Initialization of Equality Constrained LASSO

안승환

2019

Contents

1 Problem

Constrained lasso problem with only equality constraints:

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

(In this paper, we would 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 of penalty parameter ρ, an initializing value for the parameter ρ is needed. ($\mathbf{y} \in \mathbb{R}^n, \mathbf{X} \in \mathbb{R}^{n\times p}, \boldsymbol{\beta} \in \mathbb{R}^p, \mathbf{A} \in \mathbb{R}^{m\times p}$, and m is the number of constraints.)

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

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

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

2 Initialization

The stationarity of KKT conditions for (2) 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 $sign(\boldsymbol{\beta})$ is the subgradient of $||\boldsymbol{\beta}||_1$. Since $|sign(\boldsymbol{\beta})| \leq 1_p$, we can transform above stationarity condition as follows:

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

Lemma 1 *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). (This can be proved by Seperating Hyperplane Theorem?)

Corollary 1 *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 optimal solution of (3). And also, we can get the solution λ_{max} corresponding to ρ_{max} by (3).*

$$\begin{aligned} \text{minimize} \quad & \rho \\ \text{subject to} \quad & z = \mathbf{A}^T\lambda \\ & z \leq -\nabla L(\boldsymbol{\beta}) + \rho 1 \\ & z \geq -\nabla L(\boldsymbol{\beta}) - \rho 1 \\ & \rho \geq 0 \end{aligned} \tag{3}$$

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

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_{max}

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:

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

3 Completeness of Active Set

In this section, our main question is : What happens if we do NOT activate all the violated predictors of $\mathcal{A}$?

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

$$\begin{aligned}
-\mathbf{X}_{:\mathcal{B}}^T(\mathbf{y} - \mathbf{X}_{:\mathcal{B}}(\boldsymbol{\beta}_{\mathcal{B}}^{(0)} - \Delta\rho\frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}})) \quad & + \\
(\rho_{max} - \Delta\rho)sign(\boldsymbol{\beta}_{\mathcal{B}}^{(0)} - \Delta\rho\frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}}) + \mathbf{A}_{:\mathcal{B}}^T(\lambda_{max} - \Delta\rho\frac{d}{d\rho}\lambda) \quad & = \quad 0 \qquad (4) \\
|\mathbf{X}_{:\mathcal{B}^c}^T(\mathbf{y} - \mathbf{X}_{:\mathcal{B}}(\boldsymbol{\beta}_{\mathcal{B}}^{(0)} - \Delta\rho\frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}})) - \mathbf{A}_{:\mathcal{B}^c}^T(\lambda_{max} - \Delta\rho\frac{d}{d\rho}\lambda)| \quad & \leq \\
(\rho_{max} - \Delta\rho)1_{|\mathcal{B}^c|} & \qquad (5) \\
\mathbf{A}_{:\mathcal{B}}(\boldsymbol{\beta}_{\mathcal{B}}^{(0)} - \Delta\rho\frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}}) \quad & = \quad 0 \qquad (6)
\end{aligned}$$

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 1**, we have following:

$$\begin{aligned}
-\mathbf{X}_{:\mathcal{B}}^T(\mathbf{y} - \mathbf{X}_{:\mathcal{B}}\boldsymbol{\beta}_{\mathcal{B}}^{(0)}) + \rho_{max}sign(\boldsymbol{\beta}_{\mathcal{B}}^{(0)}) + \mathbf{A}_{:\mathcal{B}}^T(\lambda_{max}) \quad & = \quad 0 \qquad (7) \\
|\mathbf{X}_{:\mathcal{B}^c}^T(\mathbf{y} - \mathbf{X}_{:\mathcal{B}}\boldsymbol{\beta}_{\mathcal{B}}^{(0)}) - \mathbf{A}_{:\mathcal{B}^c}^T\lambda_{max}| \quad & \leq \quad \rho_{max}1_{|\mathcal{B}^c|} \qquad (8) \\
\mathbf{A}_{:\mathcal{B}}\boldsymbol{\beta}_{\mathcal{B}}^{(0)} \quad & = \quad 0 \qquad (9)
\end{aligned}$$

Finally, we get

$$\begin{aligned}
\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}}\Delta\rho\frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}} - \Delta\rho sign(\boldsymbol{\beta}_{\mathcal{B}}^{(0)}) + \mathbf{A}_{:\mathcal{B}}^T\Delta\rho\frac{d}{d\rho}\lambda \quad & = \quad 0 \qquad (10) \\
\mathbf{A}_{:\mathcal{B}}\Delta\rho\frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}} \quad & = \quad 0 \qquad (11) \\
|\mathbf{X}_{:\mathcal{B}^c}^T\mathbf{X}_{:\mathcal{B}}\Delta\rho\frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}} + \mathbf{A}_{:\mathcal{B}^c}^T\Delta\rho\frac{d}{d\rho}\lambda| \quad & \leq \quad \Delta\rho 1_{|\mathcal{B}^c|} \qquad (12)
\end{aligned}$$

Let's focus on first two equations (10), (11):

$$\begin{bmatrix} \mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}} & \mathbf{A}_{:\mathcal{B}}^T \\ \mathbf{A}_{:\mathcal{B}} & 0 \end{bmatrix} \begin{bmatrix} \frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}} \\ \frac{d}{d\rho}\lambda \end{bmatrix} = \begin{bmatrix} sign(\boldsymbol{\beta}_{\mathcal{B}}^{(0)}) \\ 0 \end{bmatrix} \qquad (13)$$

And,

$$\begin{bmatrix} \frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}} \\ \frac{d}{d\rho}\lambda \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}} & \mathbf{A}_{:\mathcal{B}}^T \\ \mathbf{A}_{:\mathcal{B}} & 0 \end{bmatrix}^{-1} \begin{bmatrix} sign(\boldsymbol{\beta}_{\mathcal{B}}^{(0)}) \\ 0 \end{bmatrix} \tag{14}$$

Assumption 1 Rows of $\mathbf{A}_{:\mathcal{B}}$ are linearly independent.

