Initialization of Constrained LASSO path

안승환

2019년 9월

[1] Department of Statistics
University of Seoul

Goal

Constrained lasso problem with only equality constraints:

$$\begin{aligned} &\text{minimize} \quad && L(\boldsymbol{\beta}) + \rho||\boldsymbol{\beta}||_1 \\ &\text{subject to} \quad && \mathbf{A}\boldsymbol{\beta} = 0 \end{aligned} \tag{1}$$

(We could think of $L(\boldsymbol{\beta})$ as $\frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2$.)
Since we perform path following in the decreasing direction, an initializing value for the parameter ρ is needed.
$(\mathbf{y} \in \mathbb{R}^n, \mathbf{X} \in \mathbb{R}^{n \times p}, \boldsymbol{\beta} \in \mathbb{R}^p, \mathbf{A} \in \mathbb{R}^{m \times p})$

Goal

As $\rho \to \infty$, the solution $\boldsymbol{\beta}$ to the original problem is given by

$$\begin{aligned} &\text{minimize} && ||\boldsymbol{\beta}||_1 \\ &\text{subject to} && \mathbf{A}\boldsymbol{\beta} = 0 \end{aligned} \tag{2}$$

And obviously, the solution $\hat{\boldsymbol{\beta}}$ for the above problem is 0_p.

KKT condition

The stationarity condition of KKT conditions is as follows:

$$\nabla L(\boldsymbol{\beta}) + \rho sign(\boldsymbol{\beta}) + \mathbf{A}^T \lambda = 0_p$$

, where $\lambda \in \mathbb{R}^m$ is lagrangian multiplier and $sign(\boldsymbol{\beta})$ means the subgradient of $||\boldsymbol{\beta}||_1$. And $|sign(\boldsymbol{\beta})| \leq 1_p$, we can transform above condition as follows:

$$|\nabla L(\boldsymbol{\beta}) + \mathbf{A}^T \lambda| \leq \rho 1_p$$

Lemma

For fixed ρ, $\boldsymbol{\beta}$, let $\mathcal{E}_\rho(\boldsymbol{\beta}) = \{\lambda \in \mathbb{R}^m : |\nabla L(\boldsymbol{\beta}) + \mathbf{A}^T\lambda| \leq \rho 1_p\}$, and let $\rho_{max} = inf\{\rho \in \mathbb{R} : \mathcal{E}_\rho(\boldsymbol{\beta}) \neq \varnothing\}$. Then for $\rho < \rho_{max}$, $\boldsymbol{\beta} = 0_p$ is not solution of (1).

Corollary

The minimizer of (1) for a ρ is $\boldsymbol{\beta} = 0_p$ if and only if $\rho \geq \rho_{max}$, where ρ_{max} is the solution of (3). And also, we can get the solution λ_{max} corresponding to ρ_{max} by (3).

$$\begin{aligned} \text{minimize} \quad & \rho \\ \text{subject to} \quad & z = \mathbf{A}^T \lambda \\ & z \leq -X^T y + \rho 1 \\ & z \geq -X^T y - \rho 1 \\ & \rho \geq 0 \end{aligned} \tag{3}$$

Active Set

So, we can initialize active set $\mathcal{A}$ as follows:

$$\mathcal{A} = \{j : |\nabla L(\boldsymbol{\beta})_j + a_j^T \lambda_{max}| = \rho_{max}\}$$

where $\mathbf{A} = [a_1, \cdots a_p]$.

Because If ρ is decreased very little as $\rho < \rho_{max}$, $\hat{\beta}_j = 0$ cannot be the coefficient of solution (1) for predictor x_j. So, predictor x_j must be activated.

Uniqueness of λ

λ_{max} is the unique solution to (3) if the solution for $\mathbf{A}_{\mathcal{A}}^T\tilde{\lambda} = 0$ is only $\tilde{\lambda} = 0$. ($\Leftrightarrow$ $\mathbf{A}_{\mathcal{A}}^T$ is full column).

Because we can formulate an equation for predictors which are set on boundaries of the stationarity condition(which are also included in active set $\mathcal{A}$) as follows:

$$|\nabla L(\boldsymbol{\beta})_{\mathcal{A}} + \mathbf{A}_{\mathcal{A}}^T(\lambda_{max} + \tilde{\lambda})| = \rho_{max} 1_{\mathcal{A}}$$

What if...?

Q) What happens if do NOT activate all the violated predictors($\mathcal{A}$)?

