots,\beta_p} \|\boldsymbol{y} - X\boldsymbol{\beta}\|_{\ell_2}^2 + \lambda \|\boldsymbol{\beta}\|_{\ell_1}$$

Lasso 再考

$\boldsymbol{y} = (y_1, \cdots, y_n)^t$, $X = \{x_{ij}\}$, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^t$ として

Lasso の問題 (再掲)

$$
\begin{aligned}
&\min_{\beta_1,\cdots,\beta_p} \|\boldsymbol{y} - X\boldsymbol{\beta}\|_{\ell_2}^2 + \lambda \|\boldsymbol{\beta}\|_{\ell_1} \\
=&\boldsymbol{y}^t\boldsymbol{y} + \min_{\boldsymbol{\beta}} \left(\boldsymbol{\beta}^t X^t X \boldsymbol{\beta} - 2\boldsymbol{y}^t X \boldsymbol{\beta} \right) + \lambda \sum_{j=1}^{p} |\beta_j|
\end{aligned}
$$

Lasso と拡張した問題

ここで $\beta_j = \beta_j^+ - \beta_j^-$ とふたつの成分に分ける．ただし，$\beta_j^+, \beta_j^- \geq 0$.

Lasso の問題

$$\min_{\boldsymbol{\beta}^+,\boldsymbol{\beta}^-} \left[(\boldsymbol{\beta}^+ - \boldsymbol{\beta}^-)^t X^t X (\boldsymbol{\beta}^+ - \boldsymbol{\beta}^-) - 2\boldsymbol{y}^t X(\boldsymbol{\beta}^+ - \boldsymbol{\beta}^-)\big) + \lambda(\boldsymbol{\beta}^+ + \boldsymbol{\beta}^-) \right]$$

subject to

$$\beta_j^+, \beta_j^- \geq 0 \quad j = 1, \cdots, p.$$

$\beta_j = \beta_j^+ - \beta_j^-$ となる正の β_j^+ β_j^- の組は無数にあるが，そのうちで $\beta_j^+ + \beta_j^-$ を最小にするのはどちらかが 0 となる場合に限る．

Lasso を二次計画法の基本問題へ

$\boldsymbol{\gamma} = (\beta_1^+, \cdots, \beta_p^+, \beta_1^-, \cdots, \beta_p^-)^t$ を $2p$ の長さのベクトルとおくと

Lasso の問題

$$\min_{\boldsymbol{\gamma}} \left[\boldsymbol{\gamma}^t \begin{pmatrix} X^tX & -X^tX \\ -X^tX & X^tX \end{pmatrix} \boldsymbol{\gamma} + \begin{pmatrix} -2X^t\boldsymbol{y} + \lambda\mathbf{1}_p \\ 2X^t\boldsymbol{y} + \lambda\mathbf{1}_p \end{pmatrix}^t \boldsymbol{\gamma} \right]$$

subject to

$$-E_{2p}\boldsymbol{\gamma} \leq \mathbf{0}_{2p}.$$

$\mathbf{0}_n$, $\mathbf{1}_n$ はそれぞれ n 次元の 0, 1 を成分とするベクトル，E_n は $n \times n$ の単位行列である．

Lasso を二次計画法の基本問題へ

基本形式: 再掲

$$\min_{\boldsymbol{x}} \frac{1}{2}\boldsymbol{x}^t H\boldsymbol{x} + \boldsymbol{f}^t\boldsymbol{x}$$
$$\text{subject to,} \quad A\boldsymbol{x} \leq \boldsymbol{b},$$

拡張した Lasso の問題: 再掲

$$\min_{\boldsymbol{\gamma}} \left[\boldsymbol{\gamma}^t \begin{pmatrix} X^tX & -X^tX \\ -X^tX & X^tX \end{pmatrix} \boldsymbol{\gamma} + \begin{pmatrix} -2X^t\boldsymbol{y} + \lambda\mathbf{1}_p \\ 2X^t\boldsymbol{y} + \lambda\mathbf{1}_p \end{pmatrix}^t \boldsymbol{\gamma}\right]$$
$$\text{subject to} \quad -E_{2p}\boldsymbol{\gamma} \leq \mathbf{0}_{2p}.$$

$$H=2\begin{pmatrix} X^tX & -X^tX \\ -X^tX & X^tX \end{pmatrix}, \boldsymbol{f}=\begin{pmatrix} -2X^t\boldsymbol{y} + \lambda\mathbf{1}_p \\ 2X^t\boldsymbol{y} + \lambda\mathbf{1}_p \end{pmatrix}, A=-E_{2p}, \boldsymbol{b}=\mathbf{0}_{2p}$$

