STATS 235: Modern Data Analysis
Controlling Complexity

Babak Shahbaba

Department of Statistics, UCI

Background

- Overly complex models tend to overfit the data.
- In this lecture, we mainly focus on model complexity related to the number of variables included in the model within the frequentist framework.
- 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.
- Throughout this lecture, we assume the variables are standardized to have mean zero and variance 1.

Review: Least squares estimates

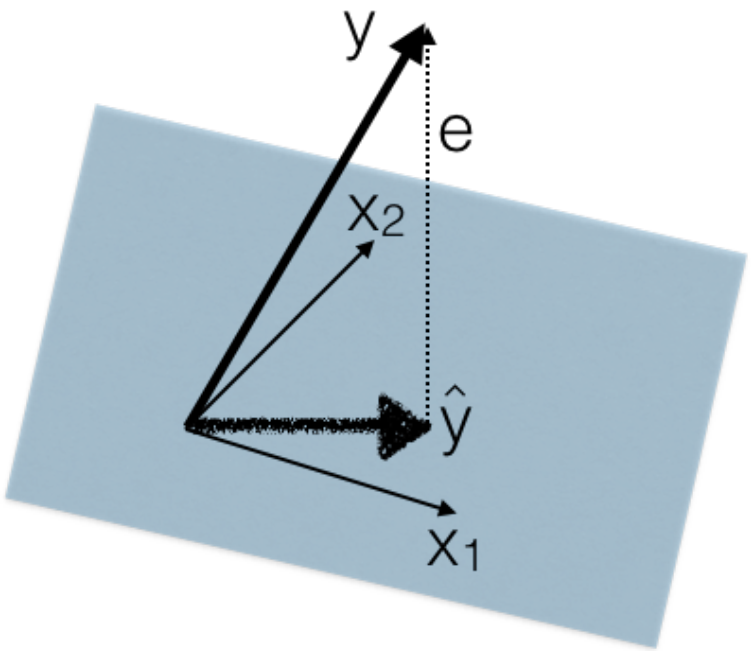
- We start with discussing least square estimates for $X\beta = y$, where X is a $n \times p$ $(n > p)$ matrix
- This doesn't have any solution: X^{-1} doesn't exist; the system is overdetermined (too many equations)
- Instead, we find a solution $\hat{\beta}$ such that

$$X\hat{\beta} = \hat{y}; \qquad y = \hat{y} + e$$

- We find the best solution $\hat{\beta}$ by making e small so y and $\hat{y}$ are "close" to each other
- We can minimize $\|e\|^2 = \|y - X\hat{\beta}\|^2 = (y - X\hat{\beta})^\top (y - X\hat{\beta})$

Review: Least squares estimates

- Geometrically, however, e would be small when it's perpendicular to $\hat{y}$ and the column space of X



Review: Least squares estimates

- e is in the null space of $X^\top$

$$\begin{aligned} X^\top e &= 0 \\ X^\top (y - \hat{y}) &= 0 \\ X^\top (y - X\hat{\beta}) &= 0 \end{aligned}$$

- From this, we get the following normal equation,

$$X^\top X \hat{\beta} = X^\top y$$

- Therefore,

$$\begin{aligned} \hat{\beta} &= (X^\top X)^{-1} X^\top y \\ \hat{y} &= X\hat{\beta} = X(X^\top X)^{-1} X^\top y = Hy \end{aligned}$$

Review: Least squares estimates

- H is symmetric ($H^T = H$) and idempotent ($H^2 = H$).
- The trace of H is its rank, which in this case is the dimension of the projection space and the number of model parameters
- Therefore, we can use $\mathrm{tr}(H)$ to capture the degree of freedom and use it as a measure of complexity
- We can also find the residual vector, e, as follows:

$$e = (I - H)y$$

- $I - H$ is also symmetric and idempotent

Principal component regression

- To control the complexity of regression models, we can use PCA to reduce the dimensionality of the observed data, and hence the number of parameters
- Consider the centered matrix of predictors, x
- As discussed before, we can find principal components and the corresponding score, z
- We then define a new set of derived predictors using the first q columns of z
- We can choose q using the scree plot or cross validation

Principal component regression

- Principal component regression (PCR) is a linear regression model that uses $z_1, \ldots, z_q$ instead of the original predictors,

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

  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 predictors.

