`glmnet`: Elastic net – Penalised regression methods

Torben

May, 2019

Introduction and (one) motivation

In analysis of high-dimensional data, we often face the situation of having more predictors, p, than observations, n, i.e $n < p$. This often the case in genomics, where the number of genetic markers by far exceeds the number of samples. We denote X the $n \times (p)$-design matrix.

For the ordinary least squares this causes a problem, as $\hat{\beta}^{ols} = (X^\top X)^{-1} X^\top y$ implies that $(X^\top X)$ needs to be invertible, i.e. X needs to have full rank, which it does not when $n < p$.

Ridge regression

One way to deal with this is to add a full-rank matrix, λI_p to $X^\top X$, where I_p is the identity matrix. The constant $\lambda \geq 0$ is a tuning parameter, that needs to be specified, e.g. by cross-validation.

A little more maths show that in fact we have

$$\hat{\beta}^{ridge} = (X^\top X + \lambda I_p)^{-1} X^\top y,$$

implying that $\hat{\beta}^{ridge} = \hat{\beta}^{ols}$ for $\lambda = 0$.

One property that is relevant to discuss is bias and variance of the estimators in OLS and Ridge Regression. Hence, let us recall that $\hat{\beta}^{ols}$ is unbiased:

$$\begin{aligned}
\mathbb{E}(\hat{\beta}^{ols}) &= \mathbb{E}\{(X^\top X)^{-1} X^\top y\} \\
&= (X^\top X)^{-1} X^\top \mathbb{E}\{y\} \\
&= (X^\top X)^{-1} X^\top X \beta \\
&= \beta,
\end{aligned}$$

where we used that $y \sim (X\beta, \sigma^2 I_n)$ by assumption.

Bias

We could repeat this argument for $\hat{\beta}^{ridge}$, however it is more instructive to rewrite $\hat{\beta}^{ridge}$ in terms of $\hat{\beta}^{ols}$:

$$\begin{aligned}
\hat{\beta}^{ridge} &= [X^\top X + \lambda I_p]^{-1} X^\top y \\
&= [(X^\top X)\{I_p + \lambda (X^\top X)^{-1}\}]^{-1} X^\top y \\
&= \{I_p + \lambda (X^\top X)^{-1}\}^{-1} (X^\top X)^{-1} X^\top y \\
&= \{I_p + \lambda (X^\top X)^{-1}\}^{-1} \hat{\beta}^{ols},
\end{aligned}$$

where we assumed that $X^\top X$ is invertible.

From this we find that $\mathbb{E}(\hat{\beta}^{ridge}) = \{I_p + \lambda (X^\top X)^{-1}\}^{-1} \beta$, i.e. an biased estimator when $\lambda > 0$.

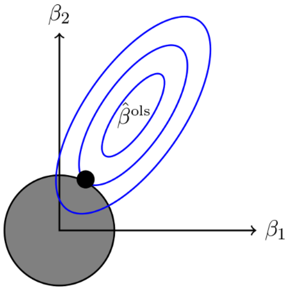

Figure 1: Ridge regression and OLS solutions. The intersection between the ellipsoid contour lines and the disc represents the Ridge solution relative to the OLS solution

However, we can also give a direct calculation of the variance for Ridge Regression and OLS, showing that

$$\text{trace}(\mathbb{V}[\hat{\beta}^{ols}]) = \sigma^2 \sum_{j=1}^{p} \frac{1}{d_j^2} \qquad\qquad \text{trace}(\mathbb{V}[\hat{\beta}^{ridge}]) = \sigma^2 \sum_{j=1}^{p} \frac{d_j^2}{(\lambda + d_j^2)^2},$$

where d_j is proportional to the sample variance in the j'th principal component.

A different view on Ridge Regression

One can also look at Ridge Regression differently, namely in terms of penalised regression, where a penalty term is applied to the squared sum of coefficient estimates:

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \|y - X\beta\|_2^2 \quad \text{subject to} \|\beta\|_2^2 \leq t,$$

for some positive constant t and $\|x\|_2^2 = \sum_{j=1}^{p} x_j^2$ is the ℓ_2-norm.

We can think of t as a "budget" for the regression, as we have to "spend" the regression parameter budget on the variables best explaining y from X. For this reason the data is also *scaled* before fitting to have zero mean and variance one (per column - `glmnet` does that automatically).

Using some Lagrange multipliers we can show that this is equivalent to

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2,$$

where λ is the same constant as before and determined through cross-validation.

Relationship between solutions to OLS and Ridge Regression

We can try to inspect the solutions graphically in two dimensions

The Lasso Regression

In the figure above we saw that the Ridge Regression contracts the solution, $\hat{\beta}^{ridge}$, towards the disc defined by the "budget"-parameter t. As a disc/sphere don't have pointy edges, is it rarely the case that any of the parameters in $\hat{\beta}^{ridge}$ are set to zero.

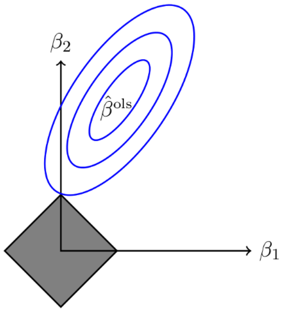

Figure 2: Lasso regression and OLS solutions. The intersection between the ellipsoid contour lines and the square represents the Lasso solution relative to the OLS solution.

When doing inference for linear regression, the elimination of insignificant terms is important. Typically we do so by successive removing terms from the model – either by some information criterion or hypothesis tests (for nested linear models).

However, the Lasso Regression makes variable selection while estimating the parameters. This is done by solving

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1,$$

where the penalty is the ℓ_1 norm, $\|x\|_1 = \sum_{j=1}^{p} |x_j|$, i.e. the sum of the absolute values.

