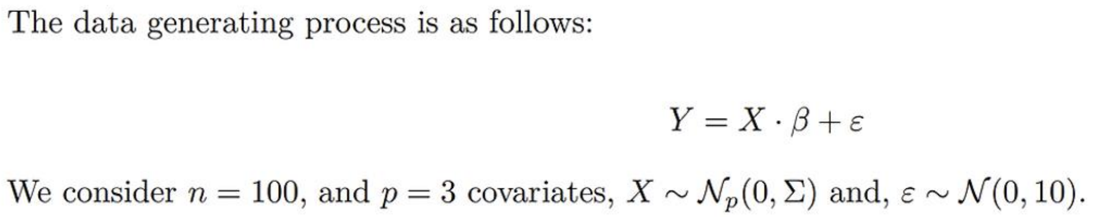
The Coefficients' Shrinkage Path of Ridge and Lasso Regression

Here we compare the shrinkage path of Ridge and Lasso regression, here the shrinkage path refers to the different coefficients of different tuning parameters. Because after we know how parameters perform, we can better understand how shrinkage methods work.

We use the following data generating process as baseline setting:



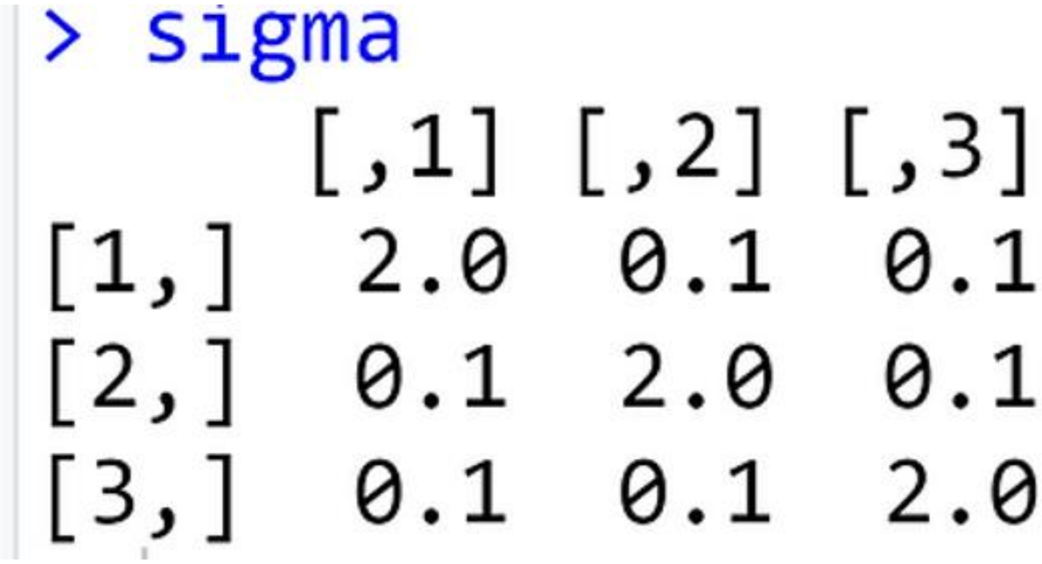
The data generating process is as follows:

$$Y = X \cdot \beta + \varepsilon$$

We consider $n = 100$, and $p = 3$ covariates, $X \sim \mathcal{N}_p(0, \Sigma)$ and, $\varepsilon \sim \mathcal{N}(0, 10)$.

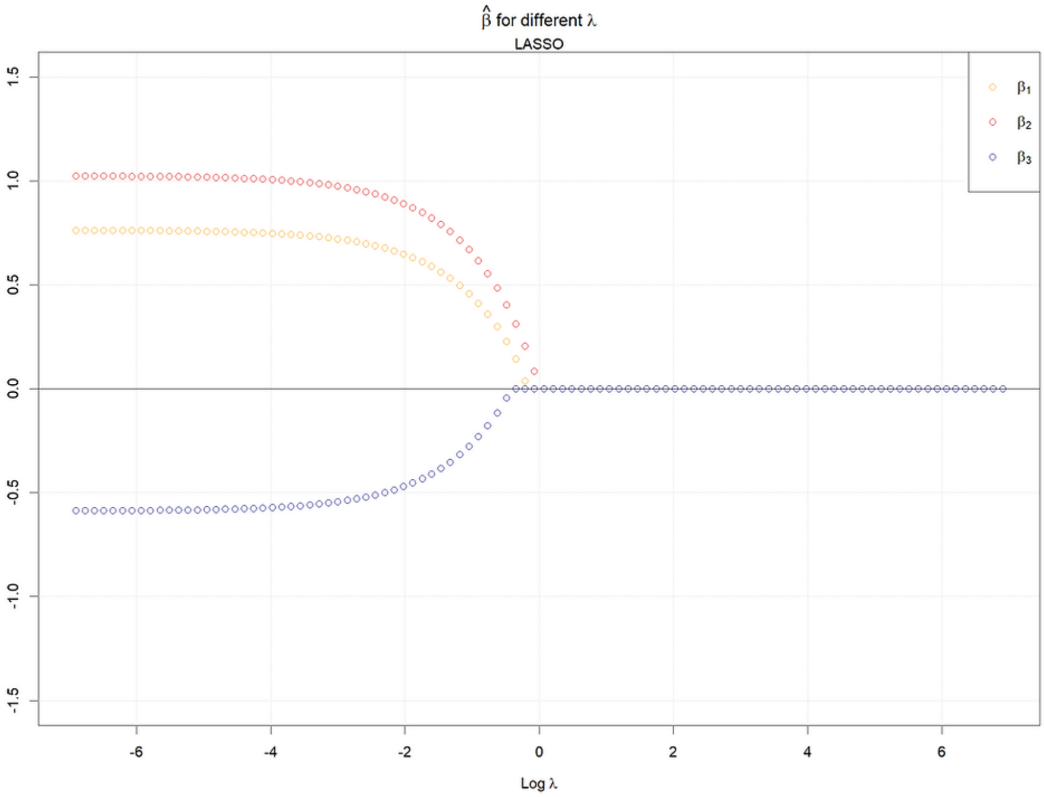
the true beta vector is (0.5 0.5 -0.5)

