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:

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

Let $A = \{i : \beta_i \notin K, i = 1, \ldots, p\}$ be the active set, we have

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

Thus, Eqn. 31 reduces to

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

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

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

where

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

Then, the ALO can be computed as

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

2 ALO for Elastic Net, Approximation in the Dual Domain

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

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

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

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

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

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

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

$$\min_{\boldsymbol{\beta}} \left(\lambda_1|\beta_i| + \lambda_2\beta_i^2 - \boldsymbol{u}^\top \boldsymbol{X}_i \boldsymbol{\beta} \right) = \begin{cases} 0 & |\boldsymbol{u}^\top \boldsymbol{X}_i| \le \lambda_1, \\ -\dfrac{(\lambda_1 - |\boldsymbol{u}^\top \boldsymbol{X}_i|)^2}{4\lambda_2} & |\boldsymbol{u}^\top \boldsymbol{X}_i| > \lambda_1. \end{cases}$$

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

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

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

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

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

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

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

$$\tilde{y}_i = y_i - \frac{\hat{u}_i}{\boldsymbol{J}_{ii}}. \tag{8}$$

The Jacobian could locally be obtained as:

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

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

3 ALO for Elastic Net, Approximation with Proximal Formulation

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

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

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

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

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

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

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

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

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

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

The ALO formula is then immediate by Thm. 5.1.

4 ALO for LASSO, with Intercept through Generalized LASSO

For the generalized LASSO:

$$\min_{\boldsymbol{\beta}}\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\| + \lambda\|\boldsymbol{D}\boldsymbol{\beta}\|_1, \tag{13}$$

the dual problem can be derived as:

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

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

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

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

$$\boldsymbol{y}^{\setminus i} = \boldsymbol{y}_i - \frac{\hat{\boldsymbol{u}}_i}{\boldsymbol{J}_{ii}}. \tag{15}$$

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

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

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

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

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

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

Thus we can add some "observations" to the data and let

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

then the elastic net becomes

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

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:

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

Furthermore, `glmnet` implicitly "standardizes 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 $\boldsymbol{y}$ by the MLE $\hat{\sigma}_y$ before fitting the model. Figure 1 shows the comparison of the ALO and LOO for different αs. Without standardizing $\boldsymbol{y}$ first, a growing discrepancy between the two curves can be observed as $\alpha \to 0$.

More precisely, `glmnet` is in fact optimizing the following problem:

$$\min_{\beta_0^*, \boldsymbol{\beta}^*} \frac{1}{2n} \sum_{j=1}^{n} \left(y_j^* - \beta_0^* - \boldsymbol{x}_j^T \boldsymbol{\beta}^* \right)^2 + \lambda\alpha \|\boldsymbol{\beta}^*\|_1 + \frac{1}{2}\lambda(1-\alpha)\|\boldsymbol{\beta}^*\|_2^2, \tag{17}$$

where

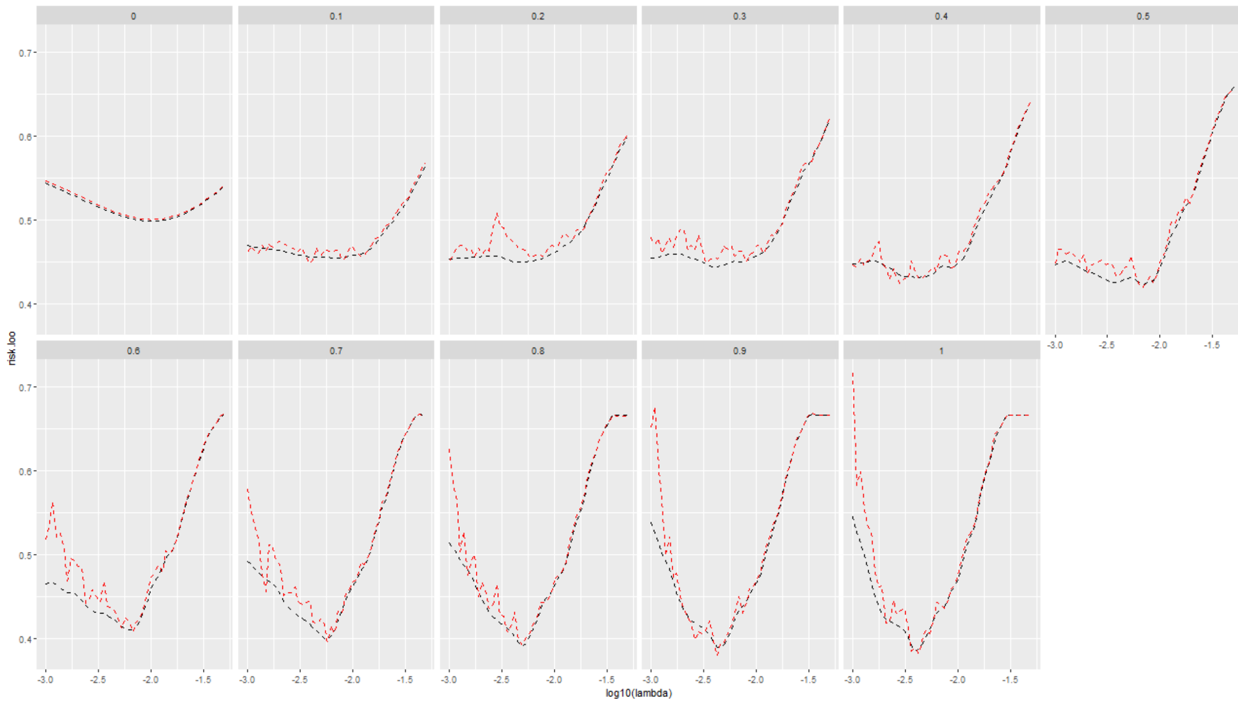
$$\boldsymbol{y}^* = \frac{\boldsymbol{y}}{\hat{\sigma}_y}, \qquad \beta_0^* = \frac{\beta_0}{\hat{\sigma}_y}, \qquad \boldsymbol{\beta}^* = \frac{\boldsymbol{\beta}}{\hat{\sigma}_y}.$$

