The Elastic Net

Subho Majumdar

Literature seminar talk
School of Statistics, University of Minnesota

Introduction

- Penalized regression setup: intended to tackle rank-deficiency of design matrix, correlated predictors and also to do variable selection.

- **Ridge regression**: upper bound on L_2-norm of coefficient vector.
  - Ridge criterion $L_1(\lambda, \beta) = |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2 + \lambda|\boldsymbol{\beta}|^2$.
  - Has closed form solution: $\hat{\boldsymbol{\beta}}_{ridge} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$
  - Reduces MSE but biased estimator, helpful when $\mathbf{X}$ is not full-rank.

- **LASSO** (*Tibshirani, 1996)*: upper bound on L_1 norm.
  - LASSO criterion: $L_2(\lambda, \beta) = |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2 + \lambda|\boldsymbol{\beta}|_1$.
  - Solution obtained through cross-validation on λ by iterative algorithms like LARS, pathwise coordinate descent etc.
  - Provides sparse solutions: useful for variable selection in $n << p$ case.

So... where's the problem??

Ridge can always provide a unique estimate with lesser MSE than OLS, but it doesn't do any variable selection like lasso. However, there are a number of difficulties associated with with lasso:

- In $p > n$ cases, lasso selects at most n variables.
- Lasso problem is not well-defined unless the bound on L_1-norm of coefficients is less than a certain value.
- **Grouping effect**: When there is a group of variables which are highly correlated pairwise, lasso selects only one variable from the group.
- Ridge outperforms lasso in the usual $n > p$ scenario (Tibshirani, 1996).

A simple solution... combine the two!

Define the **Naïve Elastic Net** (NEN) criterion (Zou and Hastie, 2005) as:

$$L(\lambda_1, \lambda_2, \beta) = |\mathbf{y} - \mathbf{X}\beta|^2 + \lambda_1|\beta| + \lambda_2|\beta|^2, \quad \lambda_1, \lambda_2 \geq 0$$

on normalized response and predictors.

A solution is obtained by viewing this as a penalized least squares:

$$\hat{\beta} = \arg\min_{\beta} |\mathbf{y} - \mathbf{X}\beta|^2, \quad \text{subject to } (1-\alpha)|\beta|_1 + \alpha|\beta|^2 \leq t \text{ for some } t$$

with $\alpha = \lambda_1/(\lambda_1 + \lambda_2)$ and $(1-\alpha)|\beta|_1 + \alpha|\beta|^2$ the elastic net penalty. We have $\alpha \in (0, 1)$, and for $\alpha = 0$ or 1 the procedure becomes lasso or ridge, respectively.

Why the name?

- For all $\alpha \in (0, 1)$ the penalty function is non-differentiable at 0 (like lasso penalty) and but strictly convex (like ridge penalty).
- By varying α we can control the proportion of ridge/lasso penalty.

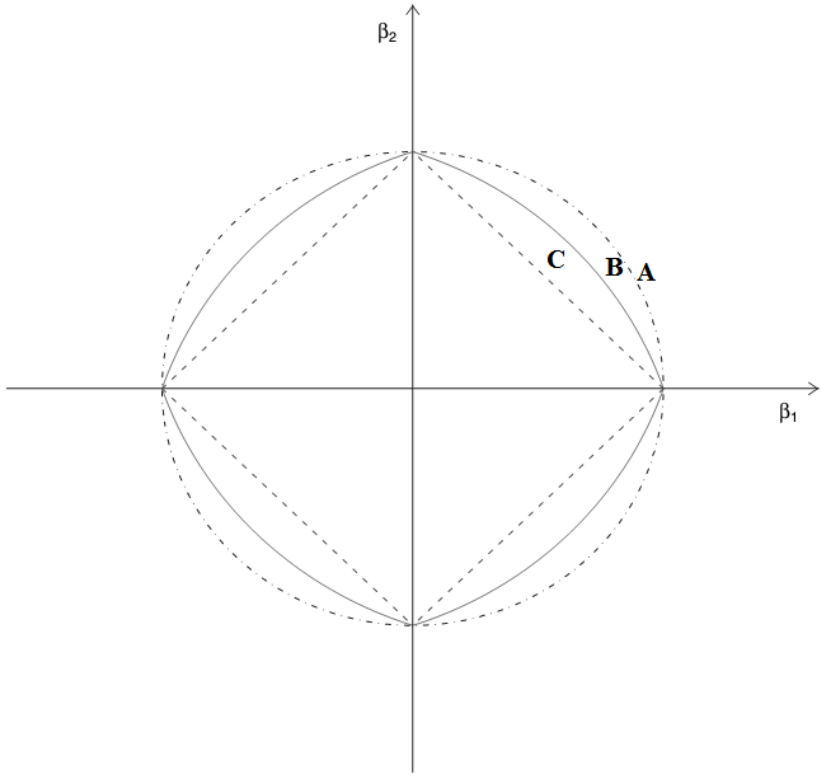

Figure : For $p = 2$, plots of (A) ridge, (B) elastic net at $\alpha = 0.5$ and (C) lasso penalty