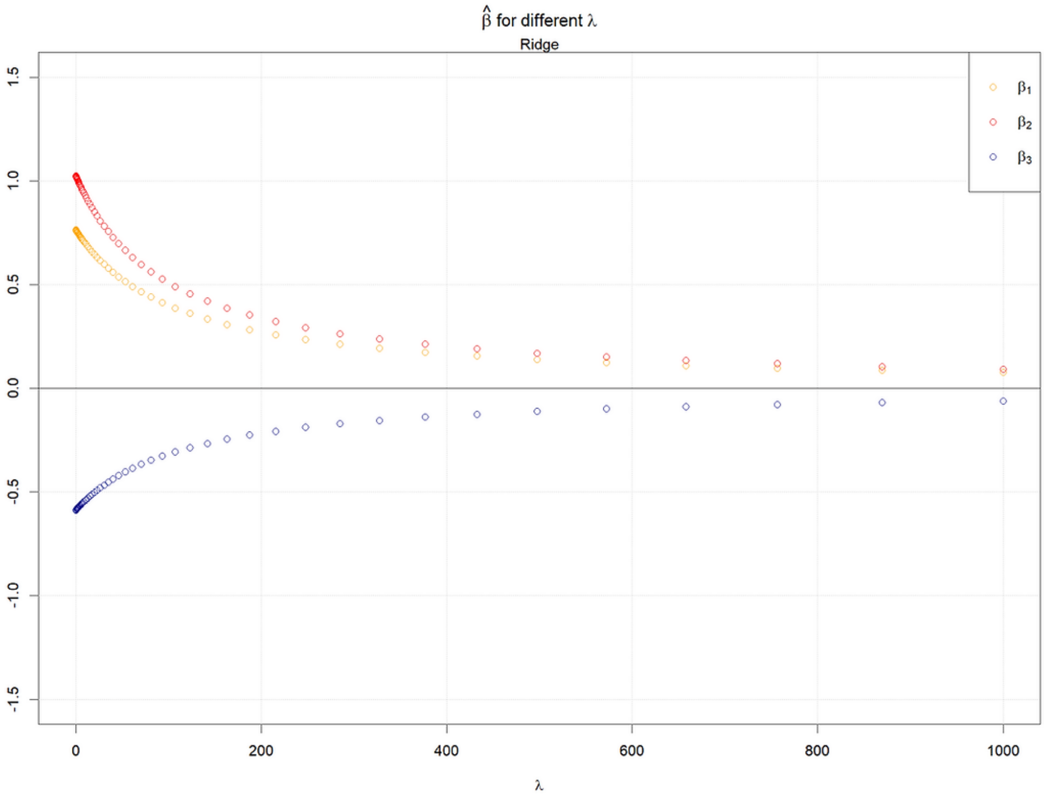
variance-covariance matrix sigma is



```
> sigma
     [,1] [,2] [,3]
[1,]  2.0  0.1  0.1
[2,]  0.1  2.0  0.1
[3,]  0.1  0.1  2.0
```

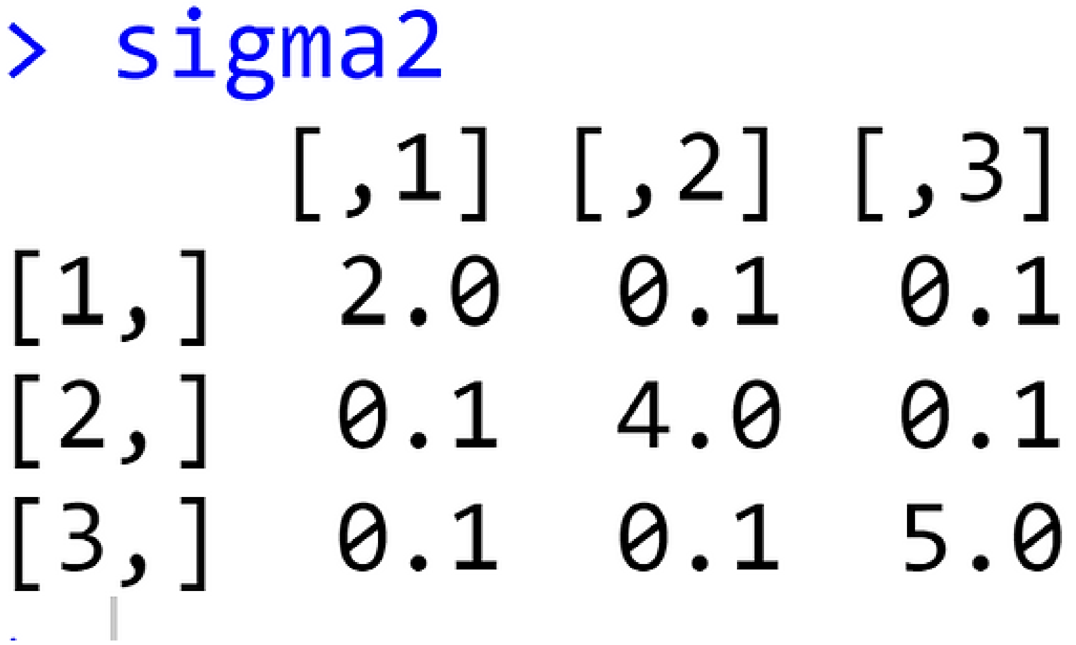
Case 1, the baseline case:
Note that the horizontal axis of the Lasso figure is the logged tuning parameter interval.