- 二次計画法と Lasso
- Lasso と Elastic net
- LARS-Lasso
- 線形計画法と Compressed Sensing

Elastic net

Elastic net の問題 (再掲)

$$
\begin{aligned}
&\min_{\beta_1,\cdots,\beta_p} \left\|\boldsymbol{y} - X\boldsymbol{\beta}\right\|_{\ell_2}^2 + \lambda_2\left\|\boldsymbol{\beta}\right\|_{\ell_2}^2 + \lambda_1\left\|\boldsymbol{\beta}\right\|_{\ell_1} \\
=&\boldsymbol{y}^t\boldsymbol{y} + \min_{\boldsymbol{\beta}}\left(\boldsymbol{\beta}^t(X^tX + \lambda_2 E_p)\boldsymbol{\beta} - 2\boldsymbol{y}^tX\boldsymbol{\beta}\right) + \lambda_1\left\|\boldsymbol{\beta}\right\|_{\ell_1}
\end{aligned}
$$

よって，X^tX を $X^tX + \lambda_2 E_p$ によって置き換えることで，Lasso と同様に扱うことができる．

- 二次計画法と Lasso
- Lasso と Elastic net
- LARS-Lasso
- 線形計画法と Compressed Sensing

Lasso の解を求める

ある λ に対して Lasso の解を求める場合は以上のように二次計画法を用いれば良い.
しかし，以下のような場合，各 λ に対して二次計画法を適用するのは手間である.

- λ を様々な値に設定しながら lasso を解きたい
- 大規模な問題を解きたい.

LARS-lasso と呼ばれるアルゴリズムを説明する.

Efron, Hastie, Johnstone, & Tibshirani, (2004). "Least angle regression," *The Annals of Statistics, 32(2)*, 407-499.

Rosset & Zhu, (2007). "Piecewise linear regularized solution paths," *The Annals of Statistics, 35(3)*, 1012–1030.

二次計画法

ある λ に対する Lasso の解を $\hat{\boldsymbol{\beta}}(\lambda)$ と書くことにする.

Lasso の問題 (再掲)

$$\hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - X\boldsymbol{\beta}\|_{\ell_2}^2 + \lambda \|\boldsymbol{\beta}\|_{\ell_1}$$

Lasso の解を求める

Lasso

$$\min_{\boldsymbol{\beta}^+,\boldsymbol{\beta}^-}\left[\|\boldsymbol{y}-X(\boldsymbol{\beta}^+-\boldsymbol{\beta}^-)\|_{\ell_2}^2+\lambda\sum_{j=1}^p(\beta_j^++\beta_j^-)\right]$$

subject to $\beta_j^+,\beta_j^-\geq 0 \quad j=1,\cdots,p.$

ラグランジュ関数

$$\min_{\boldsymbol{\beta}^+,\boldsymbol{\beta}^-}\left[L(\boldsymbol{\beta})+\lambda\sum_{j=1}^p(\beta_j^++\beta_j^-)-\sum_{j=1}^p\lambda_j^+\beta_j^+-\sum_{j=1}^p\lambda_j^-\beta_j^-\right]$$

ただし $L(\boldsymbol{\beta})=\|\boldsymbol{y}-X(\boldsymbol{\beta}^+-\boldsymbol{\beta}^-)\|_{\ell_2}^2.$

Lasso の解を求める

Karush-Kuhn-Tucker 条件 (最適性の必要条件)

$$
\begin{aligned}
\partial_{\beta_j} L(\boldsymbol{\beta}) + \lambda - \lambda_j^+ &= 0 \\
-\partial_{\beta_j} L(\boldsymbol{\beta}) + \lambda - \lambda_j^- &= 0 \\
\lambda_j^+ \beta_j^+ &= 0, \quad \lambda_j^+ \geq 0 \\
\lambda_j^- \beta_j^- &= 0, \quad \lambda_j^- \geq 0 \quad j = 1, \cdots, p.
\end{aligned}
$$

この条件を調べる．ただし

$$
\partial_{\beta_j} L(\boldsymbol{\beta}) = \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = -2(\boldsymbol{y} - X\boldsymbol{\beta})^t \boldsymbol{x}_j
$$

また，$\boldsymbol{x}_j = (x_{1j}, \cdots, x_{nj})^t$ は X の j 番目の列ベクトルである．

Lasso の解を求める

最適値を与える β_j の満たす必要条件はいくつかの場合に分けられる.

