CHAPTER 5 : SHRINKAGE METHODS

- Selection of subsets of variables in a regression context is a widely used technique. This usually provides a more interpretable model that possibly has a lower prediction error. However, this is a discrete process – variables are either retained or discarded. This process tends to have a high variance. Best subset may lead to different subsets on cross validation. In contrast, shrinkage methods are more like a continuous one and have lower variance.
- Another motivation comes from the multicollinearity perspective (ill-conditioned X). Because OLS estimates depend upon $(X'X)^{-1}$, we would have problems in computing β_{OLS} if $X'X$ were singular.
- One way out of this situation is to abandon the requirement of an unbiased estimator.

RIDGE REGRESSION:-

Suppose we are fitting the model

$$Y = \beta_0 + \sum \beta_i X_i + \epsilon$$

and we may estimate the regression coefficients by OLS that minimize

$$\text{Residual sum of squares (RSS)} = \sum_{i=1}^{n}\Big(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\Big)^2.$$

Hoerl and Kennard (1970) [Techometrics] proposed to take "loss+penalty" approach that attempts to shrink the coefficients by imposing a penalty on their size. Two alternative formulations of ridge regression are as follows.

$$\hat{\beta}_R = \min\left\{\sum_{i=1}^{n}\Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right\}; \quad \text{------} \quad (A)$$

where λ is the shrinkage parameter (tuning parameter) to de determined seperately and $\lambda \geq 0$.

When $\lambda = 0$, ridge regression reduces to OLS (MLR).

As $\lambda\uparrow$, the shrinkage becomes greater. The model becomes a null model when $\lambda \to \infty$ (the variance becomes zero and bias increases).

An equivalent formulation of ridge regression is

$$\left.\begin{aligned} &\min_{\beta}\sum_{i=1}^{n}\Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2 \\ &\text{subject to } \sum_{j=1}^{p}\beta_j^2 \leq t. \end{aligned}\right\} \quad \text{---------} \quad (B)$$

Formulation (B) makes the size constraints on the coefficients directly visible. There is a one-to-one correspondence between the tuning parameter 'λ' defined in (A) and 't' in (B). Two formulations are equivalent to each other and formulation (B) is sometimes preferred as it makes the size constraint explicit.

<u>Choice of λ:</u> In ridge regression, the tuning parameter (shrinkage) plays a vital role as stated earlier. Unlike that least square regression, ridge regression will produce a different set of coefficient estimates for each value of λ. Cross validation can be implemented to choose λ.

<u>Impact of scale:</u> The least square coefficient estimates by OLS regression are scale invariant. Thus, multiplying X_j by c simply leads to multiplying β_j by $1/c$. Hence, $\hat{\beta}_j X_j$ remains the same irrespective of scale (no matter what unit is used to measure X_j).

However, in ridge regression, change of scale would impact the estimated coefficient of the predictor and may even impact other predictors due to the sum of square constraint. This means that ridge coefficients are NOT scale invariant.

Thus, in ridge regression, the predictors are standardized using the formula:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}}.$$

By virtue of standardization, each predictor will have unit standard deviation.

<u>Centering of Variables:</u> The shrinkage penalty is applied to the coefficients $\beta_1, \beta_2, \ldots, \beta_p$ but not to intercept β_0. This is because β_0 is simply a measure of the mean value of the response when the predictors are zero.

Thus, when the predictors are centered to have mean 0, i.e., when the predictors are transformed as $x'_{ij} = x_{ij} - \bar{x}_j$, the estimated intercept takes the form $\hat{\beta}_0 = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

The remaining coefficients get estimated by a ridge regression without the intercept term.

Thus, we assume only that X's and Y have been centered so that we have no need for a constant term in the regression:

- X is an $n \times p$ matrix with centered columns,
- Y is a centered n-vector.