Partial least squares

- We could identify a set of new bases according to the relationship between predictors and the response variable.
- The partial least squares (PLS) method performs this task as follows:
  1. Find the univariate regression coefficient $\hat{\phi}_{1j}$ of y on each x_j.
  2. Obtain the first derived input $z_1 = \sum_{i=1}^{p} \hat{\phi}_{1j} x_j$.
  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 $z_1, \ldots, z_q$.

Bridge regression

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

$$\text{minimize } RSS(\beta) = \sum_i (y_i - \beta_0 - x_i^T \beta)^2$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j|^\gamma \leq s$$

- 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 $\lambda \geq 0$ is the Lagrange multiplier.

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

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

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

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

Ridge regression

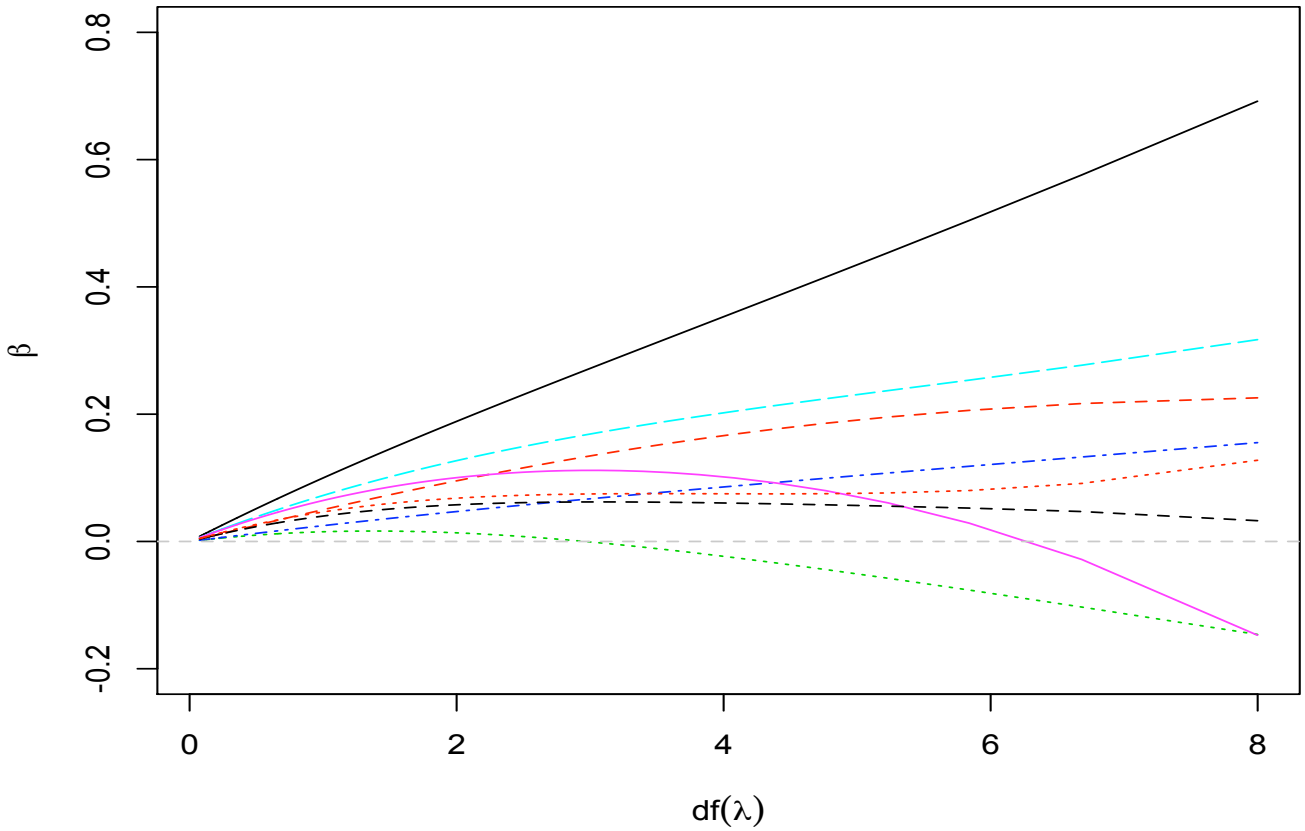
- Since $x^T x + \lambda I_p$ is non-singular as long as $\lambda > 0$, ridge regression provides a unique solution for a given λ even if $x^T x$ is not of full rank (e.g., $p > n$).
- The L_2 penalty applied to RSS shrinks the coefficients towards zero (and each other).
- The imposed penalty prevents 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.