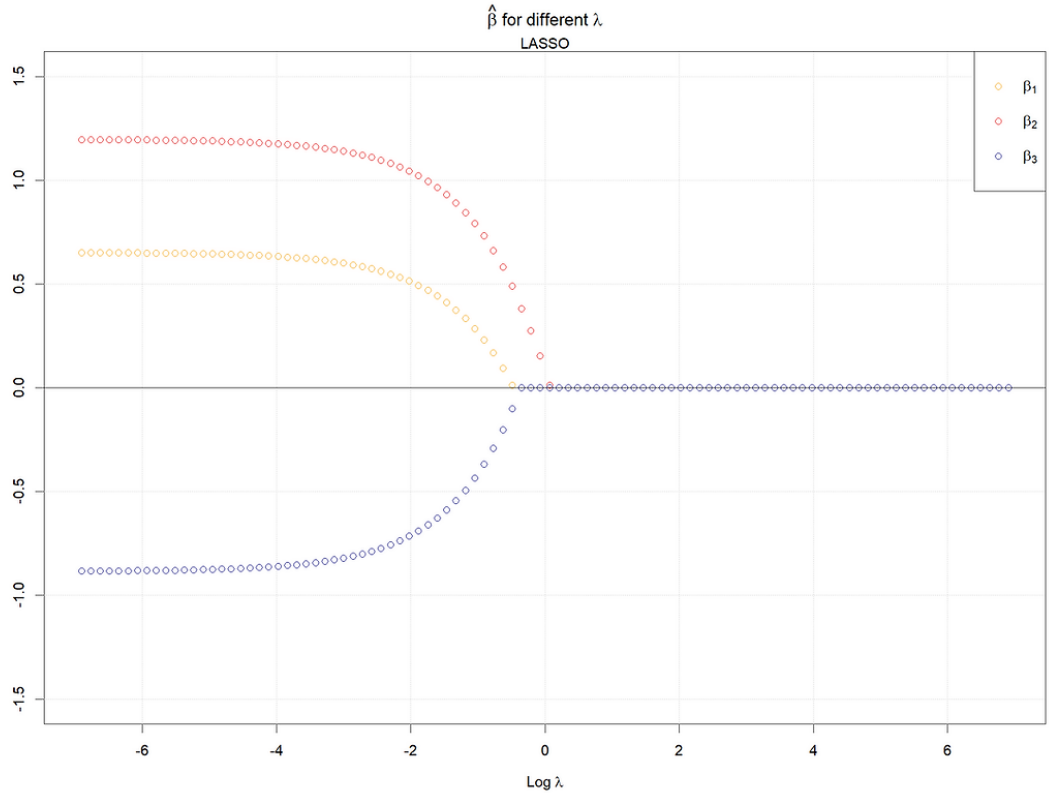


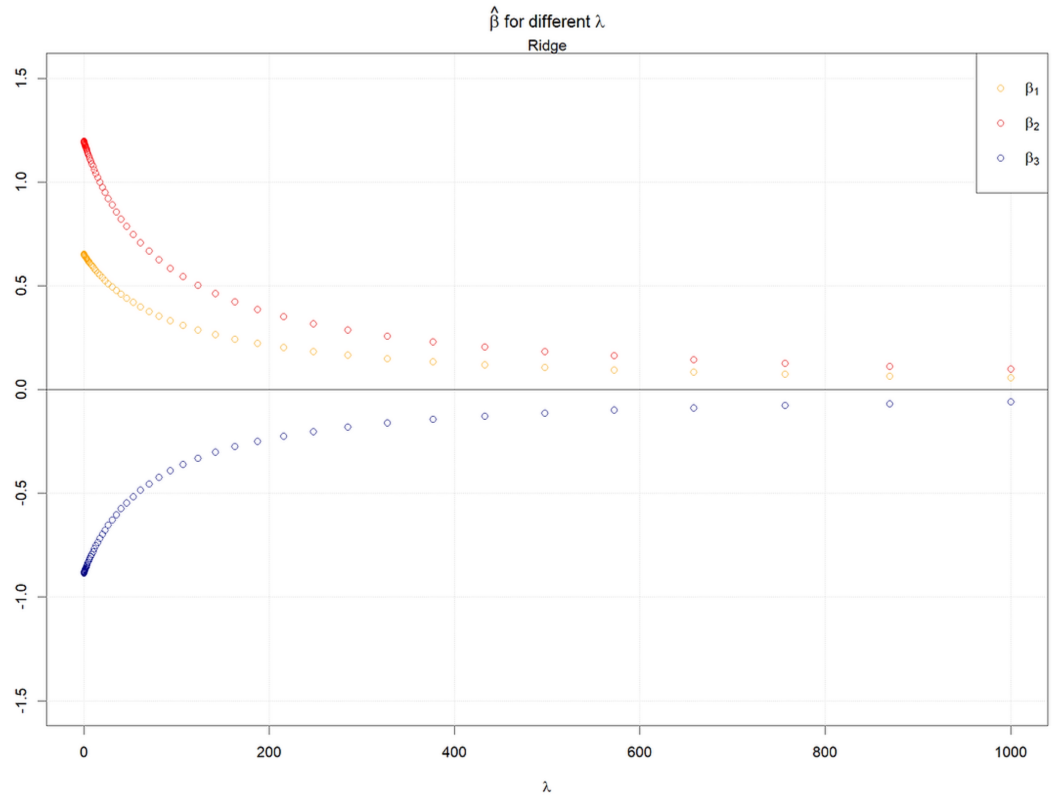
Case 2: Variables with different variance

