

Initialization of Constrained LASSO path

안승환

2019년 9월

¹Department of Statistics
University of Seoul

Problem

Constrained lasso problem with only equality constraints:

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

(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.

$(\mathbf{y} \in \mathbb{R}^n, \mathbf{X} \in \mathbb{R}^{n \times p}, \boldsymbol{\beta} \in \mathbb{R}^p, \mathbf{A} \in \mathbb{R}^{m \times p})$

Problem

As $\rho \rightarrow \infty$, the solution β to the original problem is given by

$$\begin{array}{ll} \text{minimize} & ||\beta||_1 \\ \text{subject to} & \mathbf{A}\beta = 0 \end{array} \quad (2)$$

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

The stationarity condition of KKT conditions is as follows:

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

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

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

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

Corollary

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

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

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

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

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

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

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 compose active set \mathcal{A}) as follows:

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

Uniqueness of \mathcal{A}

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

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

$$\begin{aligned} & -\mathbf{X}_{:\mathcal{B}}^T(\mathbf{y} - \mathbf{X}_{:\mathcal{B}}(\beta_{\mathcal{B}}^{(0)} - \Delta\rho \frac{d}{d\rho}\beta_{\mathcal{B}})) \quad + \\ & (\rho_{max} - \Delta\rho) \text{sign}(\beta_{\mathcal{B}}^{(0)} - \Delta\rho \frac{d}{d\rho}\beta_{\mathcal{B}}) + \mathbf{A}_{:\mathcal{B}}^T(\lambda_{max} - \Delta\rho \frac{d}{d\rho}\lambda) \quad = \quad 0 \\ & |\mathbf{X}_{:\mathcal{B}^c}^T(\mathbf{y} - \mathbf{X}_{:\mathcal{B}}(\beta_{\mathcal{B}}^{(0)} - \Delta\rho \frac{d}{d\rho}\beta_{\mathcal{B}})) - \mathbf{A}_{:\mathcal{B}^c}^T(\lambda_{max} - \Delta\rho \frac{d}{d\rho}\lambda)| \quad \leq \quad (\rho_{max} - \Delta\rho)1 \\ & \mathbf{A}_{:\mathcal{B}}(\beta_{\mathcal{B}}^{(0)} - \Delta\rho \frac{d}{d\rho}\beta_{\mathcal{B}}) \quad = \quad 0 \end{aligned}$$

Uniqueness of \mathcal{A}

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

$$\text{sign}(\beta_B^{(0)} - \Delta\rho \frac{d}{d\rho} \beta_B) = \text{sign}(\beta^{(0)})$$

. From corollary, we have following:

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

Uniqueness of \mathcal{A}

Finally, we get

$$\mathbf{X}_{:B}^T \mathbf{X}_{:B} \Delta \rho \frac{d}{d\rho} \beta_B - \Delta \rho \text{sign}(\beta_B^{(0)}) + \mathbf{A}_{:B}^T \Delta \rho \frac{d}{d\rho} \lambda = 0$$

$$\mathbf{A}_{:B} \Delta \rho \frac{d}{d\rho} \beta_B = 0$$

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

Uniqueness of \mathcal{A}

Let's focus on first two equations:

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

And,

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

(Assumption 1) $\begin{bmatrix} \mathbf{X}_{:\mathcal{B}}^T \mathbf{X}_{:\mathcal{B}} & \mathbf{A}_{:\mathcal{B}}^T \\ \mathbf{A}_{:\mathcal{B}} & 0 \end{bmatrix}$ is invertible if rows of $\mathbf{A}_{:\mathcal{B}}$ are linearly independent.

Uniqueness of \mathcal{A}

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

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

where $Z = \mathbf{A}_{:\mathcal{B}} (\mathbf{X}_{:\mathcal{B}}^T \mathbf{X}_{:\mathcal{B}})^{-1} \mathbf{A}_{:\mathcal{B}}^T$ and here, inverse of Z means not only its inverse but also its generalized inverse.

Therefore,

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

Uniqueness of \mathcal{A}

If $\frac{d}{d\rho}\beta_{\mathcal{B}} \approx 0$, for new active set \mathcal{B} , there is no direction to move $\beta_{\mathcal{B}}$ that satisfies KKT conditions.

If not, we check $\frac{d}{d\rho}\beta_{\mathcal{B}}$ and $\frac{d}{d\rho}\lambda$ for following condition:

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

(And it is easily shown that for all \mathcal{B} , above inequality is satisfied).

(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, $\frac{d}{d\rho}\beta_{\mathcal{B}} = 0$ and there is no direction to move.)

Uniqueness of \mathcal{A}

From

$$\begin{aligned}\frac{d}{d\rho}\beta_{\mathcal{B}} &= ((\mathbf{X}_{:\mathcal{B}}^T \mathbf{X}_{:\mathcal{B}})^{-1} - (\mathbf{X}_{:\mathcal{B}}^T \mathbf{X}_{:\mathcal{B}})^{-1} \mathbf{A}_{:\mathcal{B}}^T Z^{-1} \mathbf{A}_{:\mathcal{B}} (\mathbf{X}_{:\mathcal{B}}^T \mathbf{X}_{:\mathcal{B}})^{-1}) \text{sign}(\beta_{\mathcal{B}}^{(0)}) \\ &= ((\mathbf{X}_{:\mathcal{B}}^T \mathbf{X}_{:\mathcal{B}})^{-1} (\mathbf{I} - \mathbf{W})) \text{sign}(\beta_{\mathcal{B}}^{(0)})\end{aligned}$$

,

where

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

. Also, let m is the number of constraints.

Uniqueness of \mathcal{A}

Where $\mathcal{B} = \mathcal{A}$ ($|\mathcal{B}| = m + 1$)

- $\mathbf{A}_{:\mathcal{B}}(\mathbf{X}_{:\mathcal{B}}^T \mathbf{X}_{:\mathcal{B}})^{-1} \mathbf{A}_{:\mathcal{B}}^T$ 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 $\mathbf{A}_{:\mathcal{B}}^{-T} \mathbf{A}_{:\mathcal{B}}^T = \mathbf{I}$ with their right and left inverses.

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

The order of left and right inverse matrix product is changed, so $\frac{d}{d\rho} \beta_{\mathcal{B}} \neq 0$.

Uniqueness of \mathcal{A}

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

- If $|\mathcal{B}| = m$, $\mathbf{A}_{:\mathcal{B}}$ is invertible (by Assumption 1)

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

Therefore, $\frac{d}{d\rho} \beta_{\mathcal{B}} = 0$.

Uniqueness of \mathcal{A}

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 (See Appendix 1)

So, we use generalized inverse.

(**Assumption 2**: columns of $\mathbf{A}_{:\mathcal{B}}$ are linearly independent.)

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

where $(\mathbf{A})^{-}$ means generalized inverse of matrix \mathbf{A} .

Therefore, $\frac{d}{d\rho} \beta_{\mathcal{B}} = 0$.

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).

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.

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

Appendix 2

And we can change problem (4) to the problem having same purpose (get initial active set) as following:

Find λ , and minimum ρ such that

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

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(?).

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