4. Linear Model Selection and Regularization

Jesper Armouti-Hansen

University of Cologne

January 14, 2019

Course website

jeshan49.github.io/eemp2/

Today

- Lecture[1]:
  - Subset Selection
  - Shrinkage/Regularization
  - Dimension Reduction
- Tutorial:
  - Reproducing results from the lecture using:
    - Forward/Backward Subset Selection
    - Ridge and Lasso Regression
    - Principal Component and Partial Least Squares Regression

[1] Some of the figures in this presentation are 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

Introduction

- In the regression setting, the standard linear model

$$Y = \beta_0 + \beta_1 X_1 + \dots \beta_p X_p + \varepsilon \tag{1}$$

  is commonly used to describe the relationship between the response and the input

- This linear model has an obvious advantage compared to non-linear methods in terms of *model interpretability*

- In addition, it is surprisingly competitive in relation to non-linear methods in many settings in terms of *prediction accuracy*

- Today, we will discuss alternative fitting strategies to least squares that may improve the fit

Why use alternative fitting strategies over least squares?

Let us first consider *prediction accuracy*:

- Recall the bias-variance trade-off: In general, too simple (complex) models with have high (low) bias and low (high) variance
- Suppose now that the true relationship between the input and output is approx. linear
- Then, our linear model in (1) will have low bias. In addition, if $N >> p$, it will have low variance as well
- However, if $N > p$, there is a lot of variability in the fit, and hence high variance. In addition, if $N < p$, this variance is infinite
- By constraining or shrinking the coefficients, we can reduce the variance substantially at the cost of a small increase in bias

Let us now consider *model interpretability*:

- Often some or many of the p predictors are not associated with the response
- Including these predictors leads to unnecessary in the resulting model
- This is because the least square fit is extemely unlikely to yield exact zero coefficients
- By setting some of the coefficients to zero, we obtain a more easily intepretable model
- We will thus consider methods that automatically perform feature selection

Three alternative methods

Today we'll discuss three alternative methods:

1. Subset selection:
   - Identify a subset of the p predictors that we believe are related to Y. Then we fit using least squares on this subset
2. Shrinkage:
   - Fit on all p predictors using least squares subject to a constraint on the size of the coefficients
   - This shrinkage/regularization reduces the variance
3. Dimension reduction:
   - Projecting the p predictors into a M-dimensional subspace, where $M < p$
   - We then fit the model with the M predictors using least squares

Best subset selection

Algorithm 1 Best subset selection

```
Let $\mathcal{M}_0$ denote the null model, which contains no predictors.
**for** $k = 1$ to $p$ **do**
    (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 the best is defined as having the smallest RSS or highest $R^2$.
**end for**
Select a single best model from $\mathcal{M}_0, \dots, \mathcal{M}_p$ using CV, $C_p, AIC, BIC$ or Adj. $R^2$.
```

Example

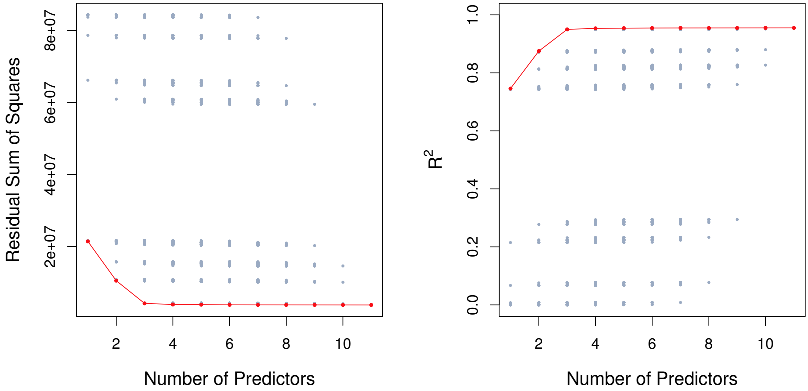

Figure: For each possible model containing a subset of ten predictors in the Credit data set, the RSS and R^2 are displayed. The red frontier tracks the best model for a given number of predictors. (See ISLR p. 206)

Some notes on best subset selection

- The same idea of best subset selection can be applied to a wide array of models, e.g. logistic regression
- While the method is simple, it suffers from computational limitations:
    - If $p = 10$ we must fit $2^{10} = 1,024$ models
    - If $p = 20$, we must fit $2^{20} = 1,048,576$ models
- Thus, best subset selection becomes unfeasible for $p > 40$
- In addition, the method may suffer from overfitting and high variance of coefficient estimates for large p
- We will now consider more computational efficient compromises, with a smaller search space - *Stepwise selection*