As a result, to match the original elastic net problem, we may rescale $\boldsymbol{y}$, $\boldsymbol{X}$, and λ, to get:

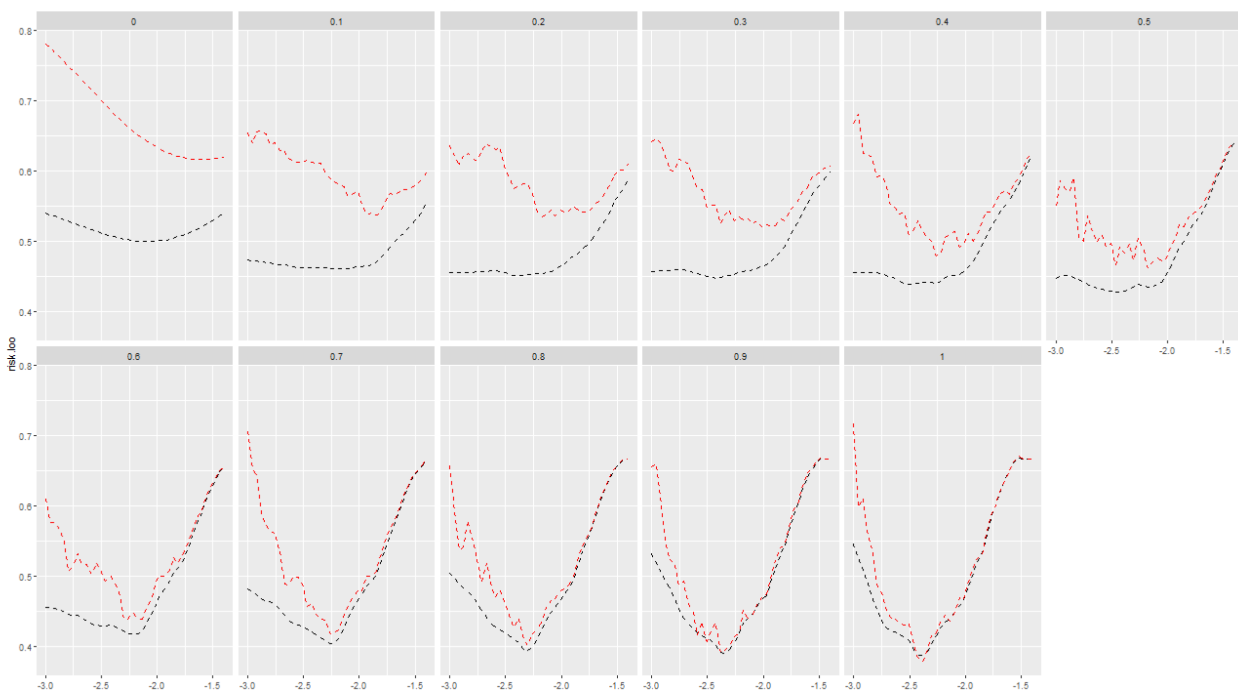
$$\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{2n} \sum_{j=1}^{n} \left(y_j^* - \beta_0^* - \boldsymbol{x}_j^{*\top} \boldsymbol{\beta} \right)^2 + \lambda^*\alpha \|\boldsymbol{\beta}\|_1 + \frac{1}{2}\lambda^*(1-\alpha)\|\boldsymbol{\beta}\|_2^2, \tag{18}$$

where

$$\boldsymbol{X}^* = \frac{\boldsymbol{X}}{\hat{\sigma}_y}, \qquad \lambda = \frac{\lambda}{\hat{\sigma}_y^2}.$$



(a) With standardization on $\boldsymbol{y}$.



(b) Without standardization on $\boldsymbol{y}$

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