Penalized Regression – biased estimators

If $J : \mathbb{R}^p \to [0, \infty)$ is any function we replace the least squares estimate by the penalized least squares estimate

$$\hat{\beta}^\lambda = \operatorname*{argmin}_\beta \mathrm{RSS}(\beta) + \lambda J(\beta).$$

The optimization problem is nicest if J is convex. The parameter $\lambda \geq 0$ determines the tradeoff between the measure of fit to data, RSS, and the penalty on the parameter, J.

The function J implements a priori preferences of some parameters over others. It is a frequentists version of a Bayesian incorporation of prior beliefs.

To a Bayesian we are computing the posterior mode when we use the prior

$$c(\lambda/2\sigma^2)^{-1} \exp\left(-\frac{\lambda}{2\sigma^2} J(\beta)\right), \quad c(\lambda) = \int \exp(-\lambda J(\beta)) \mathrm{d}\beta$$

on the mean value parameter β.

Ridge Regression

If $J(\beta) = \beta^T\beta = ||\beta||^2$ the penalized estimation method is known as ridge regression.

We need to optimize

$$(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta.$$

The function J is strictly convex with $J(\beta) \to \infty$ for $||\beta|| \to \infty$. There is always a unique minimum $\hat{\beta}^{\text{ridge}}$ when $\lambda > 0$.

Lasso

If $J(\beta) = \sum_{i=1}^{p} |\beta_i| = ||\beta||_1$ the penalized estimation method is known as lasso = least absolute shrinkage and selection operator.

We need to optimize

$$(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda \sum_{i=1}^{p} |\beta_i|.$$

The function J is convex with $J(\beta) \to \infty$ for $||\beta|| \to \infty$. If there is a unique least squares solution there is a unique minimum $\hat{\beta}^{\text{lasso}}$.

This is a convex, but non-differentiable optimization problem.

Systematic choice of λ

With $\hat{\beta}^\lambda$ the estimator for given λ choose λ by minimizing

$$\lambda \mapsto \text{EPE}(\lambda)$$

with $\text{EPE}(\lambda)$ the expected prediction error for the predictor

$$x \mapsto \beta_0^\lambda + \sum_{i=1}^{p} x_i \hat{\beta}_i^\lambda = (1, x^T)\hat{\beta}^\lambda.$$

$\text{EPE}(\lambda)$ can be estimated using an independent test/validation data set $(\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_{\tilde{N}}, \tilde{y}_{\tilde{N}})$ as

$$\hat{\text{EPE}}(\lambda) = \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} L(\tilde{y}_i, (1, \tilde{x}_i^T)\hat{\beta}^\lambda).$$

Lecture exercises

Solve 3.5 and 3.12 in the book

Restricted Estimation

If $C \subseteq \mathbb{R}^p$ the restricted estimator is the estimator

$$\hat{\beta}^C = \operatorname*{argmin}_{\beta \in C} (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta).$$

The optimization problem is nicest if C is convex. A well known situation is when C is a subspace parameterized as

$$C = \{A\delta \mid \delta \in \mathbb{R}^q\}$$

where A is a $p \times q$ rank q matrix. The solution is

$$\hat{\beta}^C = A(A^T\mathbf{X}^T\mathbf{X}A)^{-1}A^T\mathbf{X}^T\mathbf{y}.$$

Duality

If $J : \mathbf{R}^p \to [0, \infty)$ is a function we can define the sub-level sets

$$C_J(s) = \{\beta \mid J(\beta) \leq s\}.$$

If J is convex then $C_J(s)$ is convex for all s. The function

$$\lambda \to s(\lambda) := J(\hat{\beta}^\lambda)$$

is typically a continuous, strictly decreasing function with $s(\lambda) \to 0$ for $\lambda \to \infty$ mapping $[0, \infty)$ onto $(0, s(0)]$.

$$\boxed{\hat{\beta}^\lambda = \hat{\beta}^{C_J(s(\lambda))}}$$

This gives a dual viewpoint on the penalized estimator as a restricted estimator and vice versa for level set restrictions.

Figure 3.11 – Ridge and Lasso as Restricted Estimators

Ridge regression (right) is a constraint optimization problem over a set with a smooth boundary. Lasso (left) is a constraint optimization problem over a set where the boundary has corners. The corners give lasso the selection ability.

Duality

Penalization can be viewed as an implicit model restriction – but in a data dependent way through $s(\lambda)$.

The parameterized family of solutions $(\hat{\beta}^{C_J(s)})_{s \in (0, s(0)]}$ is identical to the family $(\hat{\beta}^{\lambda})_{\lambda \geq 0}$.

For lasso, optimization of

$$(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$$

subject to $||\beta||_1 \leq s$ is a quadratic optimization problem subject to linear constraints, which is a classical numerical problem.

Bridge Regression and The Elastic Net

Generalizations include

- Bridge regression

$$J(\beta) = \sum_{i=1}^{p} |\beta_i|^q$$

  for $q \in (0, \infty)$
  - $q = 2$ is ridge regression
  - $q = 1$ is lasso
  - $q < 1$ is non-convex
  - $q \to 0$ is best subset selection

- The elastic net

$$J(\beta) = \alpha \beta^T \beta + (1 - \alpha) \sum_{i=1}^{p} |\beta_i|$$

  for $\alpha \in [0, 1]$.

Figure 3.12 – Bridge and Elastic Net

For $q \leq 1$ gives corners and has the selection property. For $q < 1$ we have a non-convex problem, $q \to 0$ results in best subset selection. With $q = 1$ we get selection as well as convexity. The elastic net has selection but more convexity.

Figure 3.18 – Comparisons

For a 2 dimensional parameter we can illustrate how the chosen parameters behave for different methods and different choices of selection/regularization.

Note that only ridge and lasso provide estimates on the entire curve plottet. The other three methods provide only one alternative to the least squares estimate (4,2).