Supplementary Material

Bo Zhao [† 1] **Xinwei Sun** [† 2] **Yanwei Fu** [‡ 3 4] **Yuan Yao** [‡ 5] **Yizhou Wang** [1]

Here we discuss more accurate estimation given by MSplit LBI compared with L_1 and L_2 regularization fail in the linear model with general design matrix X, i.e.

$$y = X\beta^\star + \varepsilon,\ \boldsymbol{S} = \{i : \beta_i^\star \gtrsim \sqrt{\frac{s \log p}{n}}\} \tag{1}$$

We first discuss the bias estimation of L_1 and L_2 model in Lemma 1 and 2.

Lemma 1. *Suppose the lasso estimator*

$$\beta^{lasso} = \arg\min_\beta \frac{1}{2N}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{2}$$

Suppose the model selection consistency holds at λ_n, i.e. $\boldsymbol{S}_{\lambda_n} = \boldsymbol{S}$, then we have

$$\mathbb{E}(\beta_{\boldsymbol{S}}^{lasso}) = \beta_{\boldsymbol{S}}^\star + \lambda_n (X_{\boldsymbol{S}}^\star X_{\boldsymbol{S}})^{-1} \rho_{\boldsymbol{S}}(\lambda_n) \tag{3}$$

where $\rho(\lambda_n) \in \partial\|\beta^{lasso}(\lambda_n)\|_1$.

Proof. Take derivative of (3) w.r.t β and set it to 0, and combined with the fact that $\beta_{\boldsymbol{S}^c} = 0$, we have

$$\begin{aligned}\lambda_n \rho_{\boldsymbol{S}}(\lambda_n) &= -X_{\boldsymbol{S}}^\star (y - X\beta^{lasso}(\lambda_n)) \\ &= -X_{\boldsymbol{S}}^\star \left(X_{\boldsymbol{S}}\beta_{\boldsymbol{S}}^\star + \varepsilon - X_{\boldsymbol{S}}\beta_{\boldsymbol{S}}^{lasso}(\lambda_n)\right)\end{aligned}$$

Hence,

$$X_{\boldsymbol{S}}^\star X_{\boldsymbol{S}} \beta_{\boldsymbol{S}}(\lambda_n) - \beta_{\boldsymbol{S}}^\star) = X_{\boldsymbol{S}}^\star \varepsilon + \lambda_n \rho_{\boldsymbol{S}}(\lambda_n)$$

Then

$$\beta_{\boldsymbol{S}}(\lambda_n) = \beta_{\boldsymbol{S}}^\star + (X_{\boldsymbol{S}}^\star X_{\boldsymbol{S}})^{-1} (X_{\boldsymbol{S}}^\star \varepsilon + \rho_{\boldsymbol{S}}(\lambda_n))$$

(3) holds after we take expectation on $\beta_{\boldsymbol{S}}(\lambda_n)$. □

Lemma 2. *Denote*

$$\begin{aligned} A &= X_{\boldsymbol{S}}^\star X_{\boldsymbol{S}} + \lambda I_{\boldsymbol{S},\boldsymbol{S}} \\ B &= X_{\boldsymbol{S}}^\star X_{\boldsymbol{S}^c} \\ C &= X_{\boldsymbol{S}^c}^\star X_{\boldsymbol{S}^c} + \lambda I_{\boldsymbol{S}^c,\boldsymbol{S}^c} \end{aligned}$$

then the Ridge Regression estimator

$$\beta^{ridge} = \arg\min_\beta \frac{1}{2N}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \tag{4}$$

have that

$$\begin{aligned}\mathbb{E}(\beta_{\boldsymbol{S}}^{ridge}) &= \beta_{\boldsymbol{S}}^\star + \lambda \left[A^{-1}B(C - B^T A^{-1} B)^{-1}\right] \beta_{\boldsymbol{S}^c}^\star \\ &- \lambda \left[A^{-1} + A^{-1}B(C - BA^{-1}B^T)^{-1} B^T A^{-1}\right] \beta_{\boldsymbol{S}}^\star \end{aligned} \tag{5}$$

$$\begin{aligned}\mathbb{E}(\beta_{\boldsymbol{S}^c}^{ridge}) &= \beta_{\boldsymbol{S}^c}^\star + \lambda (C - B^T A^{-1} B)^{-1} B^T A^{-1} \beta_{\boldsymbol{S}}^\star \\ &- \lambda (C - B^T A^{-1} B)^{-1} \beta_{\boldsymbol{S}^c}^\star \end{aligned} \tag{6}$$