(i) $\lambda = 0$

$\partial_{\beta_j} L(\boldsymbol{\beta}) = 0 \qquad (\lambda_j^+, \lambda_j \geq 0$ より $\partial_{\beta_j} L(\boldsymbol{\beta}) = \lambda_j^+ = -\lambda_j^- = 0)$

(ii) $\beta_j^+ > 0, \lambda > 0$

$\partial_{\beta_j} L(\boldsymbol{\beta}) = -\lambda,\ \boldsymbol{\beta}_j^- = 0 \qquad (\lambda_j^+ = 0,\ \lambda_j^- > 0$ より$)$

(iii) $\beta_j^- > 0, \lambda > 0$

$\partial_{\beta_j} L(\boldsymbol{\beta}) = \lambda,\ \boldsymbol{\beta}_j^+ = 0 \qquad (\lambda_j^+ > 0,\ \lambda_j^- = 0$ より$)$

$|\partial_{\beta_j} L(\boldsymbol{\beta})| \leq \lambda,$

LARS-Lasso

λ によって β_j の成分のうち "Active"，すなわち 0 でない成分が変わる．0 でない成分の添字の集合を $\mathcal{A}$ とおく．$\mathcal{A} = \{j : \hat{\beta}_j(\lambda) \neq 0\}$

Active な係数の満たす条件

$$j \in \mathcal{A} \Rightarrow \partial_{\beta_j} L(\hat{\boldsymbol{\beta}}(\lambda)) = -\mathrm{sgn}(\hat{\beta}_j(\lambda))\lambda$$

Active でない係数の満たす条件

$$j \notin \mathcal{A} \Rightarrow |\partial_{\beta_j} L(\hat{\boldsymbol{\beta}}(\lambda))| \leq \lambda$$

LARS-Lasso

$$\partial_{\beta_{\mathcal{A}}}\partial_{\beta_{\mathcal{A}}} L(\hat{\boldsymbol{\beta}}(\lambda)) \frac{\partial \hat{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda)}{\partial \lambda} = -\mathrm{sgn}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda))$$

$$\frac{\partial \hat{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda)}{\partial \lambda} = -\left(\partial_{\beta_{\mathcal{A}}}\partial_{\beta_{\mathcal{A}}} L(\hat{\boldsymbol{\beta}}(\lambda))\right)^{-1} \mathrm{sgn}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda))$$

ここで添字の $\mathcal{A}$ は $j \in \mathcal{A}$ となる添字 j をとりだしたもの．また，$\partial_{\beta_{\mathcal{A}}}\partial_{\beta_{\mathcal{A}}} L(\hat{\boldsymbol{\beta}}(\lambda)) = 2X_{\mathcal{A}}^t X_{\mathcal{A}}$．したがって

$$\frac{\partial \hat{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda)}{\partial \lambda} = -\frac{1}{2}\left(X_{\mathcal{A}}^t X_{\mathcal{A}}\right)^{-1} \mathrm{sgn}\ (\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda))$$

LARS-Lasso

Algorithm

1. $\boldsymbol{\beta} = 0$, $\mathcal{A} = \{\arg\max_j |\boldsymbol{y}^t \boldsymbol{x}_j|\}$, $\boldsymbol{\gamma}_{\mathcal{A}} = -\mathrm{sgn}\ (-\boldsymbol{y}^t \boldsymbol{x}_{\mathcal{A}})$, $\boldsymbol{\gamma}_{\mathcal{A}^c} = 0$
2. $\max |\partial_{\beta_j} L(\boldsymbol{\beta})| > 0$ ならば以下の手続きを行なう.
   - i $d_1 = \min\{d > 0 : |\partial_{\beta_j} L(\boldsymbol{\beta} + d\boldsymbol{\gamma})| = |\partial_{\beta_{\mathcal{A}}} L(\boldsymbol{\beta} + d\boldsymbol{\gamma})|, j \notin \mathcal{A}\}$
     $d_2 = \min\{d > 0 : (\boldsymbol{\beta} + d\boldsymbol{\gamma})_j = 0, j \in \mathcal{A}\}$
     $d = \min(d_1, d_2)$
   - ii $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} + d\boldsymbol{\gamma}$
     $d = d_1$ ならば d_1 として選ばれた j を $\mathcal{A}$ に加える.
     $d = d_2$ ならば d_2 として選ばれた j を $\mathcal{A}$ から取り除く.
   - iii $\boldsymbol{\gamma}_{\mathcal{A}} = -(X_{\mathcal{A}}^t X_{\mathcal{A}})^{-1} \mathrm{sgn}(\partial_{\beta_{\mathcal{A}}} L(\boldsymbol{\beta}))$, $\boldsymbol{\gamma}_{\mathcal{A}^c} = \mathbf{0}$ と更新する.
   - iv ステップ 2 へ

