Regression Shrinkage and Selection via the Lasso

He Li, ***hl3420***

To better interpret model and reduce the prediction accuracy, many methods have been produced, e.g., subset selection, ridge regression and garotte estimator. Consider data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ where $\boldsymbol{x} \in \mathbb{R}^p$, ridge regression has the following form,

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = \underset{\alpha, \boldsymbol{\beta}}{\arg\min} \left\{ \sum_{i=1}^n \left(y_i - \alpha - \boldsymbol{\beta}^\top \boldsymbol{x}_i \right)^2 \right\}, \quad \|\boldsymbol{\beta}\|_2^2 \leq t, \tag{1}$$

for some hyper-parameter $t \geq 0$. The garotte estimator (Breiman, 1995) minimizes,

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = \underset{\alpha, \boldsymbol{\beta}}{\arg\min} \left\{ \sum_{i=1}^n \left(y_i - \alpha - \sum_{j=1}^p c_j \hat{\beta}_j^o x_{ij} \right)^2 \right\}, \quad c_j \geq 0, \ \sum_{j=1}^p c_j \leq t, \tag{2}$$

where $\hat{\boldsymbol{\beta}}^o = \left(\hat{\beta}_1^o, \cdots, \hat{\beta}_p^o \right)^\top$ is the OLS estimator.

This paper proposes a new method called "lasso" for the estimation task in linear model. Lasso estimator is a shrinkage estimation method and explicitly set some coefficients to 0, it has a continuous path and leads to an interpretable model. The author illustrates the advantage of lasso estimator over its counterparts under the orthonormal design case and provides parameter estimation algorithm.

The lasso estimator is defined by the following optimization problem,

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = \underset{\alpha, \boldsymbol{\beta}}{\arg\min} \left\{ \sum_{i=1}^n \left(y_i - \alpha - \boldsymbol{\beta}^\top \boldsymbol{x}_i \right)^2 \right\}, \quad \|\boldsymbol{\beta}\|_1 \leq t. \tag{3}$$

Under orthonormal design, $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}$, and we can easily derive the closed form solution for lasso estimator as,

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^o)(|\hat{\beta}_j^o| - \lambda)^+.$$

The ridge estimator,

$$\hat{\beta}_j = \frac{1}{1+\lambda} \hat{\beta}_j^o.$$

The garotte estimator,

$$\hat{\beta}_j = \left(1 - \frac{\lambda}{\hat{\beta}_j^{o2}} \right)^+ \hat{\beta}_j^o.$$

and the best subset estimator,

$$\hat{\beta}_j = \hat{\beta}_j^o \, \mathbf{1}_{|\hat{\beta}_j^o| > \lambda}.$$

Please see details in the appendix. As shown later, if the design is not orthonormal, lasso produces the same estimator while ridge and garotte ones will rely on the covariance structure of the response data vector.

For the hyper-parameter tuning, this paper provides three different methods where the first two are classical technique in model selection, namely 5-fold cross validation and the generalized

cross validation statistic. A final Stein's method is proposed with closed form solution and better computation performance over the first two.

The parameter estimation methods presented here are computationally expensive and does not utilize the piecewise linear structure of lasso's regularization path. Therefore, they cannot be compared with later algorithms like Least Angle Regression or Dantzig (Dantzig uses a different loss function).

To conclude, this paper considers a novel regularization scheme for linear regression task. Although using l_1 norm as the regularizer seems to be natural these days, it indeed was a breakthrough back then. Lasso estimator is similar to its previous counterparts, the garotte estimator, but is more stable under different OLS estimators and covariance structure of $\boldsymbol{X}$. However, the parameter estimation methods are crude and the standard error calculation for inference task is unadjusted. And it does not touch any theoretical analysis of the lasso estimator, which leaves many future work.