Proof. It's easy to verify after taking the derivative of $\frac{1}{2N}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$ and set it to 0. □

Remark 1. *For the uniqueness of $\beta^\star$, we assume the restricted convex condition, i.e. that $X_{\boldsymbol{S}}^\star X_{\boldsymbol{S}} \succcurlyeq \lambda_{\boldsymbol{S}}$, hence the $\lambda \left[A^{-1} + A^{-1}B(C - BA^{-1}B^T)^{-1} B^T A^{-1}\right] \beta_{\boldsymbol{S}}^\star$ in 5 introduced in the estimation of $\beta_{\boldsymbol{S}}^\star$ can not be ignored.*

Next, we discuss the estimation property of dense estimator of MSplit LBI. We will show that as $\nu \to \infty$, not only it can give no-bias estimation for strong signals, but also for weak signals.

It's shown in (Huang et al., 2016) that when $\kappa \to \infty$, $\alpha \to 0$, the Split LBI algorithm converges to

$$0 = -\nabla_\beta X^\star (X\beta_t - y) - \frac{D^T(D\beta_t - \gamma_t)}{\nu} \tag{7a}$$

$$\rho_t = -\frac{D^T(\gamma_t - D)}{\nu} \tag{7b}$$

$$\rho_t \in \partial\|\gamma_t\|_1, \tag{7c}$$

Then it can be shown in the following lemma that the MSplit LBI can give more accurate estimation:

[†]Equal contribution

[1]Nat'l Eng. Lab. for Video Technology; Key Lab. of Machine Perception (MoE); Cooperative Medianet Innovation Center, Shanghai; Sch'l of EECS, Peking University. Deepwise Inc.
[2]Sch'l of Mathematical Science, Peking University. Deepwise Inc.
[3]Sch'l of Data Science, Fudan University. [4]AITrics Inc.
[5]Hong Kong University of Science and Technology; Peking University.
‡Correspondence to: Yanwei Fu <yanweifu@fudan.edu.cn>, Yuan Yao <yuany@ust.hk>.

Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

Lemma 3. *Denote*

$$G = \left(I - X_{\boldsymbol{S}}(X_{\boldsymbol{S}}^{\star}X_{\boldsymbol{S}})^{-1}X_{\boldsymbol{S}}^{\star}\right) X_{\boldsymbol{S}^c}$$

Then under linear model, If there exists $\bar{t}$ in 7 satisfies that $\widetilde{\boldsymbol{S}}_t = \boldsymbol{S}$, we have

$$\mathbb{E}(\beta_{\boldsymbol{S},\bar{t}}) = \beta_{\boldsymbol{S}}^{\star} + (X_{\boldsymbol{S}}^{\star}X_{\boldsymbol{S}})^{-1}X_{\boldsymbol{S}}^{\star}X_{\boldsymbol{S}^c}\left[I + \nu X_{\boldsymbol{S}^c}^{\star}G\right]^{-1}\beta_{\boldsymbol{S}^c}^{\star} \tag{8}$$

$$\mathbb{E}(\beta_{\boldsymbol{S}^c,\bar{t}}) = \beta_{\boldsymbol{S}^c}^{\star} - \left[I + \nu X_{\boldsymbol{S}^c}^{\star}G\right]^{-1}\beta_{\boldsymbol{S}^c}^{\star} \tag{9}$$