So, $\begin{bmatrix} \mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}} & \mathbf{A}_{:\mathcal{B}}^T \\ \mathbf{A}_{:\mathcal{B}} & 0 \end{bmatrix}$ is invertible by **Assumption 1**.

Obviously, $\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}}$ is invertible. By inverse of block matrix, we have following equation:

$$\begin{bmatrix} \frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}} \\ \frac{d}{d\rho}\lambda \end{bmatrix} = \begin{bmatrix} (\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1} - (\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1}\mathbf{A}_{:\mathcal{B}}^T Z^{-1}\mathbf{A}_{:\mathcal{B}}(\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1} & (\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1}\mathbf{A}_{:\mathcal{B}}^T Z^{-1} \\ Z^{-1}\mathbf{A}_{:\mathcal{B}}(\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1} & -Z^{-1} \end{bmatrix} \begin{bmatrix} sign(\boldsymbol{\beta}_{\mathcal{B}}^{(0)}) \\ 0 \end{bmatrix} \tag{15}$$

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 $Z(Z^{-1})$ could be not only its inverse but also its generalized inverse.

Therefore,

$$\begin{aligned} \frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}} &= \left((\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1} - (\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1}\mathbf{A}_{:\mathcal{B}}^T Z^{-1}\mathbf{A}_{:\mathcal{B}}(\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1}\right) sign(\boldsymbol{\beta}_{\mathcal{B}}^{(0)}) \\ &= \left((\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1}(\mathbf{I} - \mathbf{W})\right) sign(\boldsymbol{\beta}_{\mathcal{B}}^{(0)}) \end{aligned} \tag{16}$$

,

where

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

.

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:

$$|\mathbf{X}_{:\mathcal{B}^c}^T\mathbf{X}_{:\mathcal{B}}\frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}} + \mathbf{A}_{:\mathcal{B}^c}^T\frac{d}{d\rho}\lambda| \quad \leq \quad 1_{|\mathcal{B}^c|} \tag{17}$$

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

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

$$\begin{bmatrix} \mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}} & \mathbf{A}_{:\mathcal{B}}^T \\ \mathbf{A}_{:\mathcal{B}} & 0 \end{bmatrix}^{-1} = \begin{bmatrix} 0 & \mathbf{A}_{:\mathcal{B}}^{-1} \\ (\mathbf{A}_{:\mathcal{B}}^T)^{-1} & -(\mathbf{A}_{:\mathcal{B}}^T)^{-1}\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}}\mathbf{A}_{:\mathcal{B}}^{-1} \end{bmatrix}. \tag{18}$$

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

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

(See Appendix 2)

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

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

  Because the order of left and right inverse matrix product is changed. So, $\frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}} \neq 0$.

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

- **$\mathbf{A}_{:\mathcal{B}}$ is invertible (by Assumption 1)**

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

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

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

- **$\mathbf{A}_{:\mathcal{B}}(\mathbf{X}_{:\mathcal{B}}^T\mathbf{X}_{:\mathcal{B}})^{-1}\mathbf{A}_{:\mathcal{B}}^T$ is NOT invertible** (See Appendix 1)
  So, we use generalized inverse.
  (**Assumption 2** Columns of $\mathbf{A}_{:\mathcal{B}}$ are linearly independent.)

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

  where $\mathbf{A}^-$ means generalized inverse of matrix $\mathbf{A}$. (This can be proved using the definition of generalized inverse and **Assumption 2**).
  Therefore, $\frac{d}{d\rho}\boldsymbol{\beta}_{\mathcal{B}} = 0$.

4 Appendix

4.1 Appendix 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 $H^Tx = 0$. $HH^Tx = 0$ and this means 0 is an eigenvalue of HH^T. Therefore, $|HH^T| = 0$ and HH^T is singular. (The determinant of matrix is the product of eigenvalues of the matrix).

4.2 Appendix 2

Why $|\mathcal{A}|$ is always $m+1$?

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.

$$\begin{aligned} \text{minimize} \quad & \rho \\ \text{subject to} \quad & z = \mathbf{A}^T \lambda \\ & z \leq -\nabla L(\boldsymbol{\beta}) + \rho 1 \\ & z \geq -\nabla L(\boldsymbol{\beta}) - \rho 1 \\ & \rho \geq 0 \end{aligned} \tag{19}$$

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

Find λ, and minimum ρ satisfying

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

. And predictors of $\mathcal{A}^C$ are set between lower and upper bounds of 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 $A \in \mathbb{R}^{q-1 \times p}$).