Forward stepwise selection

Algorithm 2 Forward stepwise selection

```
Let $\mathcal{M}_0$ denote the null model, which contains no predictors.
for $k = 0$ to $p - 1$ do
    (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
    (b) Pick the best among these $p-k$ models, and call it $\mathcal{M}_{k+1}$. Here the best is defined as having the smallest RSS or highest $R^2$.
end for
Select a single best model from $\mathcal{M}_0, \dots, \mathcal{M}_p$ using CV, $C_p, AIC, BIC$ or Adj. $R^2$.
```

Some notes on forward stepwise selection

- Forward stepwise selection has a clear computational advantage over best subset selection:
  - For $p = 20$, the latter fits $2^{20} = 1,048,576$ models whereas the former only fits $1 + \sum_{k=0}^{20-1}(20 - k) = 1 + 20(20 + 1)/2 = 211$ models
- Furthermore, it may perform better due to its smaller search space
- However, it may fail to selection the best model
  - Suppose that $p = 3$ and the best model is a two-variable model with X_2, X_3
  - If the best one-variable model is with X_1, then forward stepwise selection will fail in finding the best model

Backward stepwise selection

Algorithm 3 Backward stepwise selection

```
Let $\mathcal{M}_p$ denote the full model, which contains all $p$ predictors.
for $k = p$ to 1 do
    (a) Consider all $k$ models contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k - 1$ predictors.
    (b) Pick the best among these $k$ models, and call it $\mathcal{M}_{k-1}$. Here the best is defined as having the smallest RSS or highest $R^2$.
end for
Select a single best model from $\mathcal{M}_0, \dots, \mathcal{M}_p$ using CV, $C_p, AIC, BIC$ or Adj. $R^2$.
```

Some notes on backward stepwise selection

- Like forward stepwise selection, backward stepwise selection only fits $1 + p(p+1)/2$ models
    - Thus, it can be used with large p
- However, once again, it is not guaranteed to select the best model containing a subset of p predictors
- Furthermore, backward stepwise selection can only be used in settings in which $N > p$
- In contrast, we can always use forward selection up to a particular number of predictors

Choosing the optimal model

- Each preceding method yield an optimal model for $1, \ldots, p$
- Thus, at the end, we need to select one best model from these candidates
- We do not want to use RSS nor R^2 as they are directly related to the training error
  - as we know, the training error can be a poor estimate of the test error
- Thus, we consider two alternative approaches:
  1. Indirectly estimate test error by making an adjustment to the training error
  2. Directly estimate the test error by using the validation set or the cross-validation approach

C_p, AIC, BIC, and Adj. R^2

- Let d be the # of predictors and $\hat{\sigma}^2 = RSS/(N-p-1)$ an estimate of $Var[\varepsilon]$ from the full model

Mallow's C_p:

$$C_p = \frac{1}{N}(RSS + 2d\hat{\sigma}^2) \tag{2}$$

Akaike information criterion:

$$AIC = \frac{1}{N\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2) = \frac{1}{\hat{\sigma}^2} * C_p \tag{3}$$

Bayesian information criterion:

$$BIC = \frac{1}{N\hat{\sigma}^2}(RSS + log(N)d\hat{\sigma}^2) \tag{4}$$

Adjusted R^2:

$$\text{Adj. } R^2 = 1 - \frac{RSS/(N-d-1)}{TSS/(N-1)} \tag{5}$$

Example (Best subset selection)

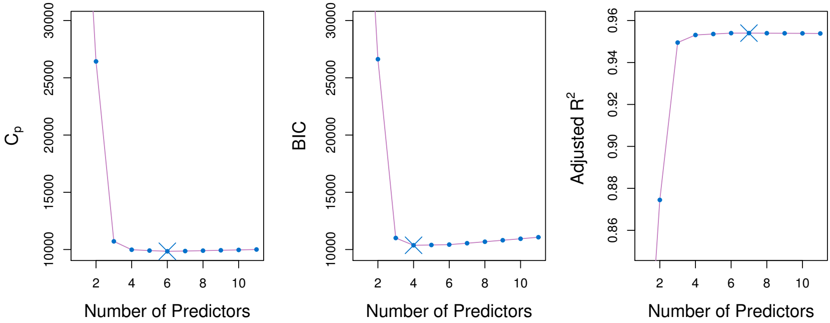

Figure: C_p, BIC, and Adj. R^2 for the best models of each size of the Credit data set. (See ISLR p. 211)