LARS-Lasso

1. $\boldsymbol{\beta} = 0$, $\mathcal{A} = \{\arg\max_j |\boldsymbol{y}^t \boldsymbol{x}_j|\}$, $\boldsymbol{\gamma}_{\mathcal{A}} = -\mathrm{sgn}\,(-\boldsymbol{y}^t \boldsymbol{x}_{\mathcal{A}})$, $\boldsymbol{\gamma}_{\mathcal{A}^c} = 0$

   初期値の設定: λ が無限に大きいとき，全てのパラメータは 0 有限の大きい値のときには，β_j として $\boldsymbol{y}^t \boldsymbol{x}_j$ を最大にする j を選ぶのが良い

2. $\max |\partial_{\beta_j} L(\boldsymbol{\beta})| > 0$ ならば以下の手続きを行なう．

   全ての $\partial_{\beta_j} L(\boldsymbol{\beta}) = 0$ ならば制約のない最適化問題となっていることを示す．このときにはさらに更新する必要はない．

LARS-Lasso

i $d_1 = \min\{d > 0 : |\partial_{\beta_j} L(\boldsymbol{\beta} + d\boldsymbol{\gamma})| = |\partial_{\beta_{\mathcal{A}}} L(\boldsymbol{\beta} + d\boldsymbol{\gamma})|, j \notin \mathcal{A}\}$
$d_2 = \min\{d > 0 : (\boldsymbol{\beta} + d\boldsymbol{\gamma})_j = 0, j \in \mathcal{A}\}$
$d = \min(d_1, d_2)$

d_1 は Active でない係数の満たす条件 $j \notin \mathcal{A} \Rightarrow \left|\partial_{\beta_j} L(\hat{\boldsymbol{\beta}}(\lambda))\right| \le \lambda$ が破綻し，ある Active でない係数が Active となる場合がいつかを調べる．
d_2 は Active な係数の満たす条件 $j \in \mathcal{A} \Rightarrow \partial_{\beta_j} L(\hat{\boldsymbol{\beta}}(\lambda)) = -\mathrm{sgn}(\hat{\beta}_j(\lambda))\lambda$ が係数が 0 となることで破綻する場合がいつかを調べる．

LARS-Lasso

ii $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} + d\boldsymbol{\gamma}$
$d = d_1$ ならば d_1 として選ばれた j を $\mathcal{A}$ に加える.
$d = d_2$ ならば d_2 として選ばれた j を $\mathcal{A}$ から取り除く.

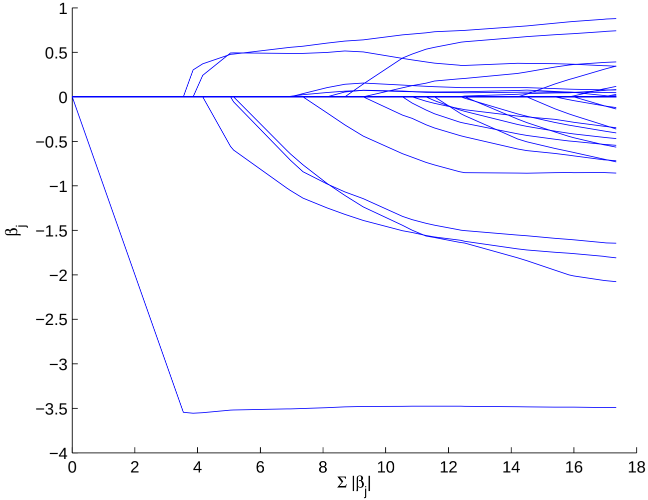
d_1 のほうが d_2 より小さければ，Active でなかった係数を Active とする
d_2 のほうが d_1 より小さければ，Active だった係数を Active な係数の集合から取り除く.

iii $\boldsymbol{\gamma}_{\mathcal{A}} = -(X^t_{\mathcal{A}} X_{\mathcal{A}})^{-1} \mathrm{sgn}(\partial_{\beta_{\mathcal{A}}} L(\boldsymbol{\beta}))$，$\boldsymbol{\gamma}_{\mathcal{A}^c} = \mathbf{0}$ と更新する.