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 predictors, it is common to standardize the predictors 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 based on the prostate cancer dataset (see my codes).



Ridge regression for prostate cancer data

- The horizontal line shows the *effective degrees of freedom* defined as follows

$$\begin{aligned} df(\lambda) &= \mathrm{tr}(H_\lambda) \\ &= \mathrm{tr}[x(x^T x + \lambda I_p)^{-1} x^T] \\ &= \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda} \end{aligned}$$

where d_j is the jth diagonal element of D obtained from the Singular Value Decomposition (SVD): $X = UDV^\top$.

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).

Lasso

- Figure 3.11 in Hastie et al. (2010) illustrates the difference between ridge regression and lasso.

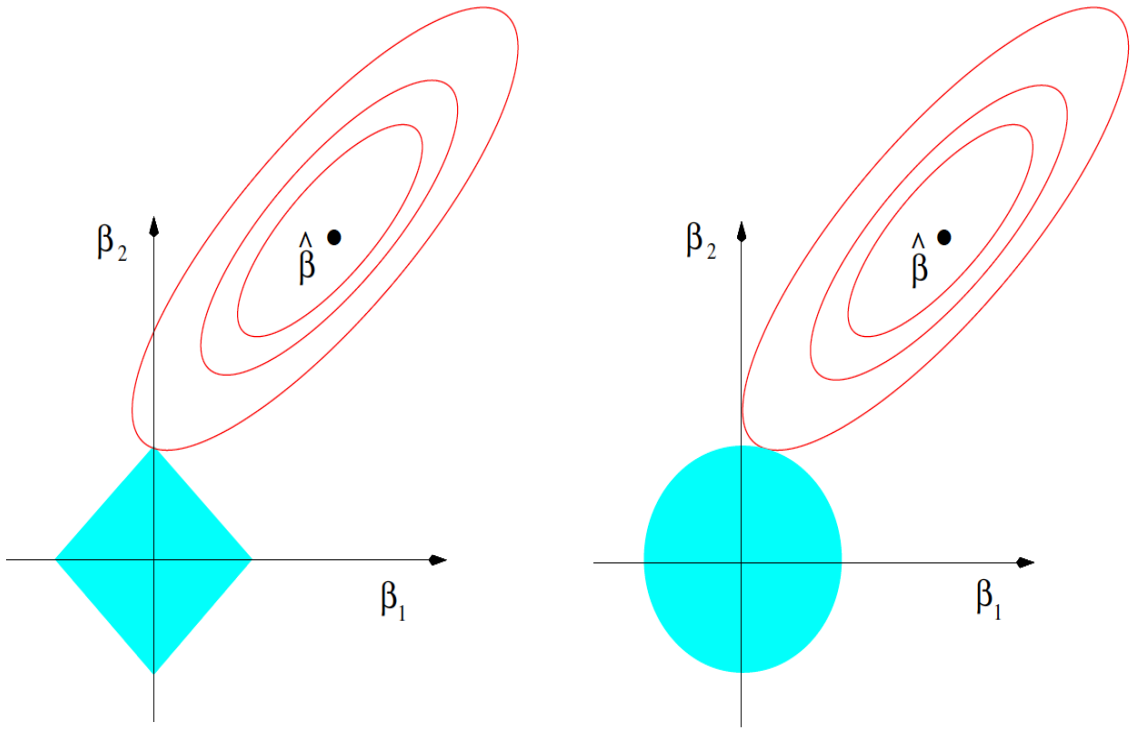
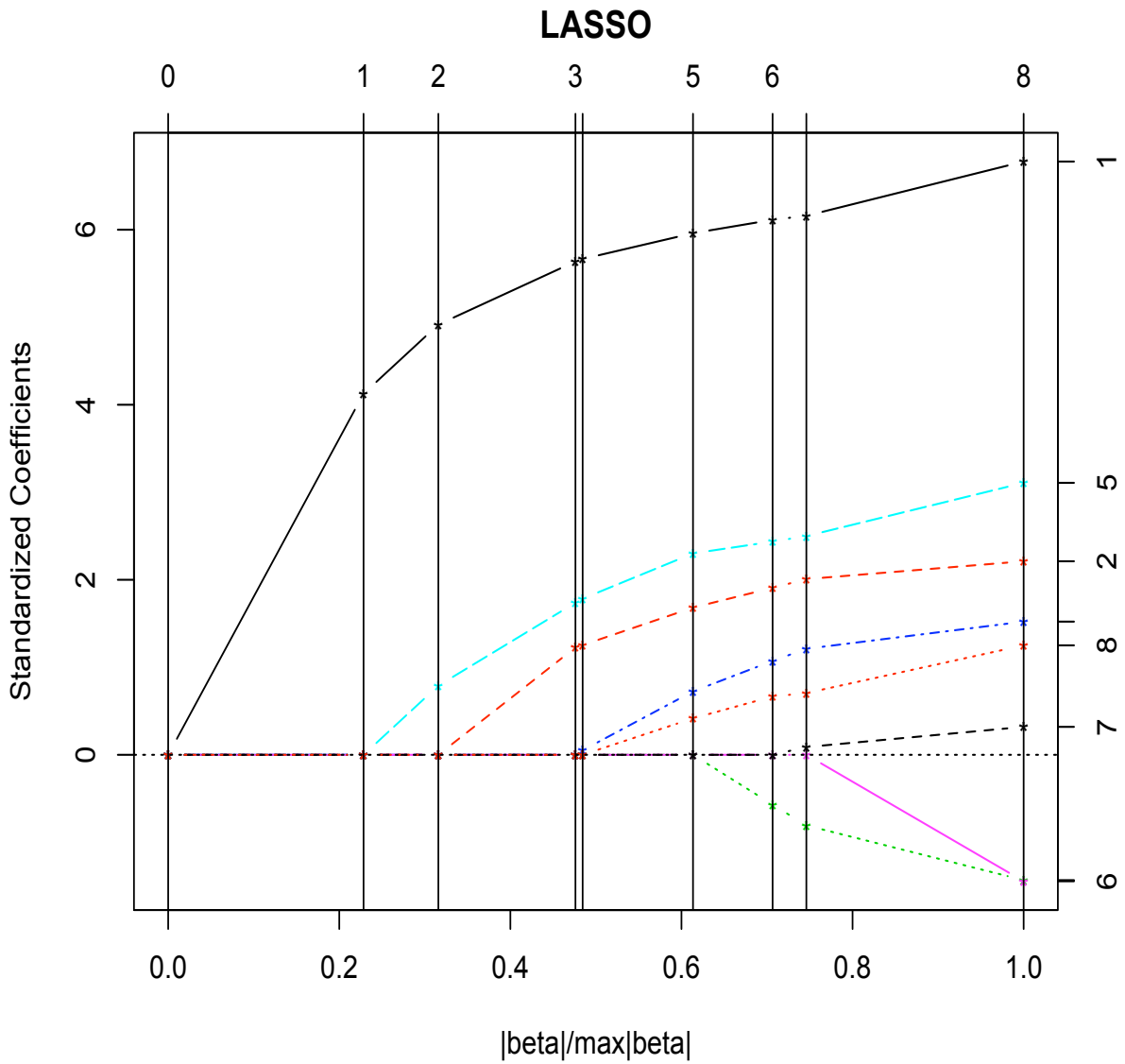