- C_p: *income, limit, rating, cards, age, student*
- BIC: *income, limit, cards, student*
- Adj. R^2: *income, limit, rating, cards, age, student, gender*

Validation and Cross-Validation

- As we already know, we can also use the validation set approach or k-fold CV for the task of model selection

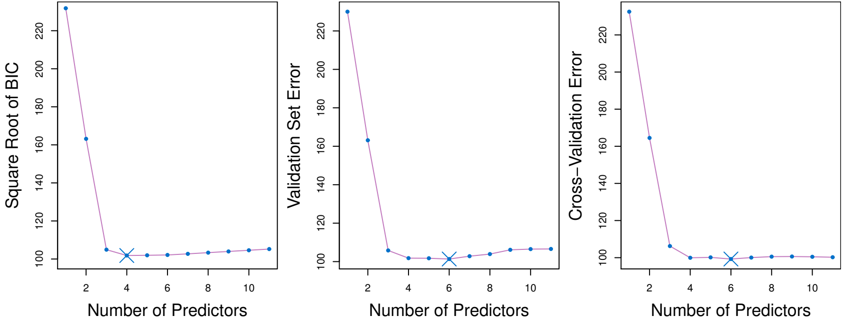

Figure: Credit data set. Left: Square root of BIC, Center: Validation set errors, Right: 10-fold CV errors (See ISLR p. 214)

Shrinkage methods - Ridge Regression

- As an alternative to subset selection, we can fit a model on all p predictors, whilst shrinking the coefficients toward zero
- We will see that this can reduce the variance of our model

Ridge regression solves:

$$\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j^2 x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \quad (6)$$

where $\lambda \geq 0$ is a tuning/hyper parameter; it controls the magnitude of regularization vs. fit

- We find an estimate $\hat{\beta}_{\lambda}^{R}$ for many λ, and then choose the optimal λ by CV - This is not computational expensive

Notes on Ridge regression

- Shrinkage is applied to $\beta_1, \ldots, \beta_p$ but not to β_0
  - This is because we want to shrink the estimated association of each variable with the response

- Standard least square coefficient estimates are scale invariant
  - $X_j \hat{\beta}_j$ will remain the same

- Ridge coefficient estimates can change substantially
  - $X_j \hat{\beta}^R_{j,\lambda}$ may not only depend on λ and its predictor's scale, but also on other predictors' scale

- Thus, it is best to apply ridge regression after standardizing the predictors

Example

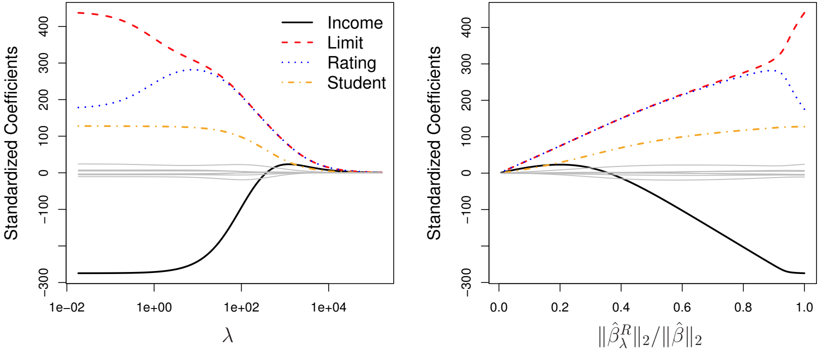

Figure: The standardized ridge regression coefficients are displayed for the Credit data set, as a function of λ and $||\hat{\beta}^R_\lambda||_2/||\hat{\beta}||_2$ (See ISLR p. 216)

Why does Ridge regression improve over Least squares?

- As λ increases, the flexibility of the fit decreases, leading to decreased variance but increased bias

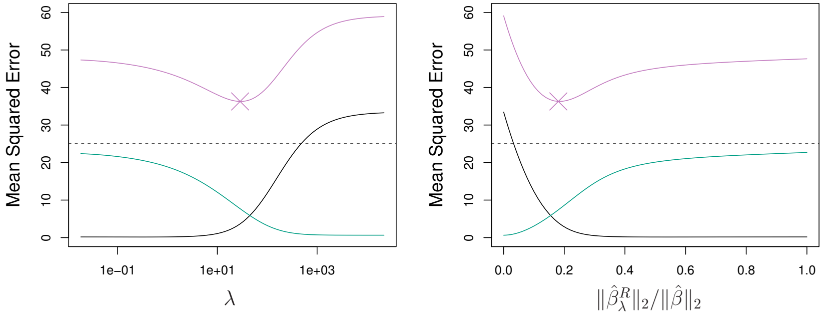

Figure: Squared bias (black), variance (green), and test MSE (purple) for the ridge regression predictions on a simulated data set, as a function of λ and $||\hat{\beta}_\lambda^R||_2/||\hat{\beta}||_2$. The dashed line indicates minimum MSE (See ISLR p. 218)