Let $L(\boldsymbol{\beta}) = \frac{1}{2}||\mathbf{y} - \mathbf{X}\beta||_2^2$. Given $\rho_{max}, \lambda_{max}$, we want $\mathcal{B} \subseteq \mathcal{A}$, $\Delta\rho, \frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}}, \frac{d}{d\rho}\lambda_{\mathcal{B}}$ such that satisfy following conditions(stationarity condition from KKT conditions, equality constraint):

$$\begin{aligned}
-\mathbf{X}_{:\mathcal{B}}^T(\mathbf{y} - \mathbf{X}_{:\mathcal{B}}(\boldsymbol{\beta}_{\mathcal{B}}^{(0)} + \Delta\rho\frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}})) &\quad + \\
(\rho_{max} - \Delta\rho)sign(\boldsymbol{\beta}_{\mathcal{B}}^{(0)} + \Delta\rho\frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}}) + \mathbf{A}_{:\mathcal{B}}^T(\lambda_{max} + \Delta\rho\frac{d}{d\rho}\lambda) &= 0 \\
|-\mathbf{X}_{:\mathcal{B}^c}^T(\mathbf{y} - \mathbf{X}_{:\mathcal{B}}(\boldsymbol{\beta}_{\mathcal{B}}^{(0)} + \Delta\rho\frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}})) + \mathbf{A}_{:\mathcal{B}^c}^T(\lambda_{max} + \Delta\rho\frac{d}{d\rho}\lambda)| & \\
&\leq (\rho_{max} - \Delta\rho) \\
\mathbf{A}_{:\mathcal{B}}(\boldsymbol{\beta}_{\mathcal{B}}^{(0)} + \Delta\rho\frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}}) &= 0
\end{aligned}$$

What if...?

ρ is in decreasing direction, so $\Delta\rho > 0$. And the moving direction of $\boldsymbol{\beta}$ is must be maintained. So,

$$sign(\boldsymbol{\beta}_{\mathcal{B}}^{(0)} + \Delta\rho \frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}}) = sign(\boldsymbol{\beta}^{(0)})$$

. From corollary, we have following:

$$\begin{aligned} -\mathbf{X}_{:\mathcal{B}}^{T}(\mathbf{y} - \mathbf{X}_{:\mathcal{B}}\boldsymbol{\beta}_{\mathcal{B}}^{(0)}) + \rho_{max} sign(\boldsymbol{\beta}_{\mathcal{B}}^{(0)}) + \mathbf{A}_{:\mathcal{B}}^{T}(\lambda_{max}) &= 0 \\ |-\mathbf{X}_{:\mathcal{B}^{\mathcal{C}}}^{T}(\mathbf{y} - \mathbf{X}_{:\mathcal{B}}\boldsymbol{\beta}_{\mathcal{B}}^{(0)}) + \mathbf{A}_{:\mathcal{B}^{\mathcal{C}}}^{T}\lambda_{max}| &\leq \rho_{max} 1_{|\mathcal{B}^{\mathcal{C}}|} \\ \mathbf{A}_{:\mathcal{B}}\boldsymbol{\beta}_{\mathcal{B}}^{(0)} &= 0 \end{aligned}$$

What if...?

Finally, we get

$$\begin{array}{rcl} \mathbf{X}_{:\mathcal{B}}^{T}\mathbf{X}_{:\mathcal{B}}\Delta\rho\frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}} - \Delta\rho sign(\beta_{\mathcal{B}}^{(0)}) + \mathbf{A}_{:\mathcal{B}}^{T}\Delta\rho\frac{d}{d\rho}\lambda & = & 0 \\ \mathbf{A}_{:\mathcal{B}}\Delta\rho\frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}} & = & 0 \\ |\mathbf{X}_{:\mathcal{B}^{c}}^{T}\mathbf{X}_{:\mathcal{B}}\Delta\rho\frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}} + \mathbf{A}_{:\mathcal{B}^{c}}^{T}\Delta\rho\frac{d}{d\rho}\lambda| & \leq & \Delta\rho 1_{|\mathcal{B}^{c}|} \end{array}$$

What if...?

