

Notes on Approximate Leave-One-Out for Elastic Net

Linyun He

Wanchao Qin

Peng Xu

Yuze Zhou

July 24, 2018

1 ALO for Elastic Net, Approximation in the Primal Domain

Recall the objective function for the elastic net problem:

$$\min_{\beta} \frac{1}{2} \sum_{j=1}^n (\mathbf{x}_j^\top \beta - y_j)^2 + \lambda \left(\alpha \|\beta\|_1 + \frac{1-\alpha}{2} \|\beta\|_2^2 \right). \quad (1)$$

Let $A = \{i : \beta_i \neq 0, i = 1, \dots, p\}$ be the active set, we have

$$\dot{\ell}(\mathbf{x}_j^\top \beta; y_j) = \mathbf{x}_j^\top \beta - y_j, \quad \ddot{\ell}(\mathbf{x}_j^\top \beta; y_j) = 1, \quad \nabla^2 R(\hat{\beta}_A) = (1-\alpha)\lambda \mathbf{I}_{A,A}.$$

Thus, Eqn. 31 reduces to

$$\mathbf{H} = \mathbf{X}_{\cdot,A} \left[\mathbf{X}_{\cdot,A}^\top \mathbf{X}_{\cdot,A} + (1-\alpha)\lambda \mathbf{I}_{A,A} \right]^{-1} \mathbf{X}_{\cdot,A}^\top. \quad (2)$$

By augmenting \mathbf{X} with an extra column of 1s, adding the intercept back to the model is straightforward, as Eqn. 31 now becomes

$$\mathbf{H} = [\mathbf{1}_n, \mathbf{X}_{\cdot,A}] \left\{ [\mathbf{1}_n, \mathbf{X}_{\cdot,A}]^\top \mathbf{D} [\mathbf{1}_n, \mathbf{X}_{\cdot,A}] + \nabla^2 R(\hat{\beta}_0, \hat{\beta}_A) \right\}^{-1} [\mathbf{1}_n, \mathbf{X}_{\cdot,A}]^\top, \quad (3)$$

where

$$\mathbf{D} = \text{diag} \left[\ddot{\ell}(\hat{\beta}_0 + \mathbf{x}_j^\top \hat{\beta}; y_j) \right]_{j \in A} = \mathbf{I}_{A,A}, \quad \nabla^2 R(\hat{\beta}_0, \hat{\beta}_A) = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & (1-\alpha)\lambda & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & (1-\alpha)\lambda \end{bmatrix}.$$

Then, the ALO can be computed as

$$\begin{bmatrix} 1 & \mathbf{x}_i^\top \end{bmatrix} \begin{bmatrix} \tilde{\beta}_0^{\setminus i} \\ \tilde{\beta}^{\setminus i} \end{bmatrix} = (\hat{\beta}_0 + \mathbf{x}_i^\top \hat{\beta}) + \frac{\mathbf{H}_{ii}}{1 - \mathbf{H}_{ii} \ddot{\ell}(\hat{\beta}_0 + \mathbf{x}_i^\top \hat{\beta}; y_i)} \dot{\ell}(\hat{\beta}_0 + \mathbf{x}_i^\top \hat{\beta}; y_i) \quad (4)$$

2 ALO for Elastic Net, Approximation in the Dual Domain

The original problem for elastic net is to solve for $\hat{\beta}$ such that:

$$\hat{\beta} = \arg \min_{\beta} \left(\frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \right) \quad (5)$$

By adding the Lagrangian, we get the formulation of L :

$$L = \frac{1}{2} \|\mathbf{y} - \mathbf{z}\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 + \mathbf{u}^\top (\mathbf{z} - \mathbf{X}\beta). \quad (6)$$

The original problem is solving the primal of the Lagrangian such that $p^* = \min_{\beta, \mathbf{z}} \max_{\mathbf{u}} L$ and the dual formulation $d^* = \max_{\mathbf{u}} \min_{\beta, \mathbf{z}} L$, to minimize over \mathbf{z} :

$$\frac{\partial L}{\partial \mathbf{z}} = \mathbf{z} - \mathbf{y} + \mathbf{u} = \mathbf{0} \implies \mathbf{y} = \mathbf{u} + \mathbf{z}.$$

