Paul Martins

Linear Model Selection and Regularization

Notes from ISLR book

Contents

Introduction

Ordinary least squares regression has low bias and provided $n \gg p$, low variance too. But if $n \sim p$ or $n < p$, there will be more variability (reaching infinity in the second case!) which will affect the prediction accuracy. Such high number of predictors also reduce the model interpretability. The solution is to constrain or shrink the estimated coefficient and/or to perform some feature selection.

1. Subset selection

1.1 Best subset selection

```
1 Start by fitting a model with no predictor
2 For $k$ in 1,2,...,$p$
    2.1 Fit all $\binom{p}{k}$ models that contain exactly $k$ predictors
    2.2 Pick the best among the $\binom{p}{k}$ models, $\mathscr{M}_k$ using $RSS$, $R^2$ or deviance[1]
3 Select the best of the $\mathscr{M}_k$ using cross-validated criterion ($C_p$, $AIC$, $BIC$, adjusted $R^2$)
```

Note that RSS or R^2 will always have smaller training error as p increases. Hence we use a different set of metrics in step 3. This method is computationally heavy: if $p = 20$, there are more than 1.10^6 models to try (2^p).

1.2 Stepwise selection

Stepwise selection can be forward or backward depending of the starting point. If we start with no predictor and continually increase the number of predictor considered, we go forward. If we start with all p predictors and progressively remove the least significant ones, we go backward.

```
1 Start with $\mathscr{M}_0$ that has no predictor
2 For $k$ in 1,2,...,$p$
    2.1 Consier all $p-k$ models that augment the predictors in $\mathscr{M}_k$ with one additional predictor
    2.2 Choose the best among the $p-k$ models called $\mathscr{M}_{k+1}$ (choose with $RSS$ or $R^2$)
3 Select the best among all the $\mathscr{M}_0, ..., \mathscr{M}_p$ with $C_p$, $AIC$,...
```

This method has the advantage of only scan $1 + p(p+1)/2$ (211 model for $p = 20$). Unlike backward, forward stepwise selection can be applied even when $p > n$, although for both, the best set of predictor is not guarantee. Hybrid method that try to benefit from best subset and stepwise can be performed where at each iteration, a a predictor is added and another can be removed.

1.3 Selection criterion

To select the mode, one can use a criterion or use cross-validated test error.

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

where d is the number of predictor and $\hat{\sigma}$ the estimate of the variance of the error ε. The Akaike information criteria is often used for model fit with maximum likelihood.

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

Similarly, The Bayesian information criterion is quite similar

$$BIC = \frac{1}{n\hat{\sigma}^2}(RSS + log(n)d\hat{\sigma}^2) \tag{3}$$

All of those have theoretical justifications and we should select low value of those. Alternatively, the adjusted R^2 tries to penalise additional noisy predictors and the higher the better.

$$\text{adjusted } R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)} \tag{4}$$

[1] equivalent of RSS for maximum-likelihood fit, -2×max(log-likelihood)

2. Shrinkage

2.1 Ridge

Ridge is similar to OLS, except the coefficients $\hat{\beta}^R$ are estimated by minimizing

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{5}$$

where the λ term is a shrinkage ℓ_2 penalty that decreases when the β_j are close to 0.

- $\lambda = 0$ corresponds to OLS
- $\lambda \to \infty$ corresponds to $\hat{\beta}_j \to 0$, reduction of variance and increase in bias.

Note that now the coefficients are not scale equivariant anymore, hence it is best to standardise the predictors before performing a ridge regression.

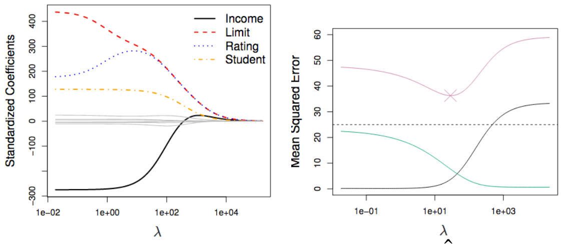

Figure 1. Left: values of standardised $\hat{\beta}_j$ as a function of λ. Right: square bias (black), variance (green) and test mean error (purple) for ridge regression.

2.2 LASSO

Ridge regression always considers all p predictors, even if their coefficients get close to 0. The lasso overcomes this because the coefficients minimise

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j| \tag{6}$$

which uses a ℓ_1 penalty that can set coefficients $\hat{\beta}_j = 0$ for a large enough λ. Hence it can be used fo variable selection. Cross-validation error can be used to select the value of λ.