Here we change the var-covariance matrix, let X2 and X3 have larger variances, and covariance keeps the same as before. However, for both shrinkage methods, standardization is recommended, and here both regressions are done with standardized data, thus case 2 makes less sense, and the explanation of cases 1 and 2 should be the same :)



```
> sigma2
     [,1] [,2] [,3]
[1,]  2.0  0.1  0.1
[2,]  0.1  4.0  0.1
[3,]  0.1  0.1  5.0
```

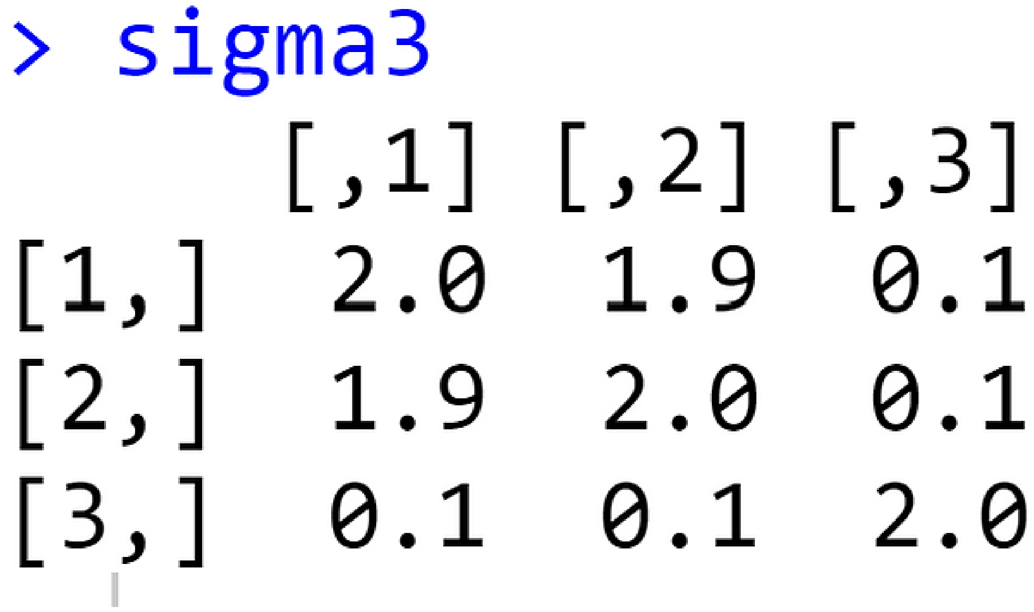




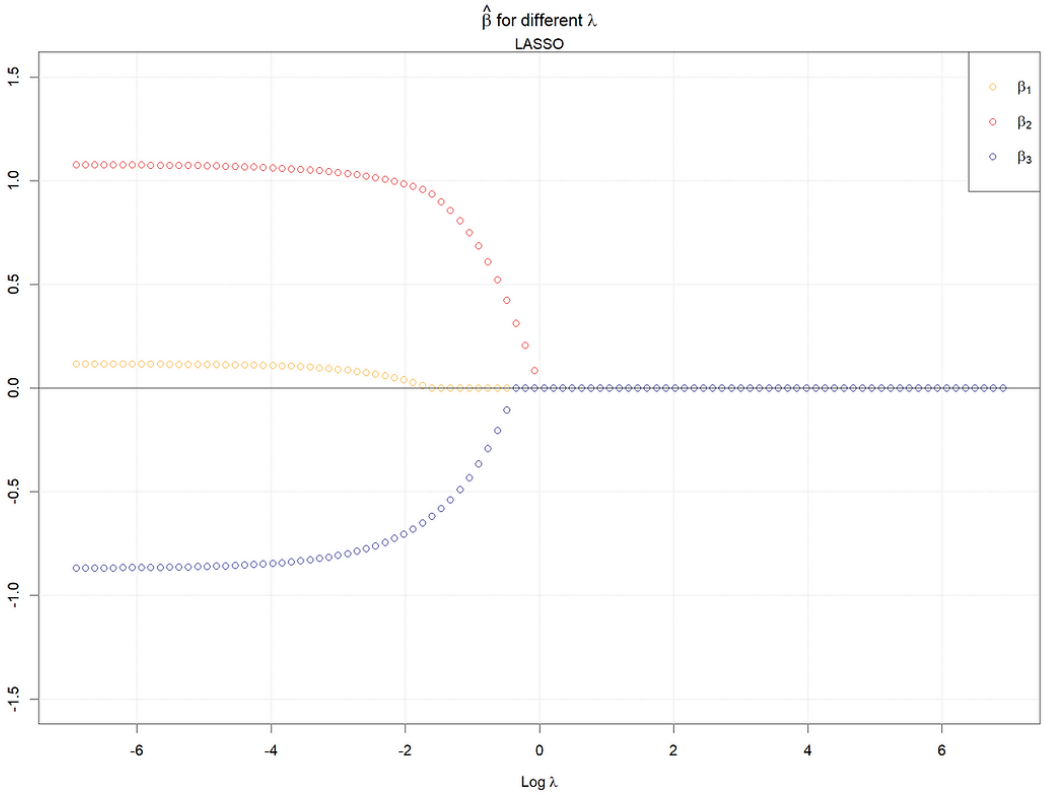
Both in cases 1 & 2, for Lasso, the larger (absolute value) the coefficients when lambda is 0, the later the coefficients shrink to 0 as lambda increases. Meanwhile, for the ridge regression, rather than shrinking to 0, their values decrease faster given a smaller (absolutely) coefficient (when lambda = 0).