Shrinkage methods - Lasso

- Ridge regression has one obvious limitation: None of the coefficient estimates will be exactly zero, unless $\lambda = \infty$

Lasso solves:

$$\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j^2 x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = RSS + \lambda \sum_{j=1}^{p} |\beta_j| \quad (7)$$

- The penalty term of the lasso has the effect of setting coefficient estimates exactly zero for finite λ
- We say that the lasso yields sparse models – models that involve only a subset of the variables
- We find an estimate $\hat{\beta}_{\lambda}^{L}$ for many λ, and then choose the optimal λ by CV

Example

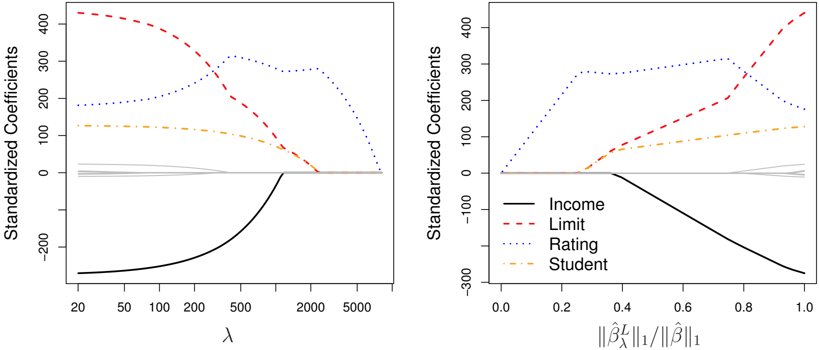

Figure: The standardized lasso coefficients are displayed for the Credit data set, as a function of λ and $||\hat{\beta}^L_\lambda||_1/||\hat{\beta}||_1$ (See ISLR p. 220)

Another representation for Ridge regression and Lasso

- One can show that (i) *best subset selection*, (ii) *ridge regression*, and (iii) *the lasso* can be formulated as follows

Best subset selection:

$$\min_{\beta}\left\{\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j^2x_{ij}\right)^2\right\}\text{ s.t. }\sum_{j=1}^{p}\mathbb{I}(\beta_j\neq 0)\leq s \quad (8)$$

Ridge regression:

$$\min_{\beta}\left\{\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j^2x_{ij}\right)^2\right\}\text{ s.t. }\sum_{j=1}^{p}\beta_j^2\leq s \quad (9)$$

Lasso:

$$\min_{\beta}\left\{\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j^2x_{ij}\right)^2\right\}\text{ s.t. }\sum_{j=1}^{p}|\beta_j|\leq s \quad (10)$$

The variable selection property of Lasso

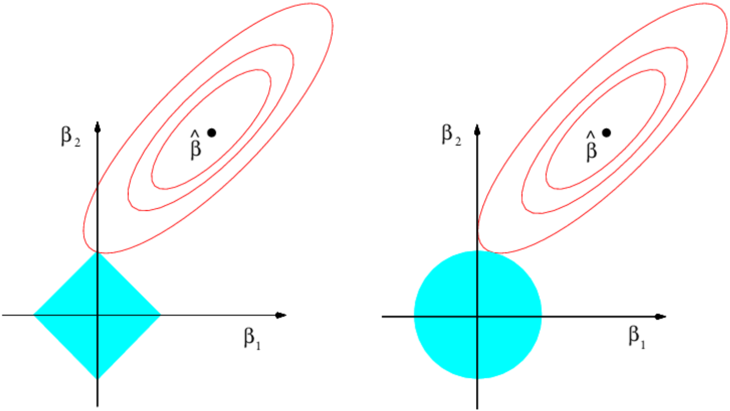

Figure: Contours of the error and constraint functions for the lasso (left) and ridge regression (right). The solid blue areas are the constraint regions, $|\beta_1| + |\beta_2| \leq s$ and $\beta_1^2 + \beta_2^2 \leq s$ (See ISLR p. 222)

