sealasso: an R package for Standard Error Adjusted Adaptive Lasso (SEA-lasso)

Wei Qian

11 December 2013

1 Introduction

The adaptive lasso (Zou, 2006) is a variable selection method that enjoys the oracle property (consistency in variable selection and asymptotic normality in coefficient estimation). It is now known that the weight applied to the l_1 penalty term of the adaptive lasso under linear regression settings can significantly influence its variable selection performance. Qian and Yang (2010) proposed two versions of the adaptive lasso named SEA-lasso and NSEA-lasso that incorporate the OLS-standard errors to the l_1 penalty term. It is shown that they outperform the usually used weight with OLS estimate only (we call the latter weight selection method OLS-adaptive lasso), especially when the condition index of the model matrix is large.

In practice, it is advisable that NSEA-lasso is used when the condition index is greater than 10 (more specifically, the condition index is defined as the natural logarithm of the ratio of the largest eigenvalue to the smallest eigenvalue for the matrix X^TX, where X is the scaled predictor matrix). In fact, NSEA-lasso is the default method in `sealasso` package. It should be pointed out that this package applies only to linear regression setting, and the number of predictors must be greater than 1 and less than the sample size. The rest of this vignette will use `diabetes` data example to introduce the two functions `sealasso` and `summary` provided in this package.

2 Diabetes data example

The `sealasso` package depends upon `lars` package, which can provide the lasso solution path by LARS algorithm (Efron et al., 2004). The `diabetes` dataset contained in `lars` package has one response and ten baseline predictors measured on 442 diabetes patients. These baseline predictors include age, sex, body mass index (bmi), average blood pressure (bp) and six blood serum measurements (tc, ldl, hdl, tch, ltg, glu).

To begin with, we only consider the main effects, and intend to find important terms from these ten predictors. With SEA-lasso method, we can obtain the following output for the predictor matrix `x` and the response vector `y`.

```
> library(sealasso)
> data(diabetes)      # use the diabetes dataset from "lars" package
> x <- diabetes$x
> y <- diabetes$y
> sealasso(x, y, method = "sealasso")
```

```
$call
sealasso(x = x, y = y, method = "sealasso")

$method
[1] "SEA-lasso"

$weight
 [1] 5.9607284 0.2549898 0.1278404 0.2014430 0.5253846 0.7103195 2.1009190
 [8] 0.9109049 0.2285473 0.9746014

$condition.index
[1] 6.2

$path
      Path Df      BIC
 [1,]    3  1 8.355525
 [2,]    9  2 8.236005
 [3,]    4  3 8.083208
 [4,]    2  4 8.083045
 [5,]    5  5 8.072306
 [6,]    8  6 8.057883
 [7,]   10  7 8.063302
 [8,]    6  8 8.069325
 [9,]    7  9 8.082954
[10,]    1 10 8.096281

$beta
         age       sex      bmi      map         tc      ldl       hdl      tch
1    0.00000   0.00000 74.44217  0.00000    0.00000   0.0000   0.00000   0.0000
2    0.00000   0.00000 78.16029  0.00000    0.00000   0.0000   0.00000   0.0000
3    0.00000   0.00000 77.31784 41.86987    0.00000   0.0000   0.00000   0.0000
4    0.00000 -14.59779 76.96890 49.04815    0.00000   0.0000   0.00000   0.0000
5    0.00000 -26.11275 76.87790 55.19241  -56.45724   0.0000   0.00000   0.0000
6    0.00000 -42.89987 73.25861 61.73496 -112.61844   0.0000   0.00000 136.6062
7    0.00000 -51.96574 70.72512 63.55161 -141.83001   0.0000   0.00000 200.0832
8    0.00000 -61.12065 66.59269 64.69064 -314.72129 235.9168   0.00000 121.0733
9    0.00000 -61.16463 66.57327 64.72640 -330.02183 251.2537  31.79894 127.2319
10 -59.67999 -61.15142 66.45651 65.34618 -416.20137 338.6419 212.28646 161.2886
         ltg      glu
1    0.00000  0.00000
2   44.06475  0.00000
3  107.65071  0.00000
4  115.63090  0.00000
5  133.32481  0.00000
6  130.32976  0.00000
7  127.06532 37.31186
```

```
8  156.22248 63.70201
9  158.51067 63.82147
10 171.70290 65.90780

$optim.beta
  age       sex      bmi      map        tc ldl hdl      tch      ltg glu
6   0 -42.89987 73.25861 61.73496 -112.6184   0   0 136.6062 130.3298   0

attr(,"class")
[1] "sealasso"
```

From the output above, we can find the weight selection method used, weight vector for the l_1 penalty term, the condition index, the solution path (with estimated degree of freedom and BIC values at the transition points), estimated coefficients at the transition points, and the optimal model selected by BIC. If we want to change the method to be NSEA-lasso, OLS-adaptive lasso or the lasso, simply replace `"sealasso"` with `"nsealasso"`, `"olsalasso"`, and `"lasso"`, respectively. Also note that the leading number under the name `optim.beta` is the step number of the optimal model in the solution path.

A `summary` function is available to print out a more succinct output, which includes only the weight selection method used, the condition index and the optimal model. Again, we use the `diabetes` dataset as an example. Besides the main effect predictors, we can also include the quadratic terms and interaction terms to generate an expanded predictor matrix `x2`. This expanded predictor matrix contains 64 predictors, including 10 baseline predictors, 45 interactions and 9 squares. The square term of sex is not included because it is a dichotomous variable. The `summary` output for the default NSEA-lasso method and the expanded matrix `x2` is shown as follows.

```
> # with quadratic terms
> x2 <- cbind(diabetes$x1,diabetes$x2)
> object <- sealasso(x2, y)
> summary(object)

$method
[1] "NSEA-lasso"

$condition.index
[1] 17.2

$optim.beta
  age       sex      bmi      map tc ldl       hdl tch      ltg glu age^2 bmi^2
8   0 -58.85721 80.74083 64.98872  0   0 -15.40572   0 17.95601   0     0     0
  map^2 tc^2 ldl^2 hdl^2 tch^2 ltg^2    glu^2  age:sex age:bmi age:map age:tc
8     0    0     0     0     0     0 52.22663 74.84565       0       0      0
  age:ldl age:hdl age:tch age:ltg age:glu sex:bmi sex:map sex:tc sex:ldl
8       0       0       0       0       0       0       0      0       0
  sex:hdl sex:tch sex:ltg sex:glu  bmi:map bmi:tc bmi:ldl bmi:hdl bmi:tch
8       0       0       0       0 59.16698      0       0       0       0
  bmi:ltg bmi:glu map:tc map:ldl map:hdl map:tch map:ltg map:glu tc:ldl tc:hdl
```

```
8       0       0      0       0       0       0       0       0      0      0
  tc:tch tc:ltg tc:glu ldl:hdl ldl:tch ldl:ltg ldl:glu hdl:tch hdl:ltg hdl:glu
8      0      0      0       0       0       0       0       0       0       0
  tch:ltg tch:glu ltg:glu
8       0       0       0
```

As a last note, we should point out that `sealasso` function automatically standardizes the model matrix so that for each column vector, the mean is 0 and l_2 norm is $\sqrt{n}$ before applying the corresponding adaptive lasso method, while the estimated coefficients are transformed back to the original scale for the output. Therefore, in practice, we do not need to standardize the model matrix for using `sealasso` function.

References

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32, 407-499.

Qian, W. and Yang, Y. (2010). Variable Selection via Standard Error Adjusted Adaptive Lasso. Technical Report, University of Minnesota.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 4225–4242.