Furthermore, we have that

$$\lim_{\nu\to\infty} \|\mathbb{E}(\beta_{\boldsymbol{S},\bar{t}}) - \beta_{\boldsymbol{S}}^{\star}\|_2^2 = 0 \tag{10}$$

$$\lim_{\nu\to\infty} \|\mathbb{E}(\beta_{\boldsymbol{S}^c,\bar{t}}) - \beta_{\boldsymbol{S}^c}^{\star}\|_2^2 = 0 \tag{11}$$

Proof. It's easy to obtain (8) and (9). To prove 10 and 11, note that

$$G\beta_{\boldsymbol{S}^c}^{\star} = X_{\boldsymbol{S}^c}\beta_{\boldsymbol{S}^c}^{\star} - \mathrm{P}_{X_{\boldsymbol{S}}}X_{\boldsymbol{S}^c}\beta_{\boldsymbol{S}^c}^{\star}$$

Then we have

$$\begin{aligned} G\beta_{\boldsymbol{S}^c}^{\star} = 0 &\iff \min_z \|X_{\boldsymbol{S}^c}\beta_{\boldsymbol{S}^c}^{\star} - X_{\boldsymbol{S}}z\|_2^2 = 0 \\ &\iff \exists\, z,\ \text{s.t. } X_{\boldsymbol{S}}z = X_{\boldsymbol{S}^c}\beta_{\boldsymbol{S}^c}^{\star} \end{aligned}$$

Therefore, for the identifiable of $\beta_{\boldsymbol{S}^c}^{\star}$, we have that $G\beta_{\boldsymbol{S}^c}^{\star} \neq 0$, i.e. $\|G\beta_{\boldsymbol{S}^c}^{\star}\|_2^2 \neq 0$, hence $\beta_{\boldsymbol{S}^c}^{\star} \in \mathrm{Im}(G^TG)$. Denote the eigenvalue-decomposition of G as $G = U\Lambda U^T$ and $\lambda_G := \Lambda_{\min}(G^TG)$, then we have

$$\begin{aligned} \left[I + \nu X_{\boldsymbol{S}^c}^{\star}G\right]^{-1}\beta_{\boldsymbol{S}^c}^{\star} &= \left(I + \nu G^TG\right)^{-1}\beta_{\boldsymbol{S}^c}^{\star} \\ &= U(I + \nu\Lambda)^{-1}U^T\beta_{\boldsymbol{S}^c}^{\star} \end{aligned} \tag{12}$$

Hence we have

$$\|U(I + \nu\Lambda)^{-1}U^T\beta_{\boldsymbol{S}^c}^{\star}\|_2 \leq \frac{1}{1 + \nu\lambda_G}\|\beta_{\boldsymbol{S}^c}^{\star}\|_2$$

If we denote

$$\begin{aligned} A =& X_{\boldsymbol{S}}^{\star}X_{\boldsymbol{S}},\ B = X_{\boldsymbol{S}}^{\star}X_{\boldsymbol{S}^c} \\ \Lambda_X :=& \sqrt{\Lambda_{\max}(X^{\star}X)}, \end{aligned}$$

then we have

$$\begin{aligned} &\left\|A^{-1}B\left(I + \nu G^TG\right)^{-1}\beta_{\boldsymbol{S}^c}^{\star}\right\|_2 \\ \leq& \|A^{-1}\|_2\|B\|_2\frac{1}{1 + \nu\lambda_G}\|\beta_{\boldsymbol{S}^c}^{\star}\|_2 \\ \leq& \frac{\Lambda_X^2}{\lambda_{\boldsymbol{S}}(1 + \nu\lambda_G)}\|\beta_{\boldsymbol{S}^c}^{\star}\|_2 \end{aligned}$$

Then 10 and 11 hold. □

References

Chendi Huang, Xinwei Sun, Jiechao Xiong, and Yuan Yao. Split lbi: An iterative regularization path with structural sparsity. advances in neural information processing systems. *Advances In Neural Information Processing Systems*, pages 3369–3377, 2016. (document)