Since β is penalized element-wisely, we can minimize over β by minimizing over each β_i , that is, we have to minimize $\lambda_1 |\beta_i| + \lambda_2 \beta_i^2 - \mathbf{u}^\top \mathbf{X}_i \beta$ for each dimension of β , where \mathbf{X}_i denotes the i th column of \mathbf{X} , therefore:

$$\min_{\beta} (\lambda_1 |\beta_i| + \lambda_2 \beta_i^2 - \mathbf{u}^\top \mathbf{X}_i \beta) = \begin{cases} 0 & |\mathbf{u}^\top \mathbf{X}_i| \leq \lambda_1, \\ -\frac{(\lambda_1 - |\mathbf{u}^\top \mathbf{X}_i|)^2}{4\lambda_2} & |\mathbf{u}^\top \mathbf{X}_i| > \lambda_1. \end{cases}$$

By taking all the above to the Lagrangian, we obtain the dual problem d^* as:

$$d^* = \min_{\mathbf{u}} \frac{1}{2} \|\mathbf{y} - \mathbf{u}\|_2^2 + \sum_{j: |\mathbf{X}_j^\top \mathbf{u}| > \lambda_1} \frac{(\lambda_1 - |\mathbf{u}^\top \mathbf{X}_j|)^2}{4\lambda_2}. \quad (7)$$

The minimizer $\hat{\mathbf{u}}$ could also be obtained from the dual problem through a proximal approach:

$$\hat{\mathbf{u}} = \mathbf{prox}_R(\mathbf{y}), \quad R(\mathbf{u}) = \sum_{j: |\mathbf{X}_j^\top \mathbf{u}| > \lambda_1} \frac{(\lambda_1 - |\mathbf{u}^\top \mathbf{X}_j|)^2}{4\lambda_2}.$$

By replacing the full data problem \mathbf{y} with $\mathbf{y}_\alpha = \mathbf{y} + (y_i^{\setminus i} - y_i)e_i$, where $y_i^{\setminus i}$ is the true LOO estimator and e_i is the i -th standard vector, and let $\mathbf{u}^{\setminus i} = \mathbf{prox}_R(\mathbf{y}_\alpha)$, we have:

$$\begin{aligned} 0 &= e_i^\top \mathbf{u}^{\setminus i} \\ &= e_i^\top \mathbf{prox}_R(\mathbf{y}_\alpha) \\ &\approx e_i^\top [\mathbf{prox}_R(\mathbf{y}) + \mathbf{J}_R(\mathbf{y})(\mathbf{y}_\alpha - \mathbf{y})] \\ &\approx \hat{u}_i + \mathbf{J}_{ii}(y_i^{\setminus i} - y_i). \end{aligned}$$

Here $\mathbf{J}_R(\mathbf{y})$ denotes the Jacobian matrix of the proximal operator at \mathbf{y} , thus the ALO estimator \tilde{y}_i is obtained as

$$\tilde{y}_i = y_i - \frac{\hat{u}_i}{\mathbf{J}_{ii}}. \quad (8)$$

The Jacobian could locally be obtained as:

$$\mathbf{J}_R(\mathbf{y}) = (\mathbf{I} + \nabla^2 R(\mathbf{prox}_R(\mathbf{y})))^{-1} = (\mathbf{I} + \nabla^2 R(\hat{\mathbf{u}}))^{-1} = \left(\mathbf{I} + \frac{1}{2\lambda_2} \mathbf{X}_E \mathbf{X}_E^\top \right)^{-1} \quad (9)$$

for $E = \{j : |\mathbf{X}_j^\top \mathbf{u}| > \lambda_1\}$.

3 ALO for Elastic Net, Approximation with Proximal Formulation

For the elastic net problem, the proximal mapping is known to be

$$\mathbf{prox}_R(\mathbf{z}) = \gamma \operatorname{sgn}(\mathbf{z}) \odot (|\mathbf{z}| - \lambda \mathbf{1}_p)_+, \quad \gamma = \frac{1}{1 + (1 - \alpha)\lambda}. \quad (10)$$

Let E be the active set, if $z_i \in E$, then

$$\frac{\partial}{\partial z_i} \gamma \operatorname{sgn}(z_i)(|z_i| - \lambda)_+ = \gamma.$$

Plug in $\mathbf{z} = \hat{\boldsymbol{\beta}} - \sum_{j=1}^n \dot{\ell}(\mathbf{x}_j^\top \hat{\boldsymbol{\beta}}; y_j) \mathbf{x}_j$, Eqn. 46 thus reduce to

$$\mathbf{H} = \gamma \mathbf{X}_{\cdot, E} \left[\gamma \mathbf{X}_{\cdot, E}^\top \mathbf{X}_{\cdot, E} + (1 - \gamma) \mathbf{I}_{E, E} \right]^{-1} \mathbf{X}_{\cdot, E}^\top. \quad (11)$$

