Summary of LASSO Paper in 1995

1. Background

Linear regression is very common and its performance improvement is a critical topic, statisticians liked to do subset selection or ridge regression before 1995. However, them both have problems, subset selection improves model interpretation but is not stable, ridge regression shows an advantage in prediction accuracy but cannot simplify models. Therefore, inspired by the garotte method, Robert Tibshirani proposed a new way: LASSO (least absolute shrinkage and selection operator), which combines good features of subset selection and ridge regression.

2. Idea & Formula

Suppose we have data $\left(\mathbf{x}^i, y_i\right), i = 1, 2, \ldots, N$, where $\mathbf{x}^i = \left(x_{i1}, \ldots, x_{ip}\right)^{\mathrm{T}}$ are predictors and y_i are responses.

In 1993, Breiman developed the brilliant non-negative garotte method:

$$min \sum_{i=1}^{N} \left(y_i - \alpha - \sum_j c_j \hat{\beta}_j^0 x_{ij} \right)^2 \quad subject\ to\ c_j \geqslant 0,\ \sum_j c_j \leqslant t$$

where $\hat{\beta}_j^0$ is the OLS (ordinary least square) estimate.

The original idea for LASSO is very simple, Robert just wanted to combine two steps above into one. The main process of LASSO can be explained as "OLS + L1 penalty":

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = \arg\min \left\{ \sum_{i=1}^{N} \left(y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right\} \quad subject\ to \sum_j |\beta_j| \leqslant t$$

If $t_0 = \sum |\hat{\beta}_j^0|,\ t < t_0$ will cause shrinkage and even set parameters as 0.

Robert said that one drawback of garotte is its use of LSE (least square estimates). Interestingly, later work showed that this actually helps garotte to outperform LASSO.

3. Properties of LASSO

3.1 Geometry Analysis

Robert has analyzed solutions of several methods in orthonormal design cases. Figure 1 clearly shows the problems of subset selection and ridge regression, and found that LASSO and garrote are almost the same except garrote did a "weighted" shrinkage. Figure 2 compares LASSO and ridge in 2-dimension cases and answers why LASSO estimates can be 0.

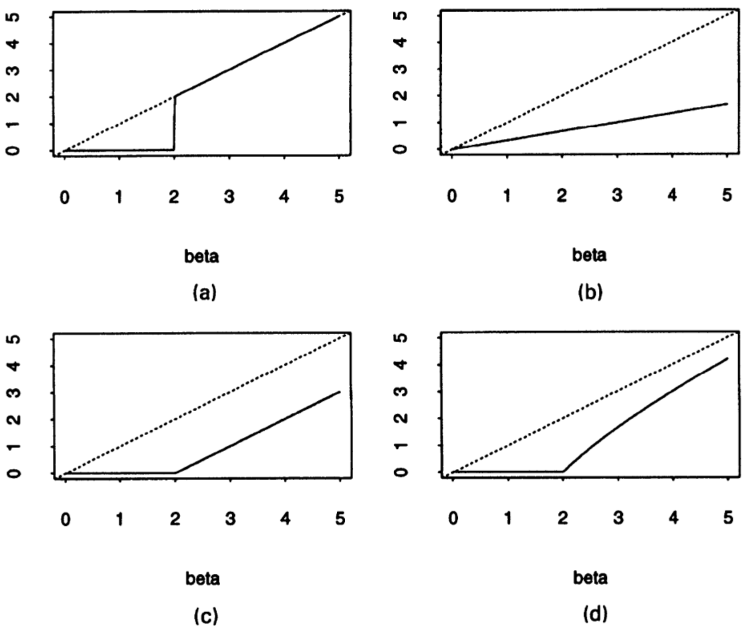

Fig. 1. (a) Subset regression, (b) ridge regression, (c) the lasso and (d) the garrote: ———, form of coefficient shrinkage in the orthonormal design case; ··········, 45°-line for reference

