Computational Statistics

26.04 and 27.04: Model selection
Methods for model selection

• Three classes:
  • Subset selection
  • Shrinkage
  • Dimension reduction (i.e. PCA and then)

• We will only discuss subset selection and shrinkage in more detail
Subset selection: all subsets

**Algorithm 6.1 Best subset selection**

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.

2. For $k = 1, 2, \ldots p$:
   
   (a) Fit all $\binom{p}{k}$ models that contain exactly $k$ predictors.

   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest $R^2$.

3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, $C_p$ (AIC), BIC, or adjusted $R^2$. 
**FIGURE 6.1.** For each possible model containing a subset of the ten predictors in the Credit data set, the RSS and $R^2$ are displayed. The red frontier tracks the best model for a given number of predictors, according to RSS and $R^2$. Though the data set contains only ten predictors, the x-axis ranges from 1 to 11, since one of the variables is categorical and takes on three values, leading to the creation of two dummy variables.
Subset selection: forward stepwise

**Algorithm 6.2** Forward stepwise selection

1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors.

2. For $k = 0, \ldots, p - 1$:
   
   (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.

   (b) Choose the best among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here best is defined as having smallest RSS or highest $R^2$.

3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, $C_p$ (AIC), BIC, or adjusted $R^2$. 
Subset selection: stepwise backward

**Algorithm 6.3 Backward stepwise selection**

1. Let $\mathcal{M}_p$ denote the **full** model, which contains all $p$ predictors.

2. For $k = p, p - 1, \ldots, 1$:
   
   (a) Consider all $k$ models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k - 1$ predictors.

   (b) Choose the best among these $k$ models, and call it $\mathcal{M}_{k-1}$. Here best is defined as having smallest RSS or highest $R^2$.

3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, $C_p$ (AIC), BIC, or adjusted $R^2$. 
Criteria for model selection

• Criteria involve trade-off between model fit and model complexity:
  • Mallows’s $C_p$
  • Akaike Information Criterion (AIC)
  • Bayesian Information Criterion (BIC)
  • Adjusted $R^2$

• Example: $BIC = -2 \ln(\hat{L}) + \log(n)d$, where $\hat{L}$ is the maximized value of the likelihood in the model, $n$ is the sample size, and $d$ is the number of parameters in the model
**FIGURE 6.2.** $C_p$, BIC, and adjusted $R^2$ are shown for the best models of each size for the Credit data set (the lower frontier in Figure 6.1). $C_p$ and BIC are estimates of test MSE. In the middle plot we see that the BIC estimate of test error shows an increase after four variables are selected. The other two plots are rather flat after four variables are included.
Shrinkage methods

• Fit model with all parameters, but “shrink” parameters towards zero:
  • Ridge regression
  • Lasso
• See board
Ridge regression

**FIGURE 6.4.** The standardized ridge regression coefficients are displayed for the Credit data set, as a function of $\lambda$ and $\|\hat{\beta}_x^R\|_2/\|\hat{\beta}\|_2$. 
FIGURE 6.5. Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of $\lambda$ and $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.
Lasso regression

**FIGURE 6.6.** The standardized lasso coefficients on the Credit data set are shown as a function of $\lambda$ and $\|\hat{\beta}_X\|_1/\|\hat{\beta}\|_1$. 
Comparing ridge and lasso

**FIGURE 6.7.** Contours of the error and constraint functions for the lasso (left) and ridge regression (right). The solid blue areas are the constraint regions, $|\beta_1| + |\beta_2| \leq s$ and $\beta_1^2 + \beta_2^2 \leq s$, while the red ellipses are the contours of the RSS.
Comparing ridge and lasso

All 45 out of 45 predictors are active (i.e., $\beta_j \neq 0$ for all $j = 1, \ldots, p$)

**FIGURE 6.8.** Left: Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso on a simulated data set. Right: Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dotted). Both are plotted against their $R^2$ on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.
Comparing ridge and lasso

FIGURE 6.9. Left: Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso. The simulated data is similar to that in Figure 6.8, except that now only two predictors are related to the response. Right: Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dotted). Both are plotted against their $R^2$ on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.
Literature

- Figures and pseudocode have been taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani"