

STATS 211: Statistical Methods II

Lecture 3: Controlling complexity

Babak Shahbaba

UCI, Winter 2012

Why we control complexity

- Overly complex models tend to overfit the data.
- That is, while they have good performance on the observed data (which we use for training), they tend to have poor performance on out-of-sample (e.g., future observations) data.
- In this lecture, we mainly focus on the model complexity related to the number of variables included in the model.
- First, we discuss methods that use few derived variables instead of using a large number of original variable.
- Next, we discuss methods that control the number of variables and magnitude of their effects by penalizing against complexity.

Principal component regression

- Consider the centered matrix of covariates, x , without the first column of 1's.
- The principal components are a set of orthogonal bases, v_1, v_2, \dots, v_p , in the column space of x such that the vectors have unit length, and they are found as follows:
 - v_1 is the basis with the largest sample variance.
 - v_2 is the basis with the second largest sample variance, and it is orthogonal to v_1 .
 - v_j is the basis with the j^{th} largest sample variance, and it is orthogonal to v_1, \dots, v_{j-1} .
- We find the principal component vectors by first finding the eigenvectors of the covariance matrix $s = x^T x / n$, and then by ordering the eigenvectors based on the descending order of their eigenvalues, λ_j , where

$$s v_j = \lambda_j v_j, \quad j = 1, \dots, p$$

and $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$.

Principal component regression

- We then obtain a new set of covariates, $z = (z_1, \dots, z_p)$ by projecting the original observed data, x , on the principal components

$$Z = XV$$

- The columns of z are known as *scores*, and the columns of v are called *loadings*.
- Note that the sample variance of z_j , i.e., the j^{th} column of z , is

$$\text{Var}(z_j) = z_j^T z_j / n = v_j^T x^T x v_j / n = v_j^T S v_j = \lambda_j$$

Principal component regression

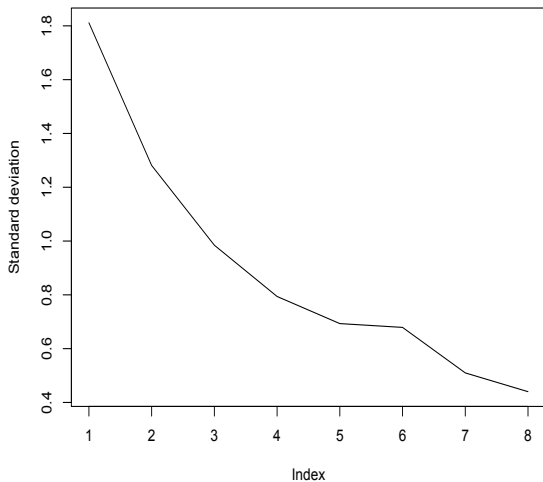
- After finding the scores, z , we define a new set of derived covariates using the first q columns of z .
- To choose q , it is common to use the corresponding *scree* plot, which is the plot of all the eigenvalues in their decreasing order.
- Principal component regression (PCR) is a linear regression model that uses z_1, \dots, z_q as covariates instead of the original x_1, \dots, x_p variables

$$y = \gamma_0 + \gamma_1 z_1 + \dots + \gamma_q z_q + \eta$$

where η is the random noise.

- PCR works well when the variation of y mainly occurs along the directions of high variance in the space of covariates.

Scree plot for prostate cancer data



Partial least squares

- Instead of defining the new set of bases according to the covariates only, we could define them with respect to the relationship between covariates and the response variable.
- Suppose the inputs are standardized to have mean 0 and standard deviation 1. The partial least squares (PLS) method performs this task as follows:
 1. Starts by finding the univariate regression coefficient $\hat{\phi}_{1j}$ of y on each x_j .
 2. Obtain the first derived input $z_1 = \sum_{i=1}^p \hat{\phi}_{1j} x_j$. This is the first PLS direction.
 3. Orthogonalize the original inputs with respect to this direction by subtracting from each x_j its projection in the direction of z_1 .
 4. We repeat the above procedure to obtain z_2 up to z_q , where $q < p$.
 5. We regress y on the new derived variables z_1, \dots, z_q .

Bridge regression

- We now consider regularized regression models, which shrink the regression coefficients by imposing a penalty on their magnitude.
- More specifically, we focus on *bridge regression* models (Frank and Friedman, 1993), where the coefficients are obtained by minimizing residual sum of squares subject to a constraint on the size of regression coefficients:

$$\begin{aligned} \text{minimize } RSS(\beta) &= \sum_i (y_i - \beta_0 - x_i^T \beta)^2 \\ \text{subject to } \sum_{j=1}^p |\beta_j|^\gamma &\leq s \end{aligned}$$

- We usually scale and center x , and center y so we don't have to deal with β_0 .