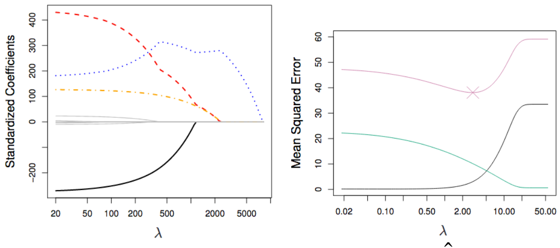

Figure 2. Left: values of standardised $\hat{\beta}_j$ as a function of λ. Right: square bias (black), variance (green) and test mean error (purple) for lasso regression.

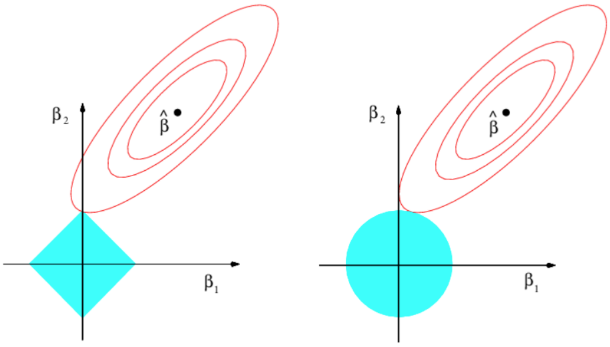

Figure 3. Contour of the RSS (red) and constraint functions for LASSO ($|\beta_1| + |\beta_2| \leq s$) and ridge ($\beta_1^2 + \beta_2^2 \leq s$)

3. Dimension reduction

Let $Z_1,...,Z_M$ represent $M < p$ linear combinations of the predictors $X_1,...,X_p$

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j \tag{7}$$

The regression model becomes

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \varepsilon_i \tag{8}$$

where

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm} \tag{9}$$

The coefficients are now constrained to be of the form of 9.

- $M \ll p$ reduces the variance of the coefficiants
- $M = p$ is equivalent to OLS

3.1 Principal components regression (PCR)

The first principal component Z_1 is the direction along which the data vary the most. The second principal component Z_2 needs to be orthogonal to Z_1. PCR consists in a least square regression on the first M principal components Z_M, hence it assumes that the directions $Z_1,...,Z_M$[2] in which the predictors $X_1,...,X_p$ vary the most are the directions that are associated with Y. It is not a subset selection since linear combination of all predictors are used to build the principal components. Similar to λ in ridge and lasso cases, M can be determined with cross-validation error and the predictors X_p need to be standardised before finding the Z.

3.2 Partial least squares (PLS)

In PCR, there is no guarantee that the Z that best explain the predictors X also best explain the response Y. Instead PLS is supervised and assigns

[2] found in an unsupervised way

$Z_1 = \sum_{j=1}^{p} \phi_{j1} X_j$ where ϕ_{j1} correspond to the coefficient of the regression of Y against X_j. Hence the highest weights are on the variables that are strongly related to the response. To calculate Z_2, we use the residuals from the regression of Y on Z_1 and process as for Z_1, and iterate until Z_M. While

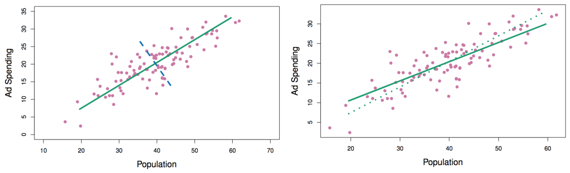

Figure 4. Left: Z_1 and Z_2 directions of PCA. Right: Z_1 direction of PLS (solid) and PCA (dotted).

PLS can reduce bias, it can increase variance.

3.3 High-dimensional data

When $p \sim n$ or $p > n$, regressions are not appropriate, as shown in Figure 5, it's too flexible. Moreover C_p, AIC, BIC, and adjusted R^2 are not good in this case as the estimate of $\hat{\sigma} = 0$.

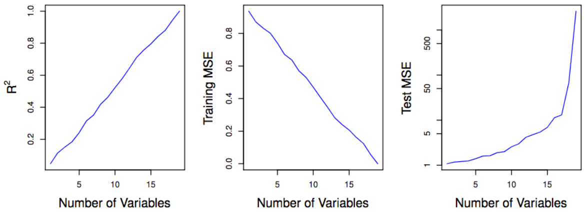

Figure 5. Least regression metrics with $n = 20$ observations and increasing number of predictors that are unrelated to the response