The Lasso penalty generally sets more parameter values to zero than the Ridge Regression, where we seldom see any terms fixed to zero.

It is not as easy to express bias and variance for the Lasso since we generally don't have closed forms solutions to the likelihood equations. However, the general picture is that the larger λ, the more bias and consequently lower variance.

As for the Ridge Regression we can visualise the Lasso solution together with the OLS solution for two dimensions. Since, the Lasso penalty can be viewed as

$$\sum_{j=1}^{p} |\beta_j| \leq t,$$

we have a 45°-rotated square centered in origo where the Lasso solution exists.

A hand-waving argument for more sparse solutions comes from the intuition that it is more likely that the corners of the hyper-cube will intersect the ellipsoid. Corners result in one or more zero-parameters.

The Baysian perspective

In Bayesian statistics, we think of the data as fixed and the parameters being random. This is different from the frequentistic approach, where we think of the parameters having some *true* value.

In the Bayesian context, the Ridge Regression results from assigning a normal distributed prior on each component in the β-vector, with zero mean and some variance, τ^2. This implies that we *a priori* assumes many of the parameters to be close to zero.

The Lasso also assigns a zero-mean distribution, but with a Laplacian distribution that decays more rapidly towards zero implying that less terms are expected to be non-zero.

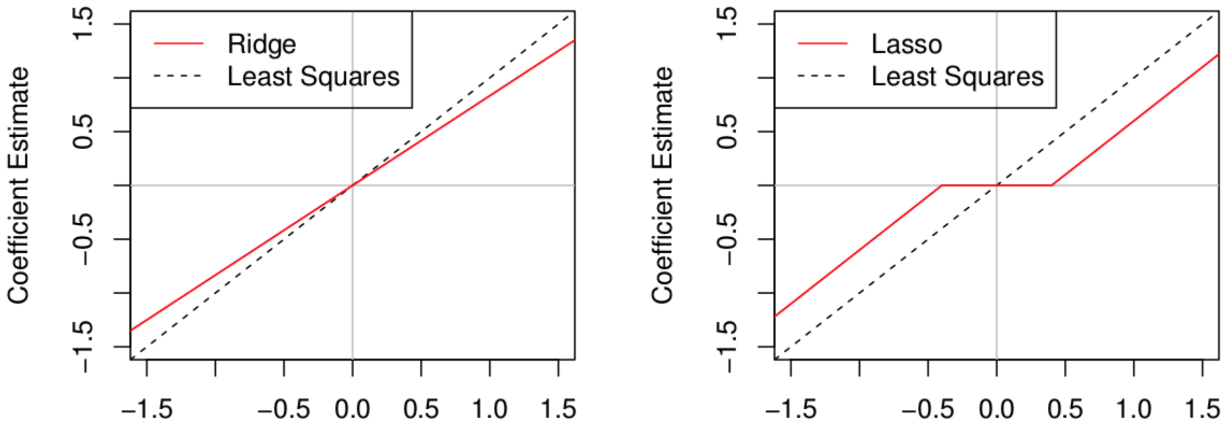

Figure 3: Ridge regession and LASSO - soft thresholding

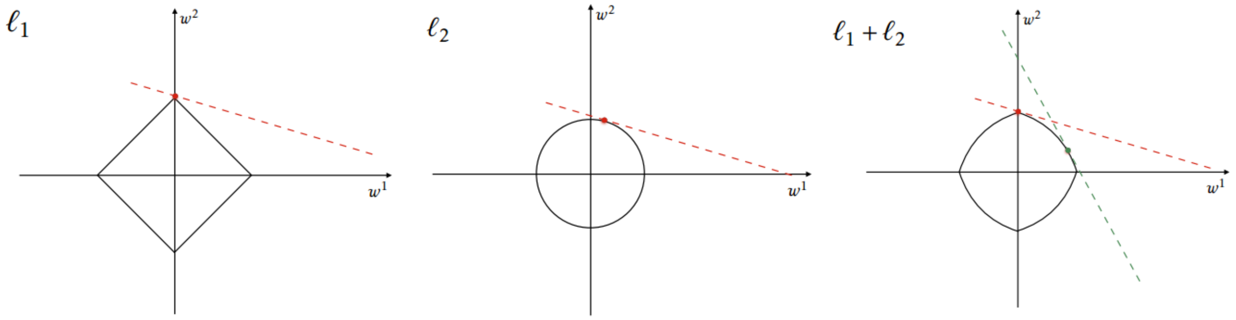

Figure 4: Lasso, Ridge and Elastic Net regularisation. Source: Wikipedia

Elastic Net – the best from two worlds?

A downside with the Lasso is that it may have difficulties when several variables are collinear, such that linear combinations of them are hard to distinguish. In such a case the Ridge Regression is better as it will typically form an average of the variables. Hence, for stable selection of variables in this case Ridge Regression may be preferred. However, Ridge Regression seldom sets any parameters to zero, i.e. no variable selection which is what we would like in the end…

The solution to the problem is Elastic Net, which incorporates both the ℓ_1 (Lasso) and ℓ_2 (Ridge) penalties in a convex way:

$$\hat{\beta}^{en} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\left(\alpha\|\beta\|_1 + \frac{1-\alpha}{2}\|\beta\|_2^2\right),$$

where α is yet another tuning parameter deciding the amount of Lasso ($\alpha = 1$) and Ridge ($\alpha = 0$) penalty that goes into the solution.

Both α and λ are selected based on cross-validation.

In the Figure below we see the three types of regularisation discussed above. The shape of the Elastic Net solution area depends on α - the closer to 1 the more square it is, and the closer to 0 the more spherical.

Further reading:

https://web.stanford.edu/~hastie/StatLearnSparsity/