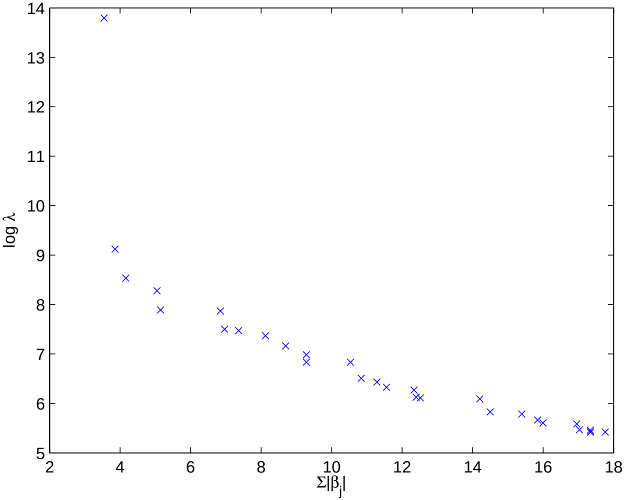
新たなパラメータの変化の方向を求める.

$$\frac{\partial \hat{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda)}{\partial \lambda} = -\left(X^t_{\mathcal{A}} X_{\mathcal{A}}\right)^{-1} \mathrm{sgn}\ (\partial_{\boldsymbol{\beta}_{\mathcal{A}}} L(\boldsymbol{\beta}))$$

パラメータの変化



λ と s の関係



- 二次計画法と Lasso
- Lasso と Elastic net
- LARS-Lasso
- 線形計画法と Compressed Sensing

線形計画法

$\boldsymbol{x} = (x_1, \cdots, x_n)^t$ を実数ベクトルとして，この $\boldsymbol{x}$ に関する次のような最適化問題を考えるとき，これを線形計画問題 (Linear Programming) と呼ぶ.

基本形式

$$\min_{\boldsymbol{x}} \boldsymbol{c}^t \boldsymbol{x}$$

subject to

$$A\boldsymbol{x} \leq \boldsymbol{b}.$$

ただし，$\boldsymbol{c}$ は n 次の実ベクトルである.

線形計画法

線形計画法を解く数値処理パッケージも沢山存在する．問題や計算機環境に応じて適切なものが選択できる．
これを用いて Compressed Sensing の問題を解く．

Compressed Sensing の問題 (再掲)

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{x}\|_{\ell_1}$$

subject to

$$A\boldsymbol{x} = \boldsymbol{y}$$

線形計画法

Lasso のときと同様 $x_j = x_j^+ - x_j^-$, $x_j^+ \geq 0$, $x_j^- \geq 0$ とおいて,

Compressed Sensing

$$\min_{\boldsymbol{x}^+, \boldsymbol{x}^-} (\boldsymbol{x}^+ + \boldsymbol{x}^-)$$

subject to

$$A(\boldsymbol{x}^+ - \boldsymbol{x}^-) = \boldsymbol{y}$$

最適な解はそれぞれの j に対して x_j^+ または x_j^- の少なくともひとつは 0 となる.

線形計画法

$\boldsymbol{\gamma} = (x_1^+, \cdots, x_p^+, x_1^-, \cdots, x_p^-)^t$ とおいて

Compressed Sensing

$$\min_{\boldsymbol{\gamma}} \mathbf{1}_{2p}^t \boldsymbol{\gamma} \quad \text{subject to} \quad \begin{pmatrix} A & -A \\ -A & A \end{pmatrix} \boldsymbol{\gamma} \leq \begin{pmatrix} \boldsymbol{y} \\ -\boldsymbol{y} \end{pmatrix}, \quad -E_{2p}\boldsymbol{\gamma} \leq \mathbf{0}_{2p}$$

これは線形計画法の問題になっている．

基本形式

$$\min_{\boldsymbol{x}} \boldsymbol{c}^t \boldsymbol{x} \quad \text{subject to} \quad A\boldsymbol{x} \leq \boldsymbol{b}.$$