Parameter estimation in Ridge Regression:

Hoerl and Kennard (1970) proposed that potential instability in the LS estimator $\hat{\beta} = (X'X)^{-1}X'Y$ could be improved by adding a small constant value λ to the diagonal entries of the matrix $X'X$ before taking its inverse. The result is the ridge regression estimator

$$\hat{\beta}_{Ridge} = (X'X + \lambda I_p)^{-1} X'Y$$

$\hat{\beta}_{Ridge}$ is chosen to minimize the penalized sum of squares:

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Thus, Minimize $e'e + \lambda\beta'\beta$ (In matrix representation)

$$\begin{aligned} R(\beta) = RSS(\beta) &= (Y - X\beta)'(Y - X\beta) + \lambda\beta'\beta \\ &= (Y' - \beta'X')(Y - X\beta) + \lambda\beta'\beta \\ &= Y'Y - \underbrace{\beta'X'Y - Y'X\beta}_{} + \beta'X'X\beta + \lambda\beta'\beta \end{aligned}$$

$$\frac{\partial R(\beta)}{\partial \beta} = -2X'Y + 2\beta X'X + 2\lambda\beta$$

Now, $\frac{\partial R(\beta)}{\partial \beta} = 0$

$\Rightarrow (X'X + \lambda I)\beta = X'Y$

$\Rightarrow \hat{\beta}_{Ridge} = (X'X + \lambda I)^{-1} X'Y$ .... (c)

NOTE:
- Traditional description of ridge regression start with (c).
- Ridge coefficients are estimated using least square methodology.
- The choice of quadratic penalty adds a (+ve) constant to the diagonal element of $X'X$. This forces a solution in all cases to mitigate the problem of multicollinearity in regression analysis. Further, the ridge solution is a linear function of y.

WHY DOES RIDGE REGRESSION IMPROVE OVER LEAST SQUARE?

Essentially due to bias-variance trade-off. As λ increases, the flexibility of the ridge regression fit decreases (leading to increasing bias but decreasing variance). At the same time, as λ increases, the shrinkage of a ridge coefficient leads to substantial reduction of variance at the cost of a small increase in bias. We may look at the mean squared error (MSE) of the validation (on test) data to choose the 'right' value of λ. We select λ that yields the smallest cross-validation prediction error.

Usage of Ridge Regression: Ridge is often used when there are many correlated variables. When the explanatory variables are highly correlated among themselves, the coefficient (regression) can become poorly determined and exhibit high variance. A pair of correlated variables often have large (+ve and (-)ve coefficients, cancelling each other. The size constraints / tuning parameter alleviates this problem.

Comment on $\hat{\beta}_{Ridge}$: 1. Whereas $\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$ are unbiased if the model is correctly specified, ridge solutions are biased, i.e.,

$$E(\hat{\beta}_{Ridge}) \neq \beta.$$

However, at the cost of bias, ridge regression reduces the variance, and thus might reduce the variance.

$$MSE = bias^2 + variance$$

2. Since the ridge estimator is linear, we can calculate the variance-covariance matrix

$$Var(\hat{\beta}_{Ridge}) = \sigma^2 (X'X + \lambda I_p)^{-1} X'X (X'X + \lambda I_p)^{-1}.$$

where $\hat{\beta}_{Ridge} = (X'X + \lambda I_p)^{-1} X'Y$.

3. $$\beta_{Ridge} \mid Y \sim N\left(\hat{\beta}_{Ridge}, \sigma^2 (X'x + \lambda I_p)^{-1} X'X (X'X + \lambda I_p)^{-1}\right),$$

confirming that the posterior mean of the Bayesian linear model corresponds to the ridge regression estimator. [From Bayesian statistics point of view]

- Ridge regression shrinks coefficient estimates towards zero. This can be explained by spectral decomposition or via singular value decomposition (SVD).

→ The SVD of the centered input matrix $X_{n \times p}$ gives us some additional insights into ridge regression.

- We write $X_{n\times p} = U_{n\times p} D_{p\times p} V'_{p\times p}$, where U and V are orthogonal matrices with $U'U = UU' = I$ and $V'V = VV' = I$.

$D_{p\times p}$ is a diagonal matrix and the diagonal elements $d_{11}, d_{22}, \ldots, d_{pp}$ are non-negative and these are the singular values of X.