Let's focus on first two equations:

$$\begin{bmatrix} \mathbf{X}_{:\mathcal{B}}^T \mathbf{X}_{:\mathcal{B}} & \mathbf{A}_{:\mathcal{B}}^T \\ \mathbf{A}_{:\mathcal{B}} & 0 \end{bmatrix} \begin{bmatrix} \frac{d}{d\rho} \boldsymbol{\beta}_{\mathcal{B}} \\ \frac{d}{d\rho} \lambda \end{bmatrix} \quad = \quad \begin{bmatrix} sign(\boldsymbol{\beta}_{\mathcal{B}}^{(0)}) \\ 0 \end{bmatrix}$$

And,

$$\begin{bmatrix} \frac{d}{d\rho} \boldsymbol{\beta}_{\mathcal{B}} \\ \frac{d}{d\rho} \lambda \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{:\mathcal{B}}^T \mathbf{X}_{:\mathcal{B}} & \mathbf{A}_{:\mathcal{B}}^T \\ \mathbf{A}_{:\mathcal{B}} & 0 \end{bmatrix}^{-1} \begin{bmatrix} sign(\boldsymbol{\beta}_{\mathcal{B}}^{(0)}) \\ 0 \end{bmatrix}$$

Answer

$\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}}$ is invertible.

$$\begin{bmatrix} \frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}} \\ \frac{d}{d\rho}\lambda \end{bmatrix} = \begin{bmatrix} (\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1} - (\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1}\mathbf{A}_{:\mathcal{B}}^T Z^{-1}\mathbf{A}_{:\mathcal{B}}(\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1} & (\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1}\mathbf{A}_{:\mathcal{B}}^T Z^{-1} \\ Z^{-1}\mathbf{A}_{:\mathcal{B}}(\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1} & -Z^{-1} \end{bmatrix} \begin{bmatrix} sign(\boldsymbol{\beta}_{\mathcal{B}}^{(0)}) \\ 0 \end{bmatrix}$$

where $Z = \mathbf{A}_{:\mathcal{B}}(\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1}\mathbf{A}_{:\mathcal{B}}^T$.

Therefore,

$$\frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}} = \left((\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1} - (\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1}\mathbf{A}_{:\mathcal{B}}^T Z^{-1}\mathbf{A}_{:\mathcal{B}}(\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1}\right) sign(\boldsymbol{\beta}_{\mathcal{B}}^{(0)})$$

Answer

If $\frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}} \approx 0$, for new active set $\mathcal{B}$, there is no direction to move $\boldsymbol{\beta}_{\mathcal{B}}$ that satisfies KKT conditions.
If not, we check $\frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}}$ and $\frac{d}{d\rho}\lambda$ for following condition:

$$|\mathbf{X}_{:\mathcal{B}^c}^T \mathbf{X}_{:\mathcal{B}} \frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}} + \mathbf{A}_{:\mathcal{B}^c}^T \frac{d}{d\rho}\lambda| \leq 1_{|\mathcal{B}^c|}$$

(Trivial: If $\mathbf{A}_{:\mathcal{B}}$ is invertible,

$$\begin{bmatrix} \mathbf{X}_{:\mathcal{B}}^T \mathbf{X}_{:\mathcal{B}} & \mathbf{A}_{:\mathcal{B}}^T \\ \mathbf{A}_{:\mathcal{B}} & 0 \end{bmatrix}^{-1} = \begin{bmatrix} 0 & \mathbf{A}_{:\mathcal{B}}^{-1} \\ (\mathbf{A}_{:\mathcal{B}}^T)^{-1} & -(\mathbf{A}_{:\mathcal{B}}^T)^{-1}\mathbf{X}_{:\mathcal{B}}^T \mathbf{X}_{:\mathcal{B}} \mathbf{A}_{:\mathcal{B}}^{-1} \end{bmatrix}.$$

So, $\frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}} = 0$ and there is no direction to move.)

Furthermore

From

$$\frac{d}{d\rho}\beta_{\mathcal{B}} = \left((\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1} - (\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1}\mathbf{A}_{:\mathcal{B}}^T Z^{-1}\mathbf{A}_{:\mathcal{B}}(\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1}\right) sign(\beta_{\mathcal{B}}^{(0)})$$

, let

$$\mathbf{W} = \mathbf{A}_{:\mathcal{B}}^T M^{-1}\mathbf{A}_{:\mathcal{B}}(\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1} = \mathbf{A}_{:\mathcal{B}}^T(\mathbf{A}_{:\mathcal{B}}(\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1}\mathbf{A}_{:\mathcal{B}}^T)^{-1}\mathbf{A}_{:\mathcal{B}}(\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1}$$

. Also, let m is the number of constraints.

