RECSM Summer School: Machine Learning for Social Sciences

Session 1.5: The Lasso

Reto Wüest

Department of Political Science and International Relations
University of Geneva

Outline

1. The Lasso
2. Comparing the Lasso and Ridge Regression
3. Selection of the Tuning Parameter

The Lasso

The Lasso

- A disadvantage of ridge regression is that it will always include all p predictors in the model.
- The ridge regression penalty $\lambda \sum_{j=1}^{p} \beta_j^2$ shrinks all coefficients towards 0, but it does not set any of them exactly to 0.
- The lasso overcomes this disadvantage by replacing the β_j^2 term in the ridge regression penalty by $|\beta_j|$.

The Lasso

- Therefore, the lasso coefficient estimates are the values that minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \text{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|. \quad (1.5.1)$$

- As with ridge regression, the lasso shrinks the estimates towards 0.
- However, when λ is sufficiently large, the lasso forces some estimates to be exactly equal to 0 (the lasso thus performs variable selection).

The Lasso

- As in ridge regression, the tuning parameter λ plays a critical role:
  - If $\lambda = 0$, then the lasso estimates are identical to the least squares estimates.
  - When λ becomes sufficiently large, the lasso estimates are set exactly equal to 0.
- Depending on the value of λ, the lasso can produce a model involving any number of variables.
- In contrast, ridge regression will always include all of the variables in the model.

Comparing the Lasso and Ridge Regression

Comparing the Lasso and Ridge Regression

- The lasso coefficient estimates solve the problem

$$\underset{\beta}{\arg\min}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\} \quad \text{s.t.} \quad \sum_{j=1}^{p}|\beta_j| \leq s. \quad (1.5.2)$$

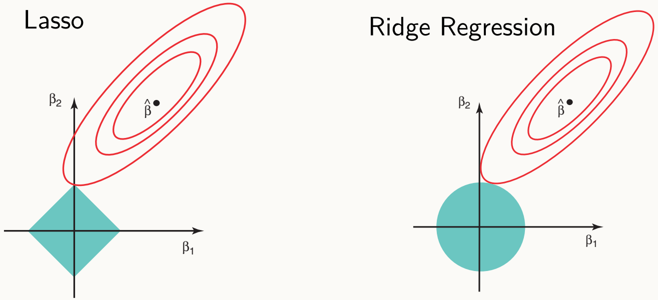
- The ridge regression coefficient estimates solve the problem

$$\underset{\beta}{\arg\min}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\} \quad \text{s.t.} \quad \sum_{j=1}^{p}\beta_j^2 \leq s. \quad (1.5.3)$$

Comparing the Lasso and Ridge Regression

- If $p = 2$, lasso tries to find the set of coefficient estimates that lead to the smallest RSS, subject to the budget constraint $|\beta_1| + |\beta_2| \leq s$.
- If $p = 2$, ridge regression tries to find the set of coefficient estimates that lead to the smallest RSS, subject to the budget constraint $\beta_1^2 + \beta_2^2 \leq s$.

Comparing the Lasso and Ridge Regression



(Source: James et al. 2013, 222)

- $\hat{\beta}$ is the least squares solution.
- The diamond and the circle are the lasso and ridge regression constraints, respectively.
- The ellipses are the set of estimates with a constant RSS. Those farther away from the least squares coefficient estimates have a larger RSS.

Comparing the Lasso and Ridge Regression

- The lasso has the advantage of producing simpler, and therefore more interpretable, models than ridge regression.
- However, which method leads to better prediction accuracy?
- Neither the lasso nor ridge regression will universally dominate the other.
  - The lasso tends to perform better when only a relatively small number of predictors have substantial coefficients.
  - Ridge regression tends to perform better when there are many predictors, all with coefficients of roughly equal size.

Selection of the Tuning Parameter

Selection of the Tuning Parameter

- Ridge regression and the lasso require us to select a value for the tuning parameter λ.
- How do we choose the optimal λ?
- Cross-validation provides a way to tackle this problem:
  - Choose a grid of λ values and compute the CV error for each value.
  - Select the tuning parameter value for which the CV error is smallest.
  - Re-fit the model using all available observations and the selected λ value.