References

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics 37*(4), 373–384.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

Appendix

Under orthonormal design, $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}$, and the OLS estimator is $\hat{\boldsymbol{\beta}}_o = \boldsymbol{X}^\top \boldsymbol{y}$. The Lagrangian form of ridge regression is,

$$\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^2 + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}.$$

Therefore,

$$\begin{aligned} 2\boldsymbol{X}^\top \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ridge}}\right) + 2\lambda \hat{\boldsymbol{\beta}}_{\text{ridge}} = 0 \\ \implies \hat{\boldsymbol{\beta}}_{\text{ridge}} = \left(\boldsymbol{X}^\top \boldsymbol{X} + \lambda I\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \frac{1}{1+\lambda} \hat{\boldsymbol{\beta}}_o. \end{aligned}$$

Similarly, for lasso estimator,

$$\begin{aligned} \arg\min_{\boldsymbol{\beta}} \frac{1}{2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_1 &= \arg\min_{\boldsymbol{\beta}} -\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{\beta} + \lambda \|\boldsymbol{\beta}\|_1 \\ &= \arg\min_{\boldsymbol{\beta}} -\hat{\boldsymbol{\beta}}_o^\top \boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{\beta} + \lambda \|\boldsymbol{\beta}\|_1. \end{aligned}$$

Therefore, the minimization task for each $j = 1, \cdots, p$ is,

$$\hat{\beta}_j^{\text{lasso}} = \arg\min_{\beta_j} -\hat{\beta}_j^o \beta_j + \frac{1}{2}\beta_j^2 + \lambda |\beta_j|.$$

WLOG, assume $\hat{\beta}_j^o > 0$, then $\hat{\beta}_j^{\text{lasso}}$ must > 0 as well or else we can always flip the sign of $\hat{\beta}_j^{\text{lasso}}$ and get a lower loss. Therefore,

$$\hat{\beta}_j^{\text{lasso}} = \arg\min_{\beta_j > 0} -\hat{\beta}_j^o \beta_j + \frac{1}{2}\beta_j^2 + \lambda \beta_j = \text{sign}\left(\hat{\beta}_j^o\right)\left(\hat{\beta}_j^o - \lambda\right)^+$$

For the garotte estimator,

$$\arg\min_{\boldsymbol{C}} \frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{C}\hat{\boldsymbol{\beta}}_o\right)^2 + \lambda \text{tr}(\boldsymbol{C}),\ \boldsymbol{C} = \text{diag}(c_1, \cdots, c_p),\ c_j \geq 0.$$

Therefore the optimization task is,

$$\begin{aligned} &\arg\min_{\boldsymbol{C}} -\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{C}\hat{\boldsymbol{\beta}}_o + \frac{1}{2}\hat{\boldsymbol{\beta}}_o^\top \boldsymbol{C}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{C}\hat{\boldsymbol{\beta}}_o + \lambda \text{tr}(\boldsymbol{C}) \\ =&\arg\min_{\boldsymbol{C}} -\hat{\boldsymbol{\beta}}_o \boldsymbol{C}\hat{\boldsymbol{\beta}}_o + \frac{1}{2}\hat{\boldsymbol{\beta}}_o^\top \boldsymbol{C}^2 \hat{\boldsymbol{\beta}}_o + \lambda \text{tr}(\boldsymbol{C}),\ \boldsymbol{C} = \text{diag}(c_1, \cdots, c_p),\ c_j \geq 0. \end{aligned}$$

The garotte estimator for each $j = 1, \cdots, p$ is,

$$\hat{\beta}_j^{\text{gar}} = \hat{\beta}_j^o c_j^{\text{gar}} = \hat{\beta}_j^o \arg\min_{c_j} -c_j \hat{\beta}_j^{o2} + \frac{1}{2}c_j^2 \hat{\beta}_j^{o2} + \lambda c_j = \left(1 - \frac{\lambda}{\hat{\beta}_j^{o2}}\right)\hat{\beta}_j^o.$$

Finally, the best subset selection choose the k response variables with larger simple linear regression coefficient. Under orthonormal design, the simple linear regression coefficient is simply $\boldsymbol{y}^\top \boldsymbol{x}_j$ where $\boldsymbol{x}_j = (x_{1j}, \cdots, x_{nj})^\top$. It is equivalent to select the k largest OLS estimation coefficient since $\hat{\boldsymbol{\beta}}_o = \boldsymbol{X}^\top \boldsymbol{y}$.