Bringing back the intercept term is straightforward as well. Noted that

$$\begin{bmatrix} \hat{\beta}_0^{\setminus i} \\ \hat{\beta}^{\setminus i} \end{bmatrix} = \mathbf{prox}_R(\mathbf{z}), \quad \mathbf{z} = \begin{bmatrix} \hat{\beta}_0^{\setminus i} \\ \hat{\beta}^{\setminus i} \end{bmatrix} - \sum_{j \neq i} \begin{bmatrix} 1 \\ \mathbf{x}_j \end{bmatrix} \dot{\ell}(\hat{\beta}_0^{\setminus i} + \mathbf{x}_j^\top \hat{\boldsymbol{\beta}}^{\setminus i}; y_j).$$

Hence, from the first-order condition $\sum_{j \neq i} \dot{\ell}(\hat{\beta}_0^{\setminus i} + \mathbf{x}_j^\top \hat{\boldsymbol{\beta}}^{\setminus i}; y_j) = 0$, we can derive that

$$\mathbf{J}_{E, E} = [\mathbf{J}(\mathbf{u})]_{E, E} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & (1 - \alpha)\lambda & 0 & \dots & 0 \\ 0 & 0 & (1 - \alpha)\lambda & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & (1 - \alpha)\lambda \end{bmatrix}^{-1}. \quad (12)$$

The ALO formula is then immediate by Thm. 5.1.

4 ALO for LASSO, with Intercept through Generalized LASSO

For the generalized LASSO:

$$\min_{\boldsymbol{\beta}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\mathbf{D}\boldsymbol{\beta}\|_1, \quad (13)$$

the dual problem can be derived as:

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \|\mathbf{y} - \boldsymbol{\theta}\|_2^2, \quad \boldsymbol{\theta} \in \{\mathbf{X}^\top \boldsymbol{\theta} = \mathbf{D}^\top \mathbf{u}, \|\mathbf{u}\|_\infty \leq \lambda\}. \quad (14)$$

The dual problem could be written in a proximal approach, such that:

$$\hat{\mathbf{u}} = \mathbf{prox}_R(\mathbf{y}), \quad R(\mathbf{u}) = \begin{cases} 0 & \boldsymbol{\theta} \in \{\mathbf{X}^\top \boldsymbol{\theta} = \mathbf{D}^\top \mathbf{u}, \|\mathbf{u}\|_\infty \leq \lambda\}, \\ \infty & \text{otherwise.} \end{cases}$$

Denote \mathbf{J} as the Jacobian of the proximal operator at the full data problem \mathbf{y} , then the ALO estimator could be obtained as:

$$\mathbf{y}^{\setminus i} = \mathbf{y}_i - \frac{\hat{\mathbf{u}}_i}{\mathbf{J}_{ii}}. \quad (15)$$

For the case of LASSO with an intercept, we could expand the \mathbf{X} with a column of ones in the first column, expand $\boldsymbol{\beta}$ with another dimension and choose $\mathbf{D} = [\mathbf{0}, \mathbf{I}]$. Let $E := \{j : |\mathbf{X}_j^\top \boldsymbol{\theta}| = \lambda\}$ denote the active set. The Jacobian is locally given as the projection onto the orthogonal complement of the span of \mathbf{X}_E and the vector of ones. Further denote $\tilde{\mathbf{X}}_E = [\mathbf{1}, \mathbf{X}_E]$, then the Jacobian is given as $\mathbf{I} - \tilde{\mathbf{X}}_E(\tilde{\mathbf{X}}_E^\top \tilde{\mathbf{X}}_E)^{-1} \tilde{\mathbf{X}}_E^\top$.

5 ALO for Elastic Net, without Penalty on Intercept through Generalized LASSO

Without penalty on intercept, the elastic net problem can be written as:

$$\begin{aligned} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}} \end{bmatrix} &= \arg \min \frac{1}{2} \|\mathbf{y} - \beta_0 - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 \\ &= \arg \min \frac{1}{2} \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta} \end{bmatrix}^\top \left(\begin{bmatrix} 1 & \mathbf{X} \end{bmatrix}^\top \begin{bmatrix} 1 & \mathbf{X} \end{bmatrix} + \lambda_2 \text{diag}(0; \mathbf{1}_p) \right) \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta} \end{bmatrix} - \mathbf{y}^\top \begin{bmatrix} 1 & \mathbf{X} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta} \end{bmatrix} + \lambda_1 \|\boldsymbol{\beta}\|_1 \end{aligned}$$

where we assume that the size of \mathbf{X} is $n \times p$. In the mean time, note the LASSO problem (also without penalty on intercept) is:

$$\begin{aligned} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}} \end{bmatrix} &= \arg \min \frac{1}{2} \|\mathbf{y} - \beta_0 - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 \\ &= \arg \min \frac{1}{2} \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta} \end{bmatrix}^\top \begin{bmatrix} 1 & \mathbf{X} \end{bmatrix}^\top \begin{bmatrix} 1 & \mathbf{X} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta} \end{bmatrix} - \mathbf{y}^\top \begin{bmatrix} 1 & \mathbf{X} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta} \end{bmatrix} + \lambda_1 \|\boldsymbol{\beta}\|_1 \end{aligned}$$