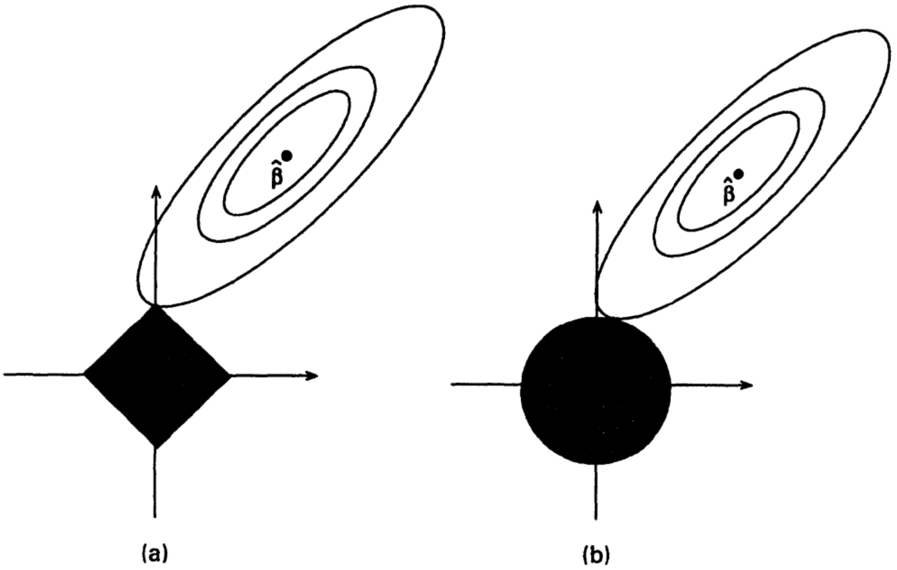

Fig. 2. Estimation picture for (a) the lasso and (b) ridge regression

Robert also stated that there is no general guarantee that signs of LASSO estimates are the same with LSE, though it is true when $p = 2$.

So why does LASSO uses L1 penalty? Robert mentioned ridge regression, whose constraint is of the form $\sum_j |\beta_j|^q \leqslant t$, he said that not only because L1 penalty ($q = 1$) is closer to subset selection ($q \rightarrow 0$)than ridge ($q = 2$) but also it is the smallest value giving a convex region (Figure 9).

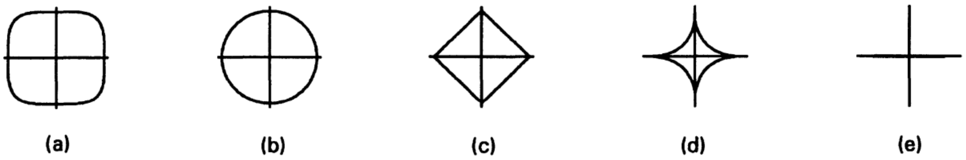

Fig. 9. Contours of constant value of $\Sigma_j|\beta_j|^q$ for given values of q: (a) $q=4$; (b) $q=2$; (c) $q=1$; (d) $q=0.5$; (e) $q=0.1$

3.2 Variance and Distribution

How to calculate variance of LASSO estimates? First, bootstrap is a general way to estimate statistics. Besides, Robert pointed out that LASSO can be viewed as a special case of ridge regression, then the approximated close form can be written down, but it always outputs 0 when lasso estimate $\beta_j = 0$.

As for the distribution, Robert showed that LASSO estimates can be seen as the Bayes posterior mode. Plot LASSO and ridge regression estimates' density, one can see that LASSO estimate density puts more mass near 0 and in the tails, which also explains why LASSO can shrinkage parameters to 0.

3.3 Orthonormal Design Case

When design matrix X is orthonormal, LASSO's solutions are exactly the form of soft thresholding, this identity partially supports LASSO's utility in linear models, but in general conditions, theorems are much harder to build.

4. LASSO Algorithm

4.1 Getting t

t is a hyperparameter in LASSO, Robert provided three ways to catch it, one is using prediction error to do cross-validation, another is to write LASSO in ridge regression form and then do generalized cross-validation, the third one is under the support of Stein's theoretical work, we can find t in closed form:

$$\hat{t} = \sum \left(\left| \hat{\beta}_j^o \right| - \hat{\gamma} \right)^+$$

Stein's method is the most efficient one, though it requires orthonormal design.

4.2 Finding Solutions

At that time, the absolute constraints in LASSO need to be tested 2^p times and thus is quite hard to solve, Robert thought about two algorithms. The first one is introducing the inequality constraints sequentially, while another algorithm writes β_j as $\beta_j^+ - \beta_j^-$, then the original constraints transform to:

$$\beta_j^+ \geqslant 0, \beta_j^- \geqslant 0 \ \ and \ \ \Sigma_j \beta_j^+ + \Sigma_j \beta_j^- \leqslant t$$

This transform brings more variables ($2p$) with fewer constraints ($2p+1$).

These two algorithms are both far from perfect in complexity. Luckily, Robert and his colleagues found a faster and beautiful algorithm: least angle regression (LARS), which promotes the application of LASSO in practice.

5. Comparison & Extension

When doing regression in prostate cancer data, LASSO did exhibit its superiority. Compared to ridge regression, LASSO sets some parameters 0; compared to subset selection, it shrinks the coefficients and their Z-scores, reflecting its stability.

4 simulation examples also show that in general cases that "small to moderate number of moderate-sized effects", LASSO is pretty good. Although ridge/subset selection may do better in some extreme settings, their performances are not as stable as LASSO.

As for extending LASSO to generalized linear regression, Robert said that the key is to see problem as an iteratively reweighted least squares procedure, then LASSO can then be applied in each iteration. Robert also tried to use LASSO in tree-based models and splines.

6. Conclusion

Robert's LASSO is a continuous shrinkage process that can produce 0 coefficients, it performs well in both theoretical research and real data analysis comparing with subset selection and ridge regression. As the first method pointing out the potential ability of absolute penalty, LASSO is a milestone in statistics. In the following two decades after 1995, many works focused on digging up the value in LASSO and received great success.