The Ridge Regression Estimator's Variance

Edric Svarte

Why is ridge regression (RR) such a powerful tool? Let's try to answer this question by reviewing some theory, and considering a quick example.

Estimators

First, what is an *estimator*? An estimator is a random variable. More importantly, an estimator is a "method" used to *estimate* some quantity. Properties of interest include bias and variance. The former is easily understood. In particular, an estimator is unbiased if it is correct, on average. Conversely, the notion of an estimator's variance is not immediately intuitive. However, it may help to remember that an estimator is a function of the sampled data. Due to random sampling, different samples may contain different data points. Consequently, an estimator may yield different estimates when applied to these various samples. If these estimates are highly variable, then the estimator in question has high variance.

Ridge Regression vs. OLS

We are now well-equipped to understand one of the main advantages of ridge regression (RR). That is, the RR estimator's variance is lower than that of the ordinary least squares (OLS) estimator. Why does variance matter? A high-variance model is more likely to overfit the data, and hence RR may outperform OLS on prediction tasks. We'll explore this later. For now, let's consider some theory.

Consider n observations of p variables stored in a design matrix, $X \in \mathbb{R}^{n \times p}$, and a vector of responses, $y \in \mathbb{R}^n$.

Now, recall the OLS estimator:

$$\tilde{\beta}_{OLS} = (X^T X)^{-1} X^T y$$

On the other hand, the RR estimator is given by:

$$\tilde{\beta}_{RR} = (X^T X + \lambda I_p)^{-1} X^T y$$

Here, $\lambda > 0$ is a tuning parameter. If $\lambda = 0$, we recover the OLS estimator. Note that I_p corresponds to the $p \times p$ identity matrix.

The variance of the first estimator is provided below:

$$\mathbb{V}_{Y|X}(\tilde{\beta}_{OLS}) = \sigma^2 (X^T X)^{-1}$$

Now, the RR estimator's variance is given by:

$$\mathbb{V}_{Y|X}(\tilde{\beta}_{RR}) = \sigma^2(X^TX + \lambda I_p)^{-1}X^TX(X^TX + \lambda I_p)^{-1}$$

Note that both $\mathbb{V}_{Y|X}(\tilde{\beta}_{OLS})$ and $\mathbb{V}_{Y|X}(\tilde{\beta}_{RR})$ are $n \times n$ variance-covariance matrices, though we generally use the term "variance" for simplicity.

Now, here is the claim: $\mathbb{V}_{Y|X}(\tilde{\beta}_{RR}) \preceq \mathbb{V}_{Y|X}(\tilde{\beta}_{OLS})$

Let's prove this proposition. First, we know that: $I_p \succ 0$

$$\implies \lambda I_p \succ 0 \implies 2\lambda I_p \succ 0$$

Recall that for any matrix A, A^TA is positive semidefinite. Hence, we have that: $X^TX \succeq 0$

Assuming X^TX is invertible,

$$X^TX \succeq 0 \implies (X^TX)^{-1} \succeq 0 \implies \lambda^2(X^TX)^{-1} \succeq 0$$

Combining both results, we obtain:

$$2\lambda I_p + \lambda^2(X^TX)^{-1} \succeq 0$$

$$\implies X^TX + 2\lambda I_p + \lambda^2(X^TX)^{-1} \succeq X^TX$$

$$\implies X^TX + \lambda I_p + \lambda I_p + \lambda^2(X^TX)^{-1} \succeq X^TX$$

$$\implies (I_p + \lambda(X^TX)^{-1})(X^TX + \lambda I_p) \succeq X^TX$$

$$\implies (X^TX + \lambda I_p)(X^TX)^{-1}(X^TX + \lambda I_p) \succeq X^TX$$

We can apply the following property of invertible matrices A and B: $A \succeq B \iff A^{-1} \preceq B^{-1}$

$$\implies (X^TX + \lambda I_p)^{-1}(X^TX)(X^TX + \lambda I_p)^{-1} \preceq (X^TX)^{-1}$$

$$\implies \sigma^2(X^TX + \lambda I_p)^{-1}(X^TX)(X^TX + \lambda I_p)^{-1} \preceq \sigma^2(X^TX)^{-1}$$

Equivalently,

$$\mathbb{V}_{Y|X}(\tilde{\beta}_{RR}) \preceq \mathbb{V}_{Y|X}(\tilde{\beta}_{OLS})$$

Example - Prediction Tasks

That's fine, but how does this affect our predictions? Let's say we want to predict the response for some (given) query point $x_0 \in \mathbb{R}^p$. The OLS model predicts $\tilde{y}_{OLS} = x_0^T \tilde{\beta}_{OLS}$, while the RR model predicts $\tilde{y}_{RR} = x_0^T \tilde{\beta}_{RR}$.

The OLS and RR estimators are random variables, and thus so are the predictions. Now, what can we say about the variance of these predictions?

Using the "inequality" from the previous section, we can obtain an inequality involving quadratic forms. We left and right "multiply" by x_0, as follows:

$$x_0^T \mathbb{V}_{Y|X}(\tilde{\beta}_{RR}) x_0 \leqslant x_0^T \mathbb{V}_{Y|X}(\tilde{\beta}_{OLS}) x_0$$

The expression above looks strange. However, recall that for a vector v, and a matrix Σ, we have that $\mathbb{V}(v^T \Sigma) = v^T \Sigma v$. Simplifying,

$$\mathbb{V}_{Y|X,x_0}(x_0^T \tilde{\beta}_{RR}) \leqslant \mathbb{V}_{Y|X,x_0}(x_0^T \tilde{\beta}_{OLS})$$

The variance is now *also* conditional on x_0 since this feature vector is fixed by assumption.

$$\implies \mathbb{V}_{Y|X,x_0}(\tilde{y}_{RR}) \leqslant \mathbb{V}_{Y|X,x_0}(\tilde{y}_{OLS})$$

Hence, the RR predictions are less variable than the OLS ones. However, remember that a model's predictive performance is determined by far more than its variance. On the other hand, the inequality above provides some intuition as to why RR is so useful in practice. Indeed, we now have statistical motivation for using this technique.