Thus we can add some “observations” to the data and let

$$\mathbf{y}^* = \begin{bmatrix} \mathbf{y} \\ \mathbf{0}_p \end{bmatrix}, \quad \mathbf{X}^* = \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda_2} \mathbf{I}_p \end{bmatrix},$$

then the elastic net becomes

$$\begin{aligned} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}} \end{bmatrix} &= \arg \min \frac{1}{2} \left\| \mathbf{y}^* - \beta_0 \begin{bmatrix} \mathbf{1}_n \\ \mathbf{0}_p \end{bmatrix} - \mathbf{X}^* \boldsymbol{\beta} \right\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 \\ &= \arg \min \frac{1}{2} \left\| \begin{bmatrix} \mathbf{y} \\ \mathbf{0}_p \end{bmatrix} - \begin{bmatrix} \mathbf{1}_n & \mathbf{X} \\ \mathbf{0}_p & \sqrt{\lambda_2} \mathbf{I}_p \end{bmatrix} \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta} \end{bmatrix} \right\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1, \end{aligned} \quad (16)$$

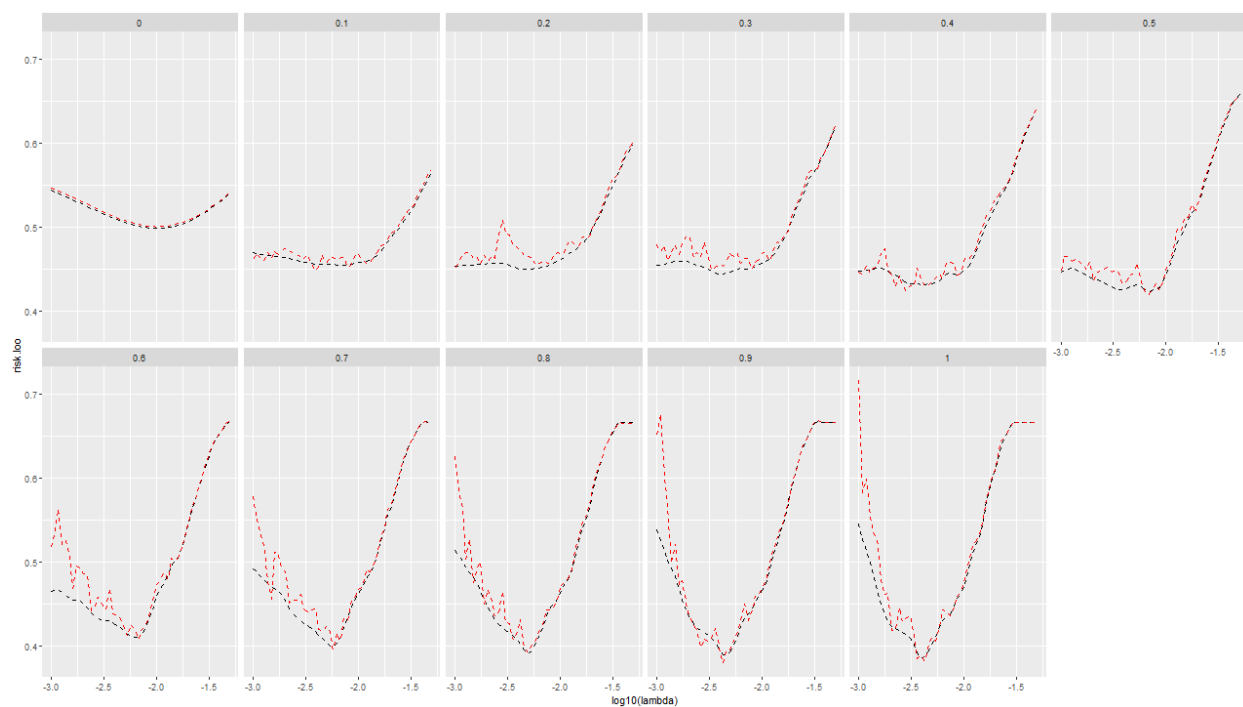
which is a special case of the general LASSO.

