Bias-Variance Tradeoff

$$
\begin{aligned}
\text{MSE}(\hat{\beta}_j) &= \mathrm{E}\left[(\hat{\beta}_j - \beta_j)^2\right] \\
&= \mathrm{E}\left[(\hat{\beta}_j - \mathrm{E}[\hat{\beta}_j])^2\right] + (\mathrm{E}[\hat{\beta}_j] - \beta_j)^2 \\
&= (\text{Variance of } \hat{\beta}_j) + (\text{Bias of } \hat{\beta}_j)^2
\end{aligned}
$$

- OLS estimates for β_j's are unbiased
- However, the variances of OLS estimates $\hat{\beta}_j$ can be large when
    - the number of predictors is large, or when
    - the predictors are multicollinear
- Is there a way to reduce the variance of $\hat{\beta}_j$, possibly at the cost of increased bias?

Shrinkage Estimates (aka. Regularization)

- OLS estimates $\hat{\beta}_j$ have no upper bound, and hence is susceptible to very high variance
- By **shrinking** the OLS estimates $\hat{\beta}_j$ toward 0, we can often substantially reduce the variance at the cost of a negligible increase in bias, substantially improving the accuracy of prediction for future observations
- **Shrinkage** is called "Regularization" in Machine Learning
- Two common shrinkage estimates are
    - Ridge regression
    - Lasso (Least Absolute Shrinkage and Selection Operator)

OLS v.s. Ridge v.s. Lasso

Ordinary Least Square minimizes:

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_p x_{ip})^2$$

Ridge Regression minimizes:

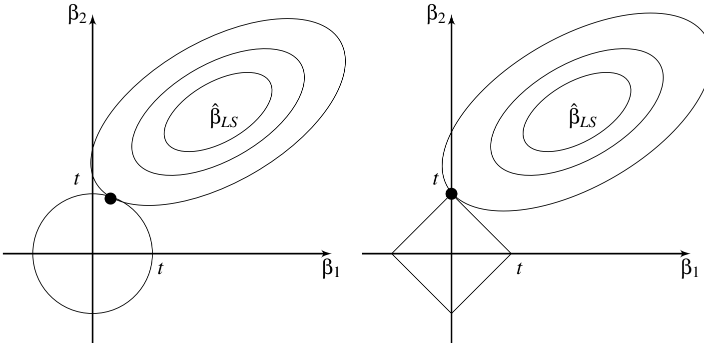
$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_p x_{ip})^2 \quad \text{with the constraint } \sum_{j=1}^{p} \hat{\beta}_j^2 \leq t$$

Lasso mininizes:

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_p x_{ip})^2 \quad \text{with the constraint } \sum_{j=1}^{p} \left|\hat{\beta}_j\right| \leq t$$

Note there is no constraint placed on the magnitude of the intercept $\hat{\beta}_0$.

Geometric Illustration of Ridge and Lasso Estimates



- Ellipses are the contours of $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2$, which centered at the OLS estimates $(\hat{\beta}_{1,OLS}\,,\ \hat{\beta}_{2,OLS}\,)$.
- (Left) Ellipse intersects the circle of radius t at the Ridge estimate.
- (Right) Ellipse intersects the square ($|\hat{\beta}_1| + |\hat{\beta}_2| < t$) at the Lasso estimate

Equivalent Forms of Ridge and Lasso

By the Lagrange multiplier methods, minimizing $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_p x_{ip})^2$ under the constraints

$$\sum_{j=1}^{p} \hat{\beta}_j^2 \leq t \quad \text{or} \quad \sum_{j=1}^{p} \left| \hat{\beta}_j \right| \leq t$$

is equivalent to

Ridge Regression, minimizing

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_p x_{ip})^2 + \lambda \sum_{j=1}^{p} \hat{\beta}_j^2$$

Lasso, minimizing:

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_p x_{ip})^2 + \lambda \sum_{j=1}^{p} \left| \hat{\beta}_j \right|$$