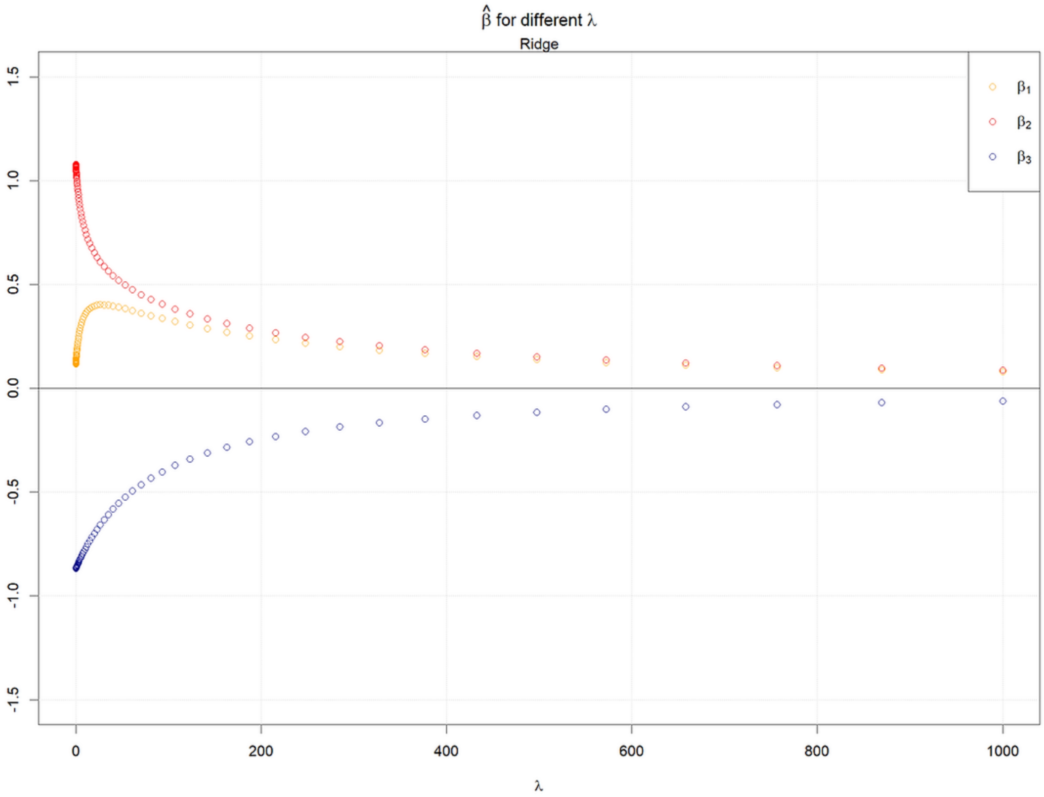
Case 3: High correlated variables

We can set X1 and X2 are highly correlated.



```
> sigma3
     [,1] [,2] [,3]
[1,]  2.0  1.9  0.1
[2,]  1.9  2.0  0.1
[3,]  0.1  0.1  2.0
```