6 Usage of ALO formulae with `glmnet` package

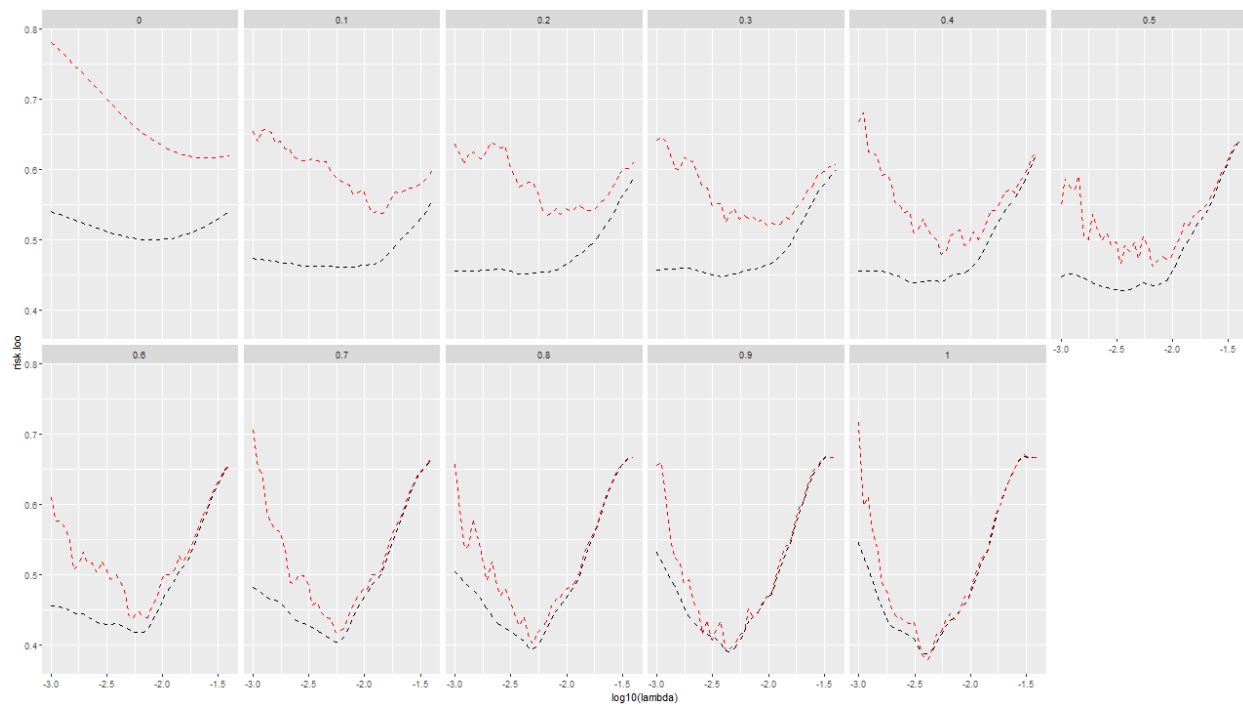
The `glmnet` package scales the elastic net loss function by a factor of $1/n$, so the ALO formulae must be adjusted accordingly, e.g. for the proximal case, we instead have:

$$\tilde{\mathbf{y}}_j^{\setminus i} = \hat{\mathbf{y}}_j + \frac{\mathbf{H}_{ii}(\hat{\mathbf{y}}_j - \mathbf{y}_j)}{n - \mathbf{H}_{ii}}, \quad \mathbf{H} = \gamma \mathbf{X}_{\cdot, E} \left[\frac{\gamma}{n} \mathbf{X}_{\cdot, E}^\top \mathbf{X}_{\cdot, E} + (1 - \gamma) \mathbf{I}_{E, E} \right]^{-1} \mathbf{X}_{\cdot, E}^\top.$$

Furthermore, `glmnet` implicitly “standardizes \mathbf{y} to have unit variance before computing its λ sequence (and then unstandardizes the resulting coefficients)” (*cf.* [Glmnet Vignette]). So to get comparable results, it is necessary to rescale \mathbf{y} by the MLE $\hat{\sigma}_y$ before fitting the model. Figure 1 shows the comparison of the ALO and LOO for different α s. Without standardizing \mathbf{y} first, a growing discrepancy between the two curves can be observed as $\alpha \rightarrow 0$.



(a) With standardization on y .



(b) Without standardization on y

Figure 1: ALO vs. LOO for elastic net with intercept, misspecification example.