- For MLR model $Y = X\beta + \epsilon$, the least square solution may be written as

$$\hat{\beta} = (X'X)^{-1} X'Y$$

$$X\hat{\beta} = X(X'X)^{-1}X'Y$$

Using SVD, we can write

$$\begin{aligned} X\hat{\beta} &= X(X'X)^{-1}X'Y \\ &= UDV'(VDU'UDV')^{-1} VDU'Y \\ &= UDV'(VDDV')^{-1} VDU'Y \\ &= (D^2)^{-1} UDV'VDU'Y \\ &= UU'Y \end{aligned}$$

- In Ridge regression, we have $\hat{\beta}_R = (X'X + \lambda I)^{-1} X'Y$

$$\begin{aligned} \hat{y} = X\hat{\beta}_{Ridge} = X\hat{\beta}_R &= X(X'X+\lambda I)^{-1} X'Y \\ &= UDV'(VD^2V' + \lambda I)^{-1} VDU'Y \\ &= UDV'(VD^2V' + \lambda VV')^{-1} VDU'Y \\ &= UDV'(V(D^2+\lambda)V')^{-1} VDU'Y \\ &= UDV'V(D^2+\lambda)^{-1} V'VDU'Y \\ &= UD(D^2+\lambda I)^{-1} DU'Y \\ &= \sum_{j=1}^{p} u_j \cdot \frac{d_j^2}{d_j^2+\lambda} \cdot u_j' Y \end{aligned}$$

$\frac{d_j^2}{d_j^2+\lambda}$ is called the shrinkage factor in Ridge regression and u_j are the normalized principal components of X.

$$\hat{\beta}_{jRidge} = \frac{d_j^2}{d_j^2+\lambda} u_j' Y \text{ and } Var(\hat{\beta}_j) = \frac{\sigma^2}{d_j^2}; \quad \epsilon_i \sim N(0, \sigma^2).$$

For large λ, the projection is shrunk in the direction of u_j.

LEAST ABSOLUTE SHRINKAGE & SELECTION OPERATOR (LASSO):

Ridge regression does not achieve parsimony. The penalty $\lambda \sum_{j=1}^{p} \beta_j^2$ shrinks all $\beta_j \to 0$. This is when the ridge solution can be hard to interpret and it is not sparse. Interpretability may also be a problem as p does not reduce in ridge formulation.
What if we constrain the L_1 norm instead of the Euclidean (L_2) norm?

Ridge subject to : $\sum_{j=1}^{p} \beta_j^2 \leq t$.

Lasso subject to : $\sum_{j=1}^{p} |\beta_j| \leq t$.

- This is a subtle, but important change. Some of the coefficients may be shrunk exactly to zero. (Tibshirani' 1996)
- A significant difficulty with Ridge is its inability to select a subset of variables. The penalty $\lambda \sum \beta_j^2$ or the constraint $\sum \beta_j^2 \leq t$ shrinks all coefficients to zero but does not set any one of them to zero unless $\lambda = \infty$.
- LASSO performs variable selection. The L_1 norm has the effect of forcing some coefficients to become zero, leading to a sparse model. In order to select a subset of variables we minimize

$$Q = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (*)$$

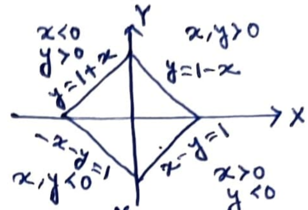
- The lasso may be alternatively formulated as

$$\left.\begin{aligned} &\underset{\beta}{\text{Minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \\ &\text{subject to} \sum_{j=1}^{p} |\beta_j| \leq t. \end{aligned}\right\} \quad (**)$$

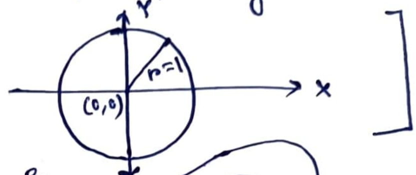
- As in Ridge, choice of λ is crucial and ususally cross validation is used.
- Formulations (*) and (**) are equivalent in the same that for every λ one can find a t that gives the same set of coefficient estimates and the other way round.

Geometric Interpretation: The lasso performs L_1 shrinkage so that there are "corners" in the constraint, which in two dimensions corresponds to a diamond. If sum of squares "hit" one of these corners, then the coefficient corresponding to the axis is shrunk to zero.

