# Regression and Shrinking via the Lasso

Asher Labovich

December 2024

**Key idea: Constrain the sum of absolute values of the coefficients**. In particular, we want to solve the problem

$$(\hat{\alpha}, \hat{\beta}) = \underset{(\alpha,\beta)}{\arg\min} \sum_{i=1}^{N} (y_i - \alpha - \sum_{j=1}^{p} \beta_j x_{ij})^2 = \underset{(\alpha,\beta)}{\arg\min} ||\mathbf{y} - \alpha - X\beta||^2 \text{ subject to} \sum_{i=1}^{p} |\beta_i| \leq t$$

**Why?** Because this can both efficiently select AND constrain coefficients.

This problem can be equivalently formulated as

$$\underset{(\alpha,\beta)}{\arg\min} L(\alpha, \beta) = \underset{(\alpha,\beta)}{\arg\min} ||\mathbf{y} - \alpha - X\beta||^2 + \lambda \sum_{i=1}^{p} |\beta_i|$$

Which is easier to work with computationally, and has a 1-to-1 correspondence between $\lambda$ and t.

If we normalize X so that $\sum_{i=1}^{N} X_{ij} = 0, \sum_{i=1}^{N} X_{ij}^2 = 1 \forall j$ (mean 0, std 1), then we see that $\frac{\partial L}{\partial \alpha} = 2(\sum_{i=1}^{N} y_i - \alpha - \sum_{j=1}^{p} B_j x_{ij}) = 0 \implies n\alpha = \sum_{i=1}^{N} y_i + \sum_{j=1}^{p} \sum_{i=1}^{N} B_j x_{ij} \implies \alpha = \overline{y} + \sum_{j=1}^{p} B_j \sum_{i=1}^{N} x_{ij} = \overline{y}$

So, we can remove $\alpha = \overline{y}$ and just set $\overline{y} = 0$.

With $\beta^0$ as the OLS estimates, we have that

$$\begin{aligned}
||y - X\beta|| = \\
||(y - X\beta^0) + (X\beta^0 - X\beta)|| = \\
[(y - X\beta^0) + (X\beta^0 - X\beta)]^T [(y - X\beta^0) + (X\beta^0 - X\beta)] = \\
||y - X\beta^0|| + ||(X\beta^0 - X\beta)|| + 2(X\beta^0 - X\beta)^T (y - X\beta^0) = \\
||y - X\beta^0|| + ||(X\beta^0 - X\beta)|| + 2(\beta^0 - \beta)^T X^T (y - X\beta^0) = \\
||y - X\beta^0|| + ||(X\beta^0 - X\beta)||
\end{aligned}$$

(Note: $X^T(y - X\beta^0) = 0$ since $X^T(y - X\beta^0) = X^T y - (X^T X)(X^T X)^{-1} X^T y = X^T y - X^T y = 0$. Intuitively, this is because linear regression projects y onto the column space of X, and thus the residuals are orthogonal to the column space of x, so $X^T r = 0$ for r $= y - X\beta^0$).

Since the first term is constant with respect to $\beta$, we have that the residual sum of squares

$$||y - X\beta|| = a + ||X\beta^0 - X\beta|| = a + (\beta^0 - \beta)^T X^T X(\beta^0 - \beta)$$

This has elliptical contours; e.g. when $||X\beta^0 - X\beta|| = c$, the shape is an ellipsoid. When $X^T X = I$ (orthonormal design matrix), then the ellipsoid contours are spherical. Otherwise, it is an ellipsoid stretched based on the eigenvalues of $X^T X$. Since $X^T X$ is symmetric, it has an orthonormal eigenbasis. Consider any vector $v$ such that $vX^T Xv =$ c. Then, represent $v = \sum_{i=1}^{p} \alpha_i v_i$, where $v_i$ is an eigenvector. Then, we have that

$$\begin{aligned}
vX^T Xv &= \\
(\sum_{i=1}^{p} \alpha_i v_i) X^T X (\sum_{i=1}^{p} \alpha_i v_i) &= \\
\sum_{i=1}^{p} \sum_{j=1}^{p} \alpha_i v_i X^T X \alpha_j v_j &= \\
\sum_{i=1}^{p} \sum_{j=1}^{p} \lambda_j \alpha_i \alpha_j v_i v_j &= \\
\sum_{i=1}^{p} \sum_{j=1}^{p} \lambda_j \alpha_i \alpha_j \delta_{ij} &= \\
\sum_{i=1}^{p} \lambda_j \alpha_i^2 &= c \implies \\
\sum_{i=1}^{p} \frac{\alpha_i^2}{\frac{1}{\lambda_j}} &= c
\end{aligned}$$

So, each axis is scaled by $\frac{1}{\sqrt{\lambda_i}}$ (interestingly, the inverse singular values of X).

The LASSO estimate occurs when $||\beta|| \leq t$, so when the ellipsoid intersects the rotated cube. This is likely to occur at a corner (when one or more coefficients are 0), whereas a ridge is not, since it intersects a sphere.

In general, LASSO performs best when there are a small to moderate number of moderate effects, but performs much worse when there are many small effects (ridge does best) or very few large effects (subset selection performs best).

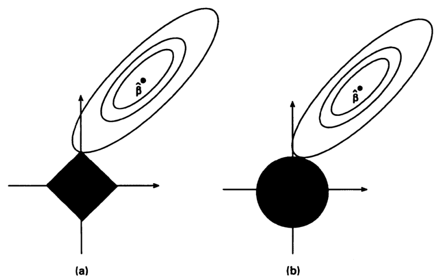

Fig. 2. Estimation picture for (a) the lasso and (b) ridge regression