3

LASSO and Bridge regression

If $\mathbf{X}$ denotes an $N \times p$ matrix of N p-dimensional covariates and $\mathbf{y}$ denotes an N-dimensional vector of observations we consider for fixed $\gamma > 0$ the penalized residual sum of squares

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

for a $\lambda > 0$. The *Bridge regression* estimate of β is defined as the β that minimizes this penalized residual sum of squares. It may not be unique. The *LASSO regression* estimate of β is the special case of the Bridge regression where $\gamma = 1$. (LASSO = "Least Absolute Shrinkage and Selection Operator").

In the formulation above, it is assumed that the matrix of covariates as well as the observation vector have been centered, that is, the average of the columns as well as of $\mathbf{y}$ equal 0.

As for ridge regression, the ordinary least squares estimate is obtained by minimizing $\text{Bridge}_0(\beta)$, and we denote by

$$t = \min_{\beta: \mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T y} \sum_{i=1}^{p} |\beta_i|^\gamma$$

the minimal value of the penalty function over the least squares solution set.

In this exercise, we will *only consider Bridge regression with* $\gamma \geq 1$

Show that $\text{Bridge}_\lambda(\beta)$ for $\lambda > 0$ is convex function, which is strictly convex for $\gamma > 1$. Show that for $\gamma > 1$ there is a unique minimizer, $\hat{\beta}(\lambda)$, with $\sum_{i=1}^{p} |\hat{\beta}_i(\lambda)|^\gamma < t$. Then show that for $\gamma = 1$ there exists a minimizer and for all minimizers, $\hat{\beta}(\lambda)$, the penalty function takes the same value,

$$s(\lambda) := \sum_{i=1}^{p} |\hat{\beta}_i(\lambda)|^\gamma < t.$$

In other words, the LASSO estimate may not be unique, but all minimizers give rise to the same penalty value. The function $s(\lambda)$ is therefore well defined as a function of $\lambda > 0$ – for $\gamma = 1$ as well as for $\gamma > 1$.

Show that $s(\lambda)$ as a function of λ on the interval $(0, \infty)$ is continuous, strictly decreasing and tends to 0 for $\lambda \to \infty$. Thus it maps $(0, \infty)$ in a one-to-one manner onto the interval $(0, t)$.

Show that the minimizers of $\text{Brigde}_\lambda(\beta)$ also minimize

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

subject to the constraint

$$\sum_{i=1}^{p} |\beta_i|^\gamma \leq s(\lambda).$$

Show on the other hand, that the minimizers of

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

subject to the constraint

$$\sum_{i=1}^{p} |\beta_i|^\gamma \leq s(\lambda)$$

also minimize $\text{Brigde}_\lambda(\beta)$.

For $\gamma = 2$ we get back ridge regression with it's explicit solution. For $\gamma = 1$ we have the LASSO estimate, which can be seen by the constraint formulation to be equivalent to a quadratic optimization problem with *linear* constraints – a very well studied class of constrained optimization problems. The penalty formulation on the other hand seems a little unpleasant due to the non-differentiable penalty function. Note, however, that the translation of the λ-penalty into a constraint is data dependent and not explicit, and the "regularization" in the LASSO estimator is typically and more conveniently given directly in terms of the constraint s rather than the penalty parameter λ. For $\gamma \neq 2$ and $\gamma > 1$ we are faced with a quadratic optimization problem over a *convex* region – or alternatively a differentiable objective function in the penalty formulation. Neither problems seems particularly nasty from a numerical point of view.

For $\gamma < 1$ we face on the contrary either a quadratic optimization problem over a *non-convex* region or a non-differentiable optimization problem, and both seem nasty from an analytic as well as a numerical point of view. We haven't even established the equivalence of the two problems!