Regularization and variable selection via the elastic net

Zou and Hastie (2003); J.R. Statist. Soc. B, 301–320

For any fixed non-negative λ1 and λ2 , we define the naıve elastic net criterion

$$L(\lambda_1,\lambda_2,\beta) = \|y-X\beta\|^2+\lambda_2\|\beta\|^2+\lambda_1|\beta|_1 \qquad (3)$$

The naıve elastic net estimator $\hat{\beta}$ is the minimizer of (3), $\hat{\beta}=arg\,min_\beta\ L(\lambda_1,\lambda_2,\beta)$. This procedure can be viewed as a penalized least squares method.

Let $\alpha=\lambda_2/(\lambda_1+\lambda_2)$;

then solving $\hat{\beta}$ in equation (3) is equivalent to the optimization problem

$$\hat{\beta} = arg\,min_\beta\ \|y-X\beta\|^2, \quad subject\,t. \quad (1-\alpha)|\beta|_1+\alpha|\beta|^2 \le t \quad for\,some\ t \qquad \textbf{(5)}$$

We call the function $(1-\alpha)\,|\beta|1 + \alpha|\beta|^2$ the *elastic net penalty*, which is a convex combination of the lasso and ridge penalty.
When α= 1, the na¨ıve elastic net becomes simple ridge regression.
Here, we consider only α< 1. For all α ∈ [0, 1) , the elastic net penalty function is singular (without first derivative) at 0 and it is strictly convex for all α> 0, thus having the characteristics of both the lasso and ridge regression. Note that the lasso penalty (α= 0) is convex but not strictly convex.

These arguments can be seen clearly from Fig. 1:

304 *H. Zou and T. Hastie*

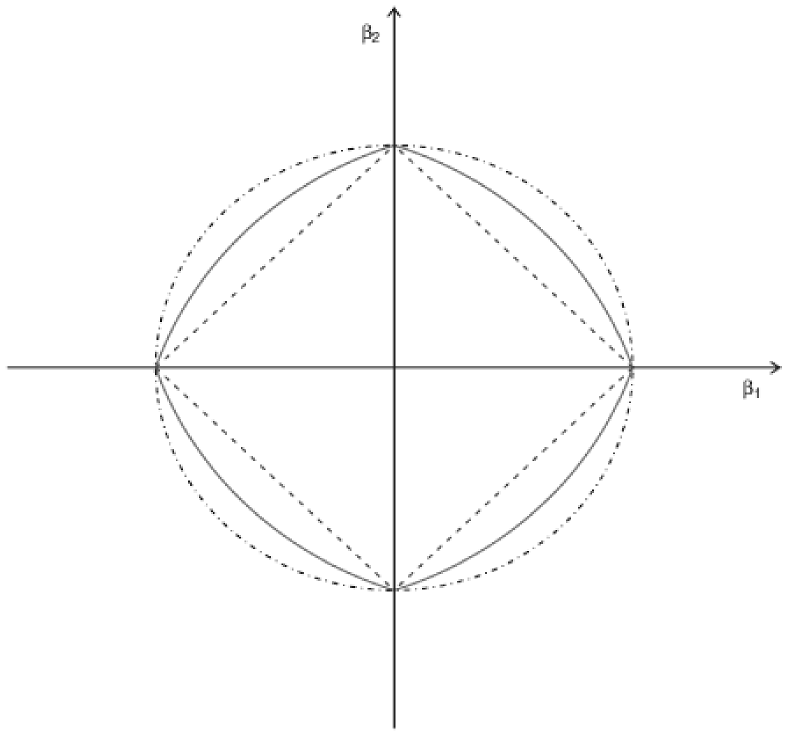

Fig. 1. Two-dimensional contour plots (level 1) (· - - - - -, shape of the ridge penalty; - - - - - - -, contour of the lasso penalty; ———, contour of the elastic net penalty with $\alpha = 0.5$): we see that singularities at the vertices and the edges are strictly convex; the strength of convexity varies with α

The Adaptive Lasso and its Oracle Properties

Hui Zou (2006)

Let us consider the weighted lasso

$$argmin_{\beta} \ \|y - \sum_{j=1}^{p} \beta_j \boldsymbol{x}_j\|^2 \ + \ \lambda \cdot \sum_{j=1}^{p} w_j \, |\beta_j|$$

where $\boldsymbol{w}$ is a known weights vector. ...show that if the weights are data-dependent and cleverly chosen, the weighted lasso can possess the oracle properties. The new methodology is called the *adaptive lasso*.
We now define the adaptive lasso. Suppose $\hat{\beta}$ is a root-n consistent estimator to β_*. For example, we can use $\widehat{\beta_{ols}}$.

Pick a $\gamma > 0$, and define the weight vector $\hat{w} = \frac{1}{\hat{\beta}^{\gamma}}$.

The *adaptive lasso estimates* $\widehat{\beta_{(n)}}$ are given by

$$\hat{\beta}_{(n)} = argmin_{\beta} \ \|y - \sum_{j=1}^{p} \beta_j \boldsymbol{x}_j\|^2 \ + \ \lambda \sum_{j=1}^{p} \hat{w}_j \, |\beta_j| \quad . \tag{6}$$

It is worth emphasizing that (6) is a convex optimization problem, thus it does not suffer from the multiple local minimal issue and its global minimizer can be efficiently solved.