Tuning Parameter λ or t

Both Ridge and Lasso have a **tunning parameter** λ (or t)

- The Ridge estimates $\hat{\beta}_{j,\lambda,Ridge}$'s and Lasso estimates $\hat{\beta}_{j,\lambda,Lasso}$ depend on the value of λ (or t)

λ (or t) is the **shrinkage parameter** that controls the size of the coefficients

- As $\lambda \downarrow 0$ or $t \uparrow \infty$, the Ridge and Lasso estimates become the OLS estimates
- As $\lambda \uparrow \infty$ or $t \downarrow 0$, Ridge and Lasso estimates shrink to 0 (intercept only model)

Ridge and Lasso Estimates Are NOT Scale Invariant

Say we change the unit of a predictor X_j from inches to feet

$$X'_j = X_j/12$$

its coefficient would be scaled as

$$\beta'_j = 12\beta_j$$

so that the product $\beta'_j X'_j = \beta_j X_j$ stays unchanged.

However, the Ridge and Lasso estimates are not scaled accordingly

$$\hat{\beta}'_{j,\lambda,Ridge} \neq 12\hat{\beta}_{j,\lambda,Ridge}, \quad \hat{\beta}'_{j,\lambda,Lasso} \neq 12\hat{\beta}_{j,\lambda,Lasso}$$

since large β's are penalized

Must Standardize Predictors Before Applying Ridge and Lasso

As Ridge and Lasso estimates are not scale invariant, by convention, we **standardize** all predictors

$$Z_j = \frac{X_j - \overline{X}_j}{s_j}, \quad j = 1, \ldots, p,$$

where s_j is the sample SD of X_j. before applying Ridge and Lasso.

That is, all predictors X_j's in Ridge and Lasso regression are assumed to have mean 0 and variance 1.

Ridge Estimates Are Biased but Have Smaller Variance

- Recall OLS estimate for $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$
- One can show Ridge estimate for $\boldsymbol{\beta}$ is $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{Y}$
  - Keep in mind that $\mathbf{X}$ is standardized
    that each predictor has mean 0 and variance 1
- Expected value for the Ridge estimate for $\boldsymbol{\beta}$ can be shown to be

$$(\mathbf{I}_p + \lambda\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{\beta} \neq \boldsymbol{\beta}$$

- If all predictors are standardized and uncorrelated,

$$\hat{\beta}_{j,\lambda,Ridge} = \frac{1}{1+\lambda}\hat{\beta}_{j,OLS}$$

- Smaller variance than OLS estimates,
- Variance of $\hat{\beta}_{j,\lambda,Ridge}$ is much smaller than $\hat{\beta}_{j,OLS}$ when the data have **multicollinearity** problem

Properties of Lasso Estimates

- No close form formula for the Lasso estimates
- Also biased (toward 0)
- Smaller variance than OLS estimates
- NOT perform as well as Ridge when data have **multicollinearity** problem
- Greatest advantage of Lasso: **Sparsity** (See next page)

Sparsity of Lasso Estimates

- In a model with many predictors

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

  we may believe many of the β_j's are actually 0.
- Hence, we seek a set of sparse solutions
- Lasso estimates will set some coefficients exactly equal to 0 when λ is large (or when t is small)

So the LASSO will perform model selection for us!

How to Choose λ?

- We need a disciplined way of choosing λ
- Obviously want to choose λ that minimizes the mean squared error
- Issue is part of the bigger problem of **variable selection**

Choosing λ Using Cross-Validation

- If we have a good model, it should predict well when we have new data
- Data are hence split into 2 parts — **training data** and **test data**
- For each λ, use the training set to fit (train) a model and than use the model to predict values in the test set and compute the rooted mean square error (RMSE)

$$\sqrt{\sum_{\text{test data}} (y_i - \hat{y}_i)^2 / n}, \quad \text{where } n = \text{size of the test data}$$

- Choose the λ that has the smallest RMSE
- The training set and test set should be chosen randomly
  - May split the whole data into several different training set and test set and compute the mean of the RMSE for different splits