Figure 3.11 in Hastie et al. (2010). Left panel: Lasso; Right panel: Ridge regression.

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 closer to zero, whereas in lasso, there is a continuing force until we reach zero.

Lasso for prostate cancer data



Bayesian interpretation

- Note that the above models have Bayesian interpretation, where the penalty term plays role of prior
- To see this, recall the Bayesian model we discussed before (without the intercept term):

$$\begin{aligned} y_i|x_i,\beta,\sigma^2 &\sim N(x_i^\top\beta,\sigma^2) \qquad i=1,\ldots,n \\ \beta_j &\sim N(0,\tau^2) \qquad j=1,\ldots,p \end{aligned}$$

- we can write the minus log-posterior (up to some constant) as follows (given τ^2 and σ^2):

$$\sum_{i=1}^{n}(y_i - x_i^\top\beta)^2 + \frac{\sigma^2}{\tau^2}\sum_{j=1}^{p}\beta_j^2$$

which is analogues to the penalized RSS in ridge regression.

Bayesian interpretation

- If instead of a normal prior we use a Laplace prior with mean zero,

$$P(\beta_j) = \frac{1}{2\tau}\exp(-\frac{|\beta_j|}{\tau}) \qquad j = 1, \ldots, p$$

  the resulting model is analogous to Lasso, but it is not the same as Lasso since it does not create sparsity.

- As we can see, in these models the regularization term plays the role of the prior.