Bridge regression

- Alternatively, we can find the estimate by solving the following optimization problem instead:

$$\text{minimize } RSS(\beta) + \lambda \sum_{j=1}^p |\beta_j|^\gamma$$

where $\lambda \geq 0$.

- That is, we minimize a penalized residual sum of squares.
- In the matrix form,

$$\min_{\beta} (y - X\beta)^T (y - X\beta) + \lambda \sum_{j=1}^p |\beta_j|^\gamma$$

- When $\gamma = 0$, the bridge regression becomes equivalent to best subset selection.

Ridge regression

- When $\gamma = 2$, we obtain a special case of the bridge regression known as the *ridge regression* (Hoerl and Kennard, 1970)

$$\text{minimize } RSS(\beta) + \lambda \sum_{j=1}^p \beta_j^2$$

- In ridge regression, the estimates are shrunk towards zero and each other.
- The ridge regression solutions are

$$\hat{\beta}^{\text{ridge}} = (x^T x + \lambda I_p)^{-1} x^T y$$

- Assignment: Find the relationship between the estimate of $\hat{\beta}^{\text{ridge}}$ and $\hat{\beta}^{\text{LS}}$. Show that $\hat{\beta}^{\text{ridge}}$ is biased.

Ridge regression

- Since $x^T x + \lambda I_p$ is non-singular as long as $\lambda > 0$, ridge regression provides a unique solution for a given λ even if $x^T x$ is not of full rank (e.g., $p > n$).
- The L_2 penalty applied to RSS shrinks the coefficients towards zero (and each other).
- The imposed penalty prevents the estimates of regression coefficients to become large.
- This is of course based on our belief that very large values of β are not very likely and should be discouraged.
- For example, if there are two highly correlated variable in the regression model, a large positive coefficient on one of them can be canceled by a large negative coefficient on the other one. This can be prevented by using penalized RSS.

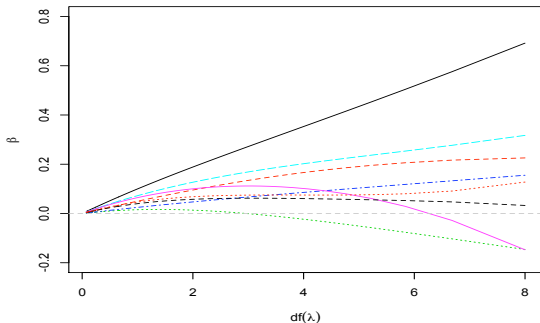
Ridge regression

- The larger the value of λ , the greater the amount of shrinkage.
- However, since the effect of the penalty depends on the scale of covariates, it is common to standardize the covariates so they all have standard deviation 1.
- The estimates from ridge regression are biased but they have lower variance compared to least-squares estimates.
- The overall performance of course depends on how well we choose λ . To choose an appropriate λ , it is common to use cross validation.

Ridge regression for prostate cancer data

- The following plot shows the estimate of parameters for different values of λ . The horizontal line shows the *effective degrees of freedom* defined as follows

$$df(\lambda) = \text{tr}[x(x^T x + \lambda I_p)^{-1} x^T]$$



Lasso

- When $\gamma = 1$, the bridge regression becomes equivalent to the *lasso* (least absolute shrinkage and selection operator).
- Lasso is similar to ridge regression, but instead of L_2 penalty, we use the L_1 penalty $\sum_{j=1}^p |\beta_j|$

$$\text{minimize } RSS(\beta) + \lambda \sum_{j=1}^p |\beta_j|$$

- As before, the penalty results in the shrinkage of coefficients towards zero.
- However, by using the the L_1 penalty and a large enough λ , some of the coefficients could become exactly zero (i.e., become excluded from the model).
- Unlike ridge regression, $\hat{\beta}^{\text{lasso}}$ do not have closed form. Therefore, we need to use optimization techniques.

Lasso

- Figure 3.12 in The Elements of Statistical Learning illustrates the difference between ridge regression and lasso.
- It is clear that the L_1 penalty allows for some of the coefficients to be exactly zero.
- This is also clear from the fact that the derivative of the lasso penalty with respect to β remains constant for all $\beta > 0$, whereas in ridge regression the penalty is proportional to β .
- As the result, in ridge regression the effect of penalties reduces as β moves close to zero, whereas in lasso, there is a continuing force to move β towards zero.

Lasso for prostate cancer data