[Recall the graph of $|x|+|y| \leq 1$:



Recall the graph of $x^2+y^2 \leq 1$:



]

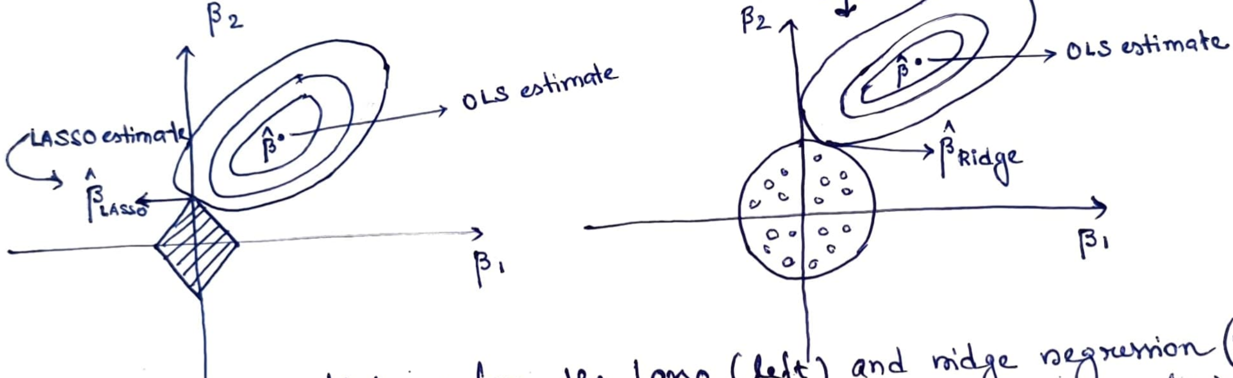


Fig: Estimation picture for the Lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The areas are the constraint regions $|\beta_1|+|\beta_2| \leq t$ and $\beta_1^2+\beta_2^2 \leq t^2$, respectively, while the ellipses are the contours of the least squares error function. The figures are when $p=2$, lasso coefficients correspond to the least RSS for (β_1, β_2) falling in the diamond described by $|\beta_1|+|\beta_2| \leq t$ whereas ridge regression estimates have the smallest RSS out of all points that lie within the circle defined by $\beta_1^2+\beta_2^2 \leq t^2$.

- As p increases, the multidimensional diamond has an increasing number of corners, and so it is highly likely that some coefficients will be set equal to zero. Hence, the lasso performs shrinkage and (effectively) subset selection.
- Subset selection:

$$\underset{\beta}{\text{Minimize}} \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \sum_{j=1}^{p} I(\beta_j \neq 0).$$

In contrast with subset selection, Lasso performs a soft thresholding: as the smoothing parameter is varied, the sample path of the estimates moves continuously to zero.

Comparing Ridge and Lasso: — No clean winner.

- When some of the case when there is the response can be expressed as a function of only a relatively small number of predictors, Lasso performs better.
- Unlike Ridge regression, there is no analytical solution for the Lasso because the solution is nonlinear in Y.
- Ridge regression shrinks all regression coefficients towards zero; the lasso tends to give a set of zero regression coefficients and leads to a sparse solution.

$\boxed{\text{For research-oriented reading only}}$

Inference for Lasso Estimation:

The ordinary lasso does not address the uncertainty of parameter estimation; standard errors for β's are not immediately available. For inference using the lasso estimator, various standard error estimators have been proposed:

- Tibshirani (1996) suggested the bootstrap (Efron, 1979) for the estimation of standard errors and derived an approximate closed-form estimate.
- Fan and Li (2001) derived the sandwich formula in the likelihood setting as an estimator for the covariance of the estimates.
- However, these two formulation approximate covariance matrices give an estimated variance of 0 for predictors with $\hat{\beta}_j = 0$.
- The "Bayesian lasso" of Park and Casella (2008) provides valid standard errors for β and provides more stable point estimates by using the posterior median.

Prediction models in Shrinkage / regularization models:

When fitting linear shrinkage/regularization models (ridge and lasso), the predictors, X, should be standardized $\left(\frac{X - \text{mean}}{\text{s.d.}}\right)$.

For a brand-new X, the prediction model is

$$\hat{y}_{new} = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j \frac{X_{new,j} - \text{mean}(X_{train,j})}{sd(X_{train,j})},$$

where $\hat{\beta}$'s are estimated from the training data. The R function "glmnet" performs the standardization by default.

Elastic Net: A compromise between Ridge & LASSO by Zou & Hastie (2006)

$$\hat{\beta}_{EN} = \text{loss} + \lambda \sum_{j=1}^{p} \left(\alpha \beta_j^2 + (1-\alpha)|\beta_j|\right)$$