Econ 31703: Assignment 2

Due date: May 5, 2021

Exercise 1

Let us build our own LASSO function step by step and compare LASSO with Ridge and OLS estimation. Lastly, we will see how LASSO estimates change along with the penalty parameter.

(a) Write a function that calculates the LASSO objective function for given coefficients:

$$\hat{\beta}^{lasso} = \arg\min_{\mathbf{b}} \left[\frac{1}{2} \sum_{i=1}^{N} (Y_i - \mathbf{X}_i{}^{\intercal}\mathbf{b})^2 + \lambda ||\mathbf{b}||_1 \right].$$

```
lasso.objective <- function(b,data,lambda){
## b is a p × 1 coefficient vector
## data is a n × (p+1) matrix,
## where the first column denotes y and the rest denotes x's.
## lambda is a scalar penalty parameter.
⋮
return(a scalar lasso objective)
}
```

(b) Write a function that updates the coefficient by its each coordinate.

```
lasso.update <- function(b,data,lambda){
for (k in 1:p) {
    ## k indicates which coefficient to update.
    ## update k-th coefficient using the closed-form solution.
   ⋮
}
⋮
return(a p × 1 coefficient vector)
}
```

(c) Write a wrapper function, which uses your `lasso.objective` and `lasso.update`. For stopping criterion, stop the iteration when the number of iteration passes a set maximum `max`, or the updates in the estimates are smaller than than a precision parameter `eps`:

$$\max_k \left\{ \left| b_1{}' - b_1 \right|, \cdots, \left| b_p{}' - b_p \right| \right\} < \epsilon.$$

```
lasso <- function(b_initial=rep(0,p),data,lambda,eps=1e-06, max=1000){
## standardize every regressor in data:  sum_{i=1}^N x̃_ip = 0, sum_{i=1}^N x̃_ip^2 = 1
## loop using the two stopping criteria
⋮
ans <- list(estimate =, ## the final estimates
            objective =, ## the sequence of objectives updated
            status =, ## which stopping criterion is used?
)
return(ans)
}
```

(d) Generate 10,000 samples as in Assignment 1, **1-(a)**. Compute $\hat{\beta}^{lasso}$ for each sample with the penalty parameter $\lambda = 20$, using all $p = 90$ variables. Out of 10,000 samples, how many times do we get $\hat{\beta}_1^{lasso} \neq 0$? Compare the averages of $\hat{\beta}_1^{lasso}$ and $\hat{\beta}_2^{lasso}$ across the simulated samples where $\hat{\beta}_1^{lasso} \neq 0$ with that of $\hat{\beta}_1(90)$, the OLS estimate using all 90 regressors. .

(e) Using the same 10,000 samples from **1-(a)**, compute $\hat{\beta}^{ridge}$ for each sample with $\lambda = 20$, using all $p = 90$ variables:

$$\hat{\beta}^{ridge} = \arg\min_{\mathbf{b}} \left[\frac{1}{2} \sum_{i=1}^{N} (Y_i - \mathbf{X}_i{}^{\intercal}\mathbf{b})^2 + \frac{1}{2}\lambda \left|\left|\mathbf{b}\right|\right|_2^2 \right].$$

Compare the averages of $\hat{\beta}_1^{ridge}$ and $\hat{\beta}_2^{ridge}$across the simulated samples with that of $\hat{\beta}_1(90)$.

(f) Randomly choose a sample from the 10,000 samples. Repeat computing $\hat{\beta}^{lasso}$ while varying $\lambda = 0.01\lambda_{\max}, 0.02\lambda_{\max}, \cdots, \lambda_{\max}$ where $\lambda_{\max}$ is the smallest value of λ which gives us $\hat{\beta}^{lasso} = \mathbf{0}$. Take $\hat{\beta}_1^{lasso}, \cdots, \hat{\beta}_5^{lasso}$ and plot the five estimates as functions of λ.