Comparing Ridge regression and Lasso

- Lasso has an advantage over ridge regression in terms of model interpretability
- But what about prediction accuracy?
- Lasso implicitly assumes that some of the predictors are unrelated to the response
- If this assumption holds, then the lasso can perform better. If not, then ridge regression will in general perform better
- It is possible to combine both approaches in one method
    - ElasticNet regression: a convex combination of ridge regression and lasso

Dimension reduction methods

- Methods that transform the p predictors onto M dimensions and then fit a least square model on the transformed variables
- Let $Z_1, \dots Z_M$ represent $M < p$ linear combinations of the p predictors

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j \quad \forall m \tag{11}$$

  for some constants $\phi_{1m}, \dots, \phi_{pm}$

- We then fit the linear regression model by least squares:

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \varepsilon_i, i = 1, \dots, n \tag{12}$$

- With this approach, we are reducing the dimension of the problem from $p + 1$ to $M + 1$

Principal Component Regression (PCR)

- To perform PCR, we apply principal component analysis (PCA)
  - PCA is a technique for reducing the dimension of our data
- Our goal: Ending up with $M < p$ principal components which summarizes the majority of the variation in our data
- **Assumption**: The direction in which $X_1, \ldots, X_p$ show the most variation are the directions that are associated with the response
- 1st PC (Z_1): The linear combination of predictors with the largest variance
  - Or: the line that is as close as possible to the data
- $i + 1$th PC (Z_{i+1}): The linear combination of predictors with the largest variance subject to being uncorrelated with the ith PC
  - Or: The line that is as close as possible to the data subject to being orthogonal to the ith PC

Example

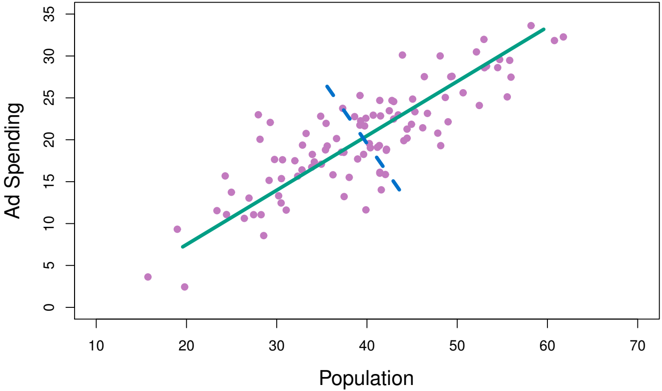

Figure: The population size and ad spending for 100 different cities are shown as purple circles. First PC (green), second PC (blue, dashed) (See ISLR p. 230)

Example (cont'd)

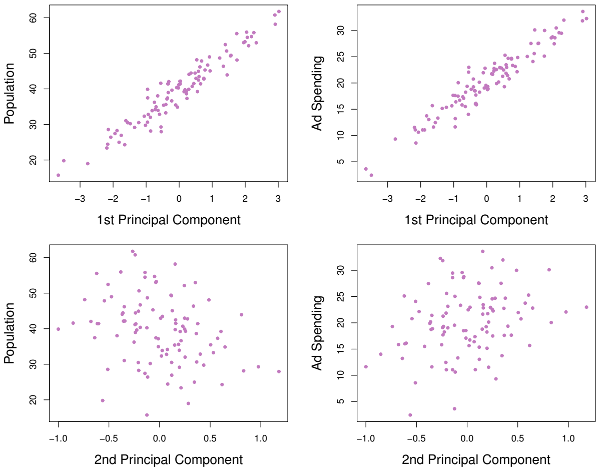

Figure: Plots of the first (top) and second (bottom) PC scores vs. population (left) and ad spending (right) (See ISLR pp. 233-234)

Partial Least Squares (PLS)

- PCR identifies $Z_1, \dots, Z_M$ in an unsupervised way – i.e. without considering the response

- Thus, it may be that the directions/components are not the best predictors of the response

- Unlike PCR, PLS defines $Z_1, \dots, Z_M$ in a supervised way:

- PLS computes Z_1 by defining each ϕ_{j1} from (11) to be the coefficient from the simple linear regression of Y onto X_j
  - The coefficient is proportional to the correlation
  - Thus, PLS places most weight on variables that are strongly correlated with the response

- Subsequent directions are found by taking residuals and repeating the process

Notes on PCR and PLS

- With both PCR and PLS, we standardize the predictors before applying the methods
- With both PCR and PLS, we generally locate the optimal number of directions by cross-validation
- In general, the supervised dimension reduction by PLS can reduce bias
- However, this can come at the cost of an increase in variance
- PCR, PLS and ridge regression performs in practice similar in terms of prediction accuracy

References

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). **Chapter 6**