Form of the solution

Given dataset $(\mathbf{y}, \mathbf{X})$ and tuning parameters (λ_1, λ_2) define an artificial dataset $(\mathbf{y}^*, \mathbf{X}^*)$ as

$$\mathbf{X}^*_{(n+p)\times p} = \frac{1}{\sqrt{1+\lambda_2}} \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda_2}\mathbf{I}_p \end{pmatrix}, \quad \mathbf{y}^*_{(n+p)} = \begin{pmatrix} \mathbf{y} \\ 0_p \end{pmatrix}$$

Let $\gamma = \lambda_1/\sqrt{1+\lambda_2}, \beta^* = \sqrt{1+\lambda_2}\beta$. Then the NEN criterion can be written as

$$L(\gamma, \beta) = L(\gamma, \beta^*) = |\mathbf{y}^* - \mathbf{X}^*\beta^*|^2 + \gamma|\beta^*|_1$$

The elastic net solution in original setup turns out to be just a scalar multiple of the solution to the above lasso minimization problem in transformed setup, i.e. $\hat{\beta}^*$ as

$$\hat{\beta} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\beta}^*, \quad \text{with } \hat{\beta}^* = \arg\min_{\beta^*} L(\gamma, \beta^*)$$

Lemma

Consider the general penalized regression setup with objective function $|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2 + \lambda J(\boldsymbol{\beta})$, with $\hat{\boldsymbol{\beta}}$ being its minimizer. Assume that $\mathbf{x}_i = \mathbf{x}_j$ for some $i, j \in \{1, ..., p\}$. Then

(a) If $J(.)$ is strictly convex, then $\hat{\beta}_i = \hat{\beta}_j$ for all $\lambda > 0$.

(b) If $J(\boldsymbol{\beta}) = |\boldsymbol{\beta}|_1$, then $\hat{\beta}_i \hat{\beta}_j \geq 0$, and $\hat{\boldsymbol{\beta}}^*$ is another minimizer of the above objective function, then

$$\hat{\beta}_k^* = \begin{cases} \hat{\beta}_k & \text{if } k \neq i \text{ and } k \neq j \\ s(\hat{\beta}_i + \hat{\beta}_j) & \text{if } k = i \\ (1-s)(\hat{\beta}_i + \hat{\beta}_j) & \text{if } k = j \end{cases}$$

for any $s \in [0, 1]$.

This means that for the extreme case of two predictors being exactly equal, lasso doesn't give a unique solution for their coefficients.

Since the elastic net penalty is strictly convex for any $\alpha \in (0, 1)$, it doesn't suffer from this deficiency of lasso.

The actual elastic net estimate

Prediction performance of Naïve Elastic Net is not good unless it is very close to Ridge or Lasso.

The NEN estimate is obtained by first getting ridge coefficients for a fixed λ_2 then using that to obtain a lasso solution to the modified problem. The *ridge shrinkage followed by lasso shrinkage* increases the bias of the estimate but the variance doesn't come down much, thus overall prediction error increases.

The final **elastic net (EN) estimate** is obtained by multiplying the NEN estimate with $1 + \lambda_2$:

$$\hat{\beta}_{EN} = (1 + \lambda_2)\beta_{NEN}$$

This reverts back the ridge shrinkage. Empirical evidence suggests this estimate performs well in variable selection and prediction compared to both lasso and ridge.

Computation of estimate

- Computation is done by **LARS-EN**, a modification of the LARS algorithm (Efron *et al*, 2004) formulated to obtain a solution of the lasso problem.
- Solution is obtained by cross-validation over a grid of values for (λ_1, λ_2).
- Iteratively updates the fits for each coordinate of the coefficient vector, just like LARS. For m iterations, it takes $O(m^3 + pm^2)$ FLOPs.
- The algorithm involves obtaining laso fits over the modified data $(\mathbf{y}^*, \mathbf{X}^*)$ as described before. The modified matrix of predictors $\mathbf{X}^*$ has dimension $(n + p) \times p$, which increases computational burden for $n << p$ scenarios. In these cases early stopping might be required.

Discussion

- **Connection with bridge regression**: both Bridge regression and elastic net are generalizations of lasso/ ridge, but elastic net can produce sparse solutions like lasso, while bridge cannot.
- Application in classification problems: Li and Jia, 2010; Hong, Chen and Harris, 2013.

References

Zou H.; Hastie T. Regularization and variable selection via the elastic net. *J. R. Statist. Soc.B*, **2005**, 67, 301-320.

Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.B*, **1996**, 58, 267-288.

Efron B.; Hastie T.; Johnstone I.; Tibshirani R. Least angle regression. *Ann. Statist.*, **2004**, 32, 407-840.

Li J-T.; Jia Y-M. An Improved Elastic Net for Cancer Classification and Gene Selection. *Act. Aut. Sinica*, **2010**, 36, 976-981.

Hong X.; Chen S.; Harris C.J. Elastic-Net Prefiltering for Two-Class Classification. *IEEE Transactions on Cybernetics*, **2013**, 43, 286-295.

THANK YOU!