Exercise 2

Consider the following DGP: with $\rho \in (0,1)$, $X_t \in \mathbb{R}^{50}$,

$$\begin{aligned} X_0 &\sim \mathcal{N}\left(\mathbf{0}, \left(1-\rho^2\right)^{-1}\Sigma\right), \\ X_{t+1} &= \rho X_t + \varepsilon_{t+1}, \quad \varepsilon_{t+1} \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \Sigma), \\ Y_t &= X_t^{\intercal}\beta + \eta_t, \quad \eta_t \overset{iid}{\sim} \mathcal{N}(0,1), \end{aligned}$$

where $X_0, \varepsilon_1, \cdots, \varepsilon_T, \eta_0, \cdots, \eta_T$ are jointly independent of each other. Set $T = 200$ and $\rho = 0.9$

An econometrician does not have knowledge of the DGP and wants to estimate a forecasting model as follows: with some $h \in \mathbb{N} \cup \{0\}$, they estimate

$$Y_{t+h} = \beta(L)X_t + u_t.$$

Even with a moderate choice on the number of lags to use, the forecasting model cannot be estimated with OLS.

(a) Write a code that experiments with the penalty parameter and returns the 'best' in terms of forecasting error.

```
cross.validation <- function(data,lambda_seq,tau,h){
## lambda_seq is a sequence of values to experiment with.
## tau = (tau1, tau2) is a vector of n. of obs. in training set and in
test set, respectively.
## h is the number of periods we predict ahead.
⋮
return(sequence of MSFE)
}
```

Let $B = T - \tau_1 - h + 1$. The function should make B forecasts. Firstly, the function estimates

$$Y_{t+h} = X_t^{\intercal}\beta + u_{t+h}, \quad t = 0, \cdots, \tau_1 - 1.$$

.

with LASSO with penalty parameter λ. Using $\hat{\beta}$ from this estimation, make τ_2 forecasts:

$$\hat{Y}_{t+h} = X_t^{\intercal}\hat{\beta}, \quad t = \tau_1, \cdots, \tau_1 + \tau_2 - 1,$$

which completes one cycle of training and test. Then the function estimates using observations from $t = \tau_2, \cdots, \tau_1 + \tau_2 - 1$ and make forecasts for $Y_{\tau_1+\tau_2+h}, \cdots, Y_{\tau_1+2\tau_2+h-1}$. Repeat this until the function makes a forecast for Y_T. The penalty parameter λ is evaluated in terms of the mean squared forecasting error (MSFE):

$$\frac{1}{B}\sum_{b=0}^{B-1}\left(Y_{\tau_1+b+h} - \hat{Y}_{\tau_1+b+h}\right)^2.$$

(b) Simulate a sample of the DGP with $\Sigma = (1-\rho^2)\mathbf{I}_{50}$ and β such that $\beta_1 = 5$ and $\beta_k = 0$ for $k = 2, \cdots, 50$. With each sample, use your code in **(a)** to cross-validate across your choice of (at least 20) λs when $h = 1, \tau = (100, 10)$[1]. Plot the averages of the MSFE as a functions of λ. Report the estimation result using the full sample under λ minimizing the MSFE.

(c) Repeat **(b)** when $\Sigma = \mathbf{I}_5 \otimes \tilde{\Sigma}$ and $\tilde{\Sigma}$ is a 10×10 symmetric matrix where the diagonal elements are $1 - \rho^2$ and the off-diagonal elements are $0.8 \cdot (1 - \rho^2)$.

[1]The procedure described in **(a)** for the given τ is as follows; use $(Y_1, X_0), \cdots, (Y_{100}, X_{99})$ for the first training set and $(Y_{101}, X_{100}), \cdots, (Y_{110}, X_{109})$ for the first test set; then, use $(Y_{10}, X_9), \cdots, (Y_{110}, X_{109})$ for the second training set and $(Y_{111}, X_{110}), \cdots, (Y_{120}, X_{119})$ for the second training set; repeat this until you have no more observation. Then compute the MSFE with $Y_{101}, \cdots, Y_{200}$.