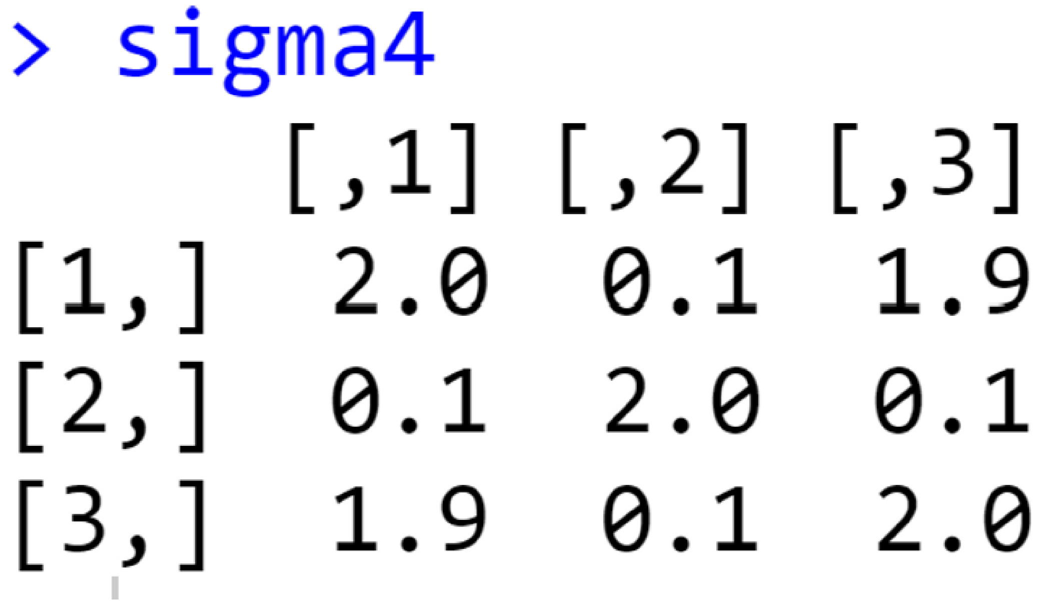




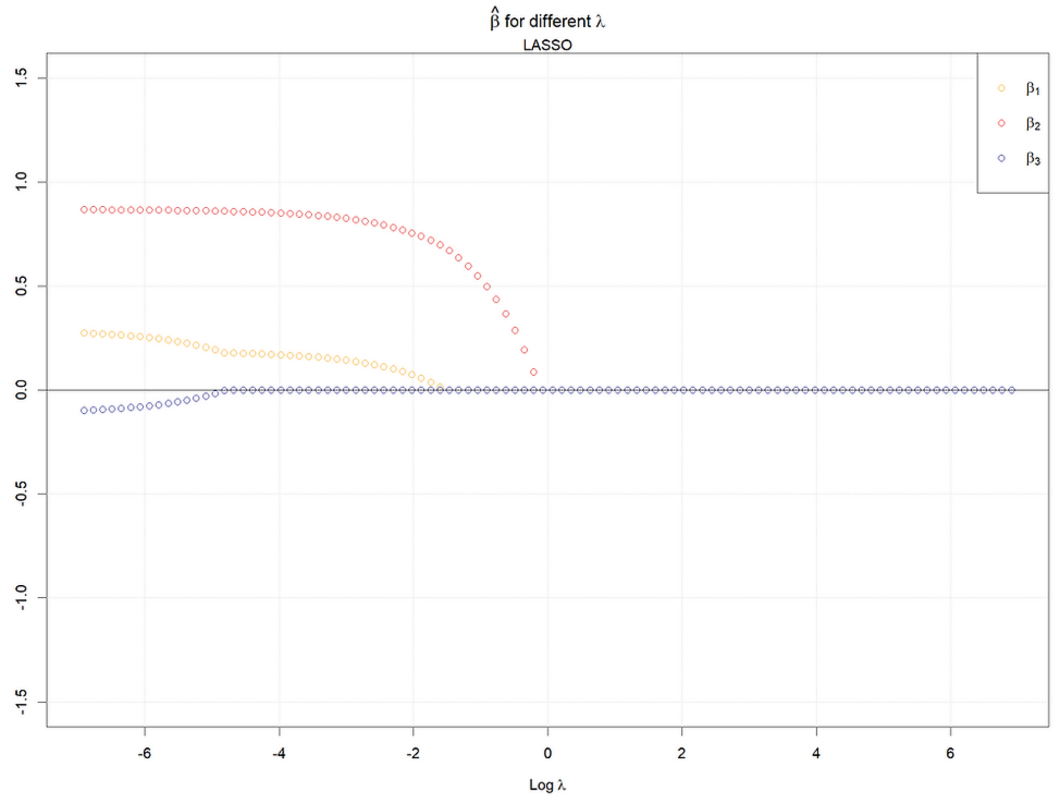
Here we see that X1's lambda 0 coefficient becomes very small and X2's becomes around 1, which makes sense because either X1 or X2 can reflect most information given high covariance.