Furthermore

Where $\mathcal{B} = \mathcal{A}$

- $\mathbf{A}_{:\mathcal{B}}(\mathbf{X}_{:\mathcal{B}}^T \mathbf{X}_{:\mathcal{B}})^{-1}\mathbf{A}_{:\mathcal{B}}^T$ **is invertible.**
  So, $\mathbf{A}_{:\mathcal{B}}$ is right-invertible, and $\mathbf{A}_{:\mathcal{B}}^T$ is left-invertible. And let $\mathbf{A}_{:\mathcal{B}}\mathbf{A}_{:\mathcal{B}}^{-1} = \mathbf{I}$, and $\mathbf{A}_{:\mathcal{B}}^{-T}\mathbf{A}_{:\mathcal{B}}^T = \mathbf{I}$

$$\begin{aligned} \mathbf{Z} &= \mathbf{A}_{:\mathcal{B}}^T \mathbf{A}_{:\mathcal{B}}^{-T}(\mathbf{X}_{:\mathcal{B}}^T \mathbf{X}_{:\mathcal{B}})\mathbf{A}_{:\mathcal{B}}^{-1}\mathbf{A}_{:\mathcal{B}}(\mathbf{X}_{:\mathcal{B}}^T \mathbf{X}_{:\mathcal{B}})^{-1} \\ &\neq \mathbf{I} \end{aligned}$$

The order of matrix product is changed, so $\frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}} \neq 0$.

Furthermore

Where $|\mathcal{B}| = m$

- **If $|\mathcal{B}| = m$, $\mathbf{A}_{:\mathcal{B}}$ is invertible**

$$\begin{aligned}\mathbf{Z} &= \mathbf{A}_{:\mathcal{B}}^T \mathbf{A}_{:\mathcal{B}}^{-T} (\mathbf{X}_{:\mathcal{B}}^T \mathbf{X}_{:\mathcal{B}}) \mathbf{A}_{:\mathcal{B}}^{-1} \mathbf{A}_{:\mathcal{B}} (\mathbf{X}_{:\mathcal{B}}^T \mathbf{X}_{:\mathcal{B}})^{-1} \\ &= \mathbf{I}\end{aligned}$$

Therefore, $\frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}} = 0$.

Furthermore

Where $|\mathcal{B}| < m$

- **If $\mathbf{A}_{:\mathcal{B}}(\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1}\mathbf{A}_{:\mathcal{B}}^T$ is NOT invertible**
  (**Assumption**: columns of $\mathbf{A}_{:\mathcal{B}}$ are linearly independent.)

$$\begin{aligned}\mathbf{Z} &= \mathbf{A}_{:\mathcal{B}}^T Ginv(\mathbf{A}_{:\mathcal{B}}(\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1}\mathbf{A}_{:\mathcal{B}}^T)\mathbf{A}_{:\mathcal{B}}(\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1} \\ &= \mathbf{I}\end{aligned}$$

  where $Ginv(\mathbf{A})$ means generalized inverse of matrix $\mathbf{A}$.
  Therefore, $\frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}} = 0$.

Future work...

1. For $H \in \mathbb{R}^{m \times n}$, HH^T is not invertible, where $m > n$

proof)

The columns of H^T are linearly dependent. So, there exists $x \neq 0$ such that $H^T x = 0$. $HH^T x = 0$ and this means 0 is an eigenvalue of HH^T. Therefore, $|HH^T| = 0$ and HH^T is nonsingular. (The determinant of matrix is the product of eigenvalues of the matrix).

Future work...

2. Why $|\mathcal{A}| = m + 1$?

For setting initial active set, we solve following problem and find predictors which set on the boundary of inequalities. And these predictors are chose for initial active set.

$$\begin{aligned} \text{minimize} \quad & \rho \\ \text{subject to} \quad & z = \mathbf{A}^T \lambda \\ & z \leq -X^T y + \rho 1 \\ & z \geq -X^T y - \rho 1 \\ & \rho \geq 0 \end{aligned} \tag{4}$$

(Application: If you want to activate only q predictors at the initialization step, then you should set $q - 1$ constraints (it means $A \in \mathbb{R}^{q-1 \times p}$).

Future work...

And we can change problem (4) to the problem having same purpose(get initial active set) as following:
Find λ, minimum ρ such that

$$\begin{bmatrix} \mathbf{A}_{\mathcal{A}}^{T} & \pm 1_{|\mathcal{A}|} \end{bmatrix} \begin{bmatrix} \lambda \\ \rho \end{bmatrix} = \begin{bmatrix} -\mathbf{X}_{\mathcal{A}}^{T} y \end{bmatrix}$$

And predictors of $\mathcal{A}^{C}$ are set between inequalities. The unknown variable $\begin{bmatrix} \lambda \\ \rho \end{bmatrix}$ is $m+1$ dimension. So, when $|\mathcal{A}| = m+1$, this linear programming has unique solution.