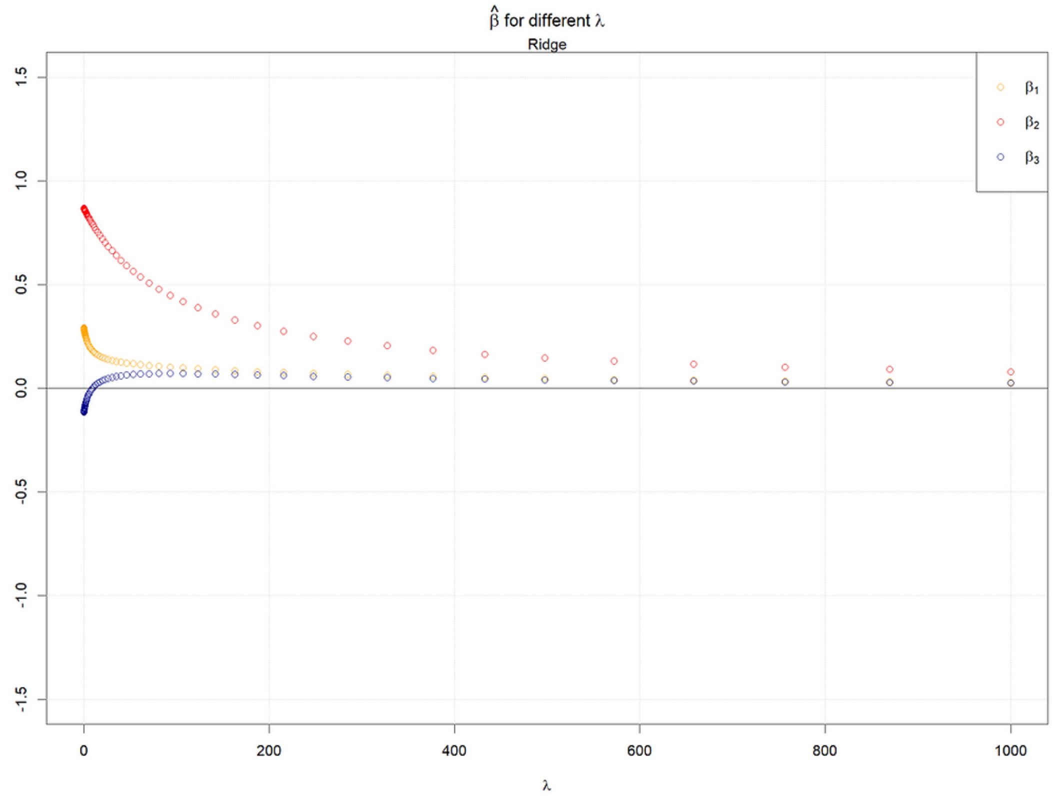
And here we can observe the difference between Ridge and Lasso. For ridge, the two coefficients' path converges to each other, but for the Lasso case, the coefficient of X1 quickly goes to 0. Here we can see the different logic of shrinkage methods of these two. In summary, Ridge makes two coefficients simultaneously decrease, but Lasso drops one of them.

Case 4: High correlation of X1 and X3



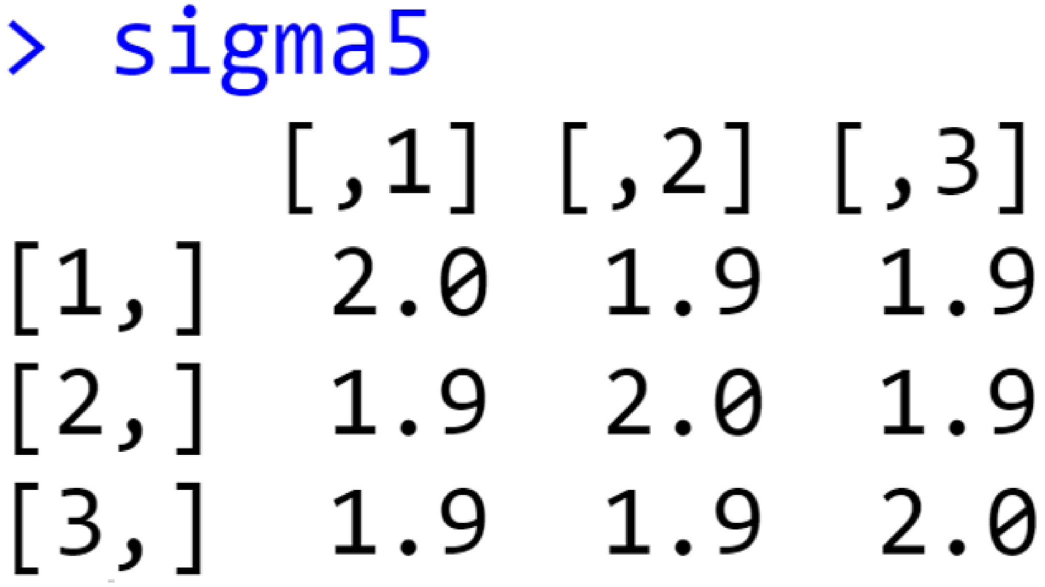
```
> sigma4
     [,1] [,2] [,3]
[1,]  2.0  0.1  1.9
[2,]  0.1  2.0  0.1
[3,]  1.9  0.1  2.0
```



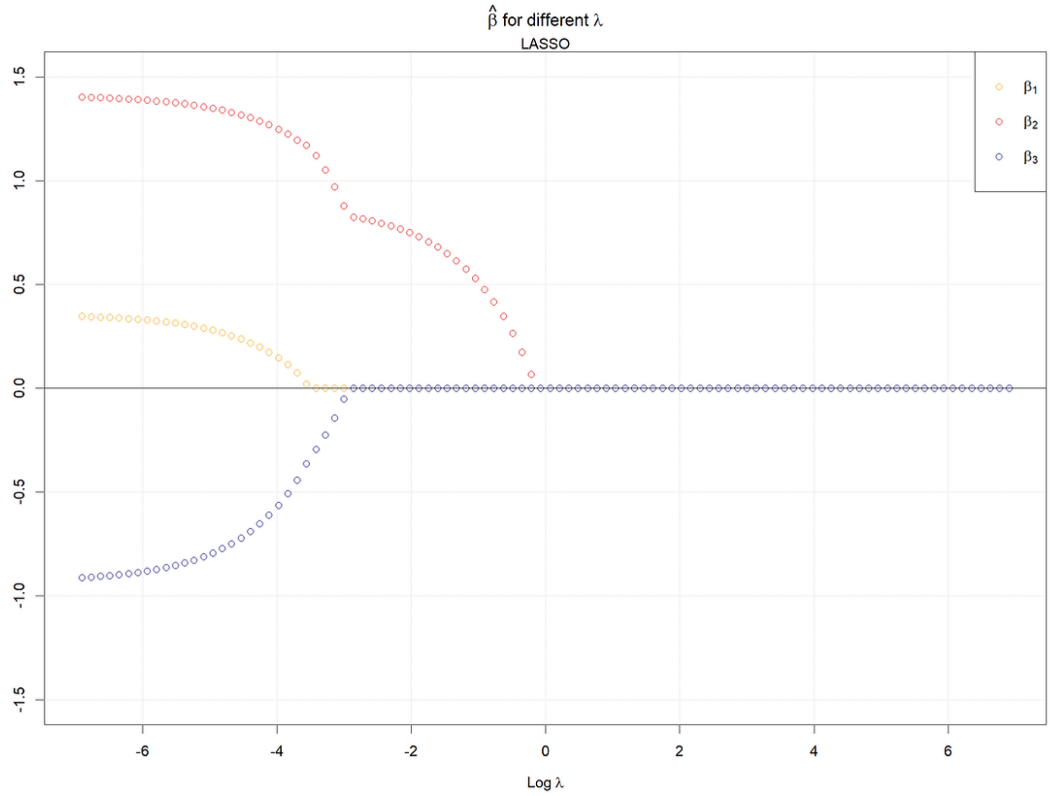


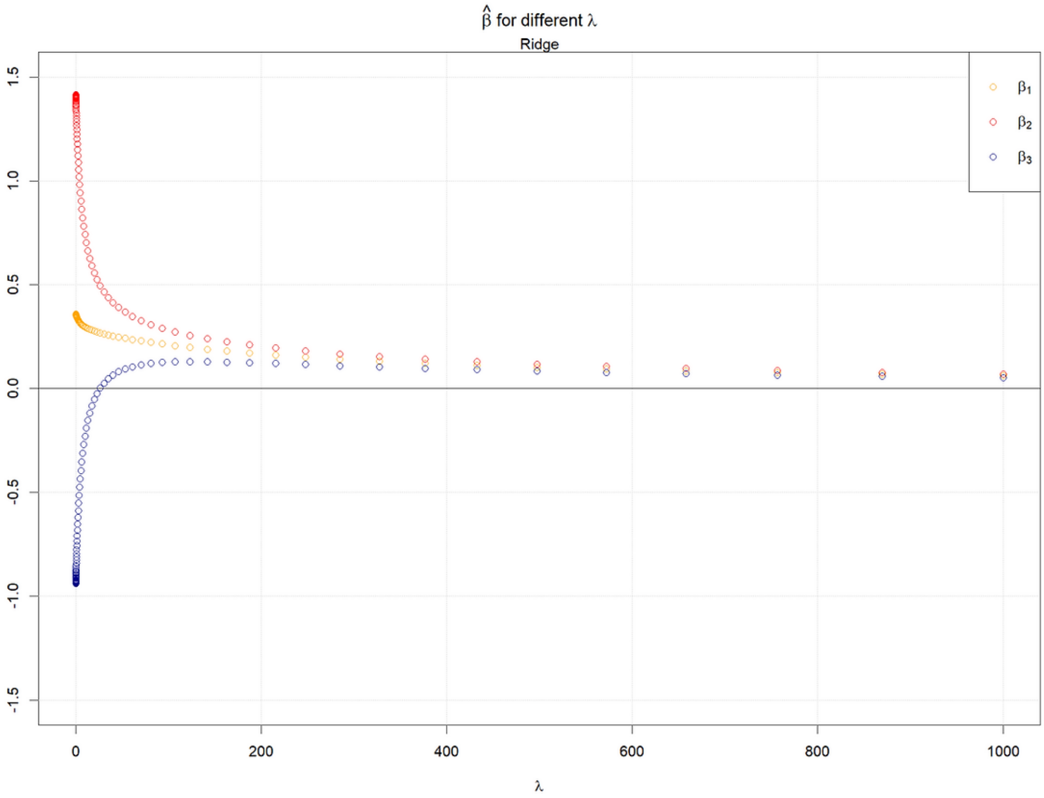
Here we observe the same pattern as in case 3. However, for the Lasso case, we can see that when the coefficient of X3 becomes 0, the path of X1 changes its decrease pattern, because when X3 plays no role, the highly correlated X1 takes more influence, then we see the steep decrease pattern suddenly becomes to a smother pattern.

Case 5: All variables are highly correlated



```
> sigma5
     [,1] [,2] [,3]
[1,]  2.0  1.9  1.9
[2,]  1.9  2.0  1.9
[3,]  1.9  1.9  2.0
```





As in case 4, we can see the same pattern, the kink also exists in the Lasso case.

In conclusion, the larger the starting points of coefficients (refers to 0 lambda), the later the coefficients shrink to 0 or small values. Meanwhile, for highly correlated variables, the Ridge will make them converge to the same value as lambda increases, and in the Lasso case, one of the set of highly correlated variables will shrink to 0 firstly and quickly, at the same time a kink will appear at the paths left.