Data-driven Lambda for the LASSO

Shuofan Zhang

January 25, 2019

In this report, only the significant lags chosen by AR were included in the X matrix.

Since the plots for yr2, yr3, yr4 and yr5 are similar, so here I just plotted yr2.

The M.S.E of the training dataset increases with λ, so it is not feasible to choose a λ according to the performance in training dataset.

Two solutions:

1, Same as before, separate the data set into training set (80%) and test set (20%); then choose λ according to the smallest M.S.E in the test set directly and use it for the comparison.

2, Separate the data set into three sets, training (60%), validation (20%) and test (20%); then choose λ according to the smallest M.S.E in the validation set, use the M.S.E from the test set for comparison.

(Continued from 12th Dec):

1, When we have high-dimentional data, and set *singular.ok = TRUE*, then run OLS with *lm* function in R, the x matrix will be truncated to have an appropriate number of columns. Example: n=100, p=200, the 100^{th}-200^{th} columns will be removed before running the regression. The actual data used in the regression will be n=100, p=99.

2, When we have high-dimentional data, and set λ as a series of values including zero, then run LASSO with glmnet function in R, it gives non-zero estimations to every variables (when $\lambda = 0$). In fact, the function does not allow $\lambda = 0$ so it uses some very small value instead. (verified by data experiments as shown in the scatter plot, when the smallest lambda is smaller than 1E-6, the scatter plot almost reduce to a straight line, code check is ongoing)

3, When the data is not high-dimentional, the estimations of LASSO (when $\lambda = 0$) becomes closer and closer to the OLS estimations with decreasing "thresh" values.

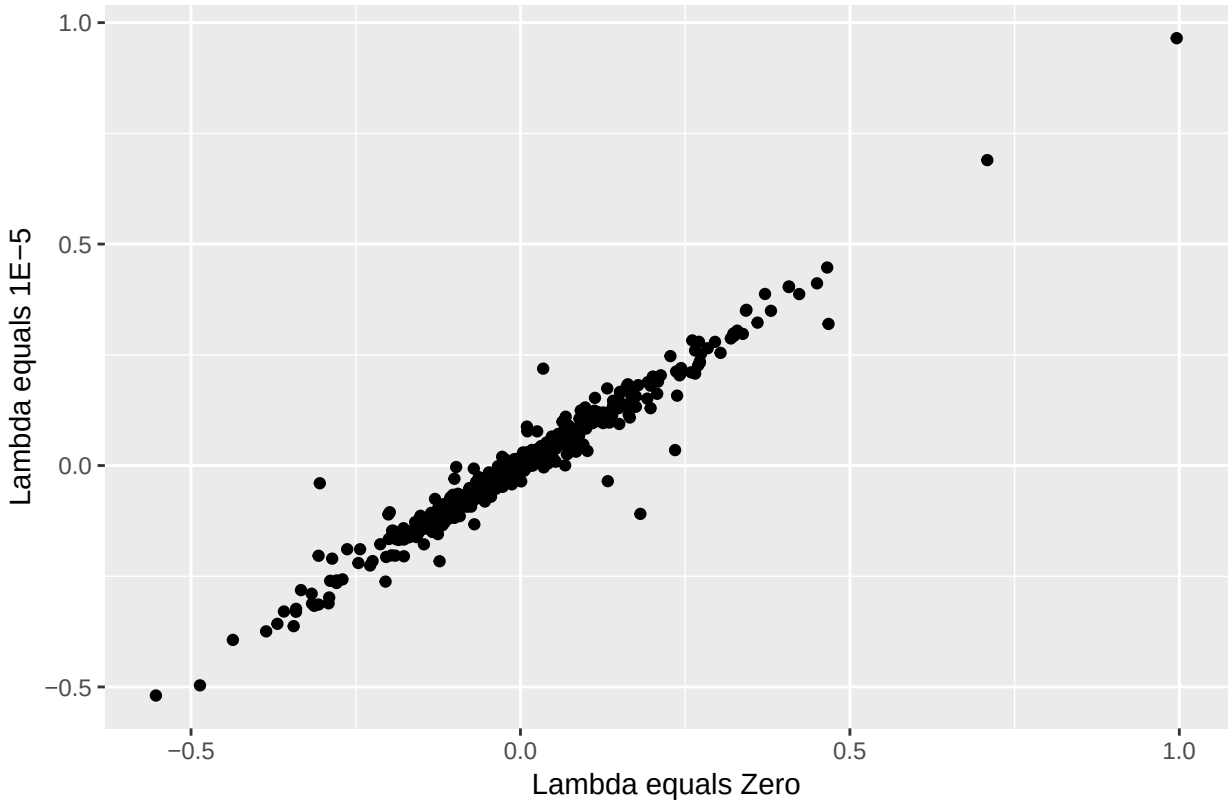
some references:

How Correlations Influence Lasso Prediction

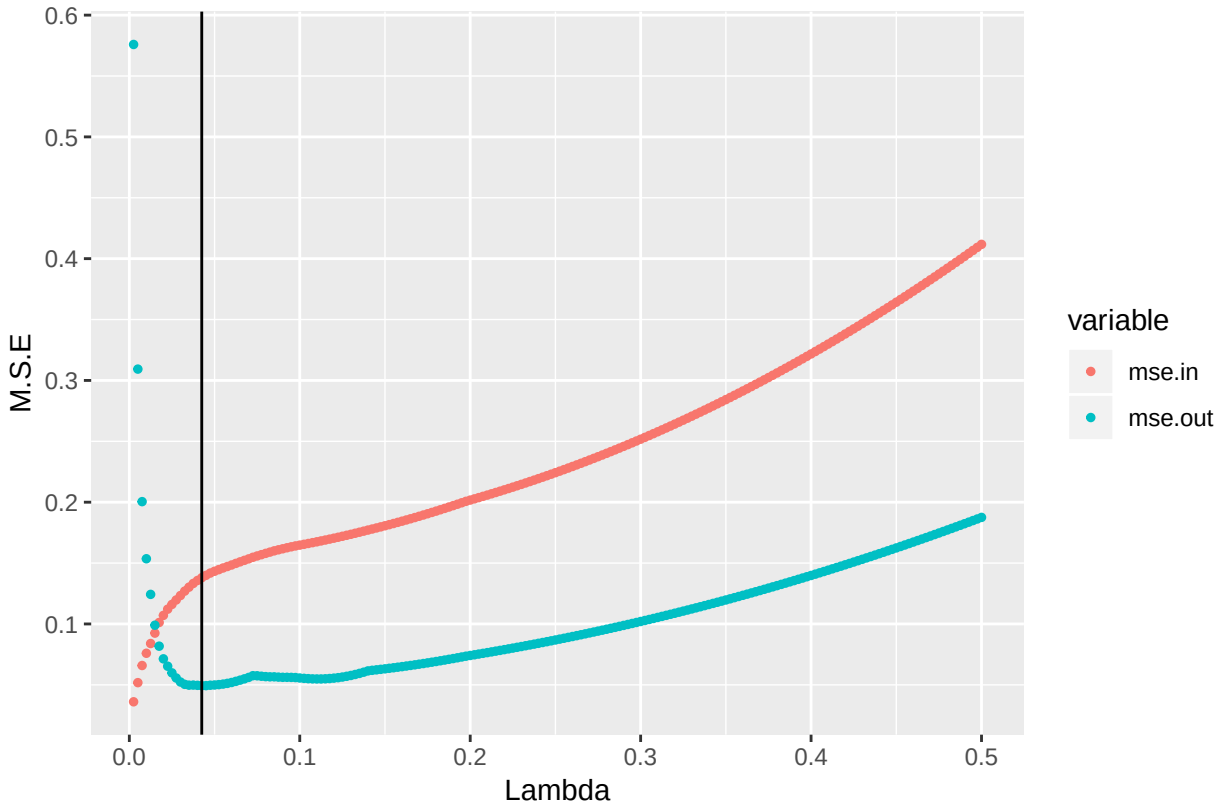
This paper argues correlation is not problematic for LASSO prediction, but it has influence on suitable λ. The higher the correlation, the smaller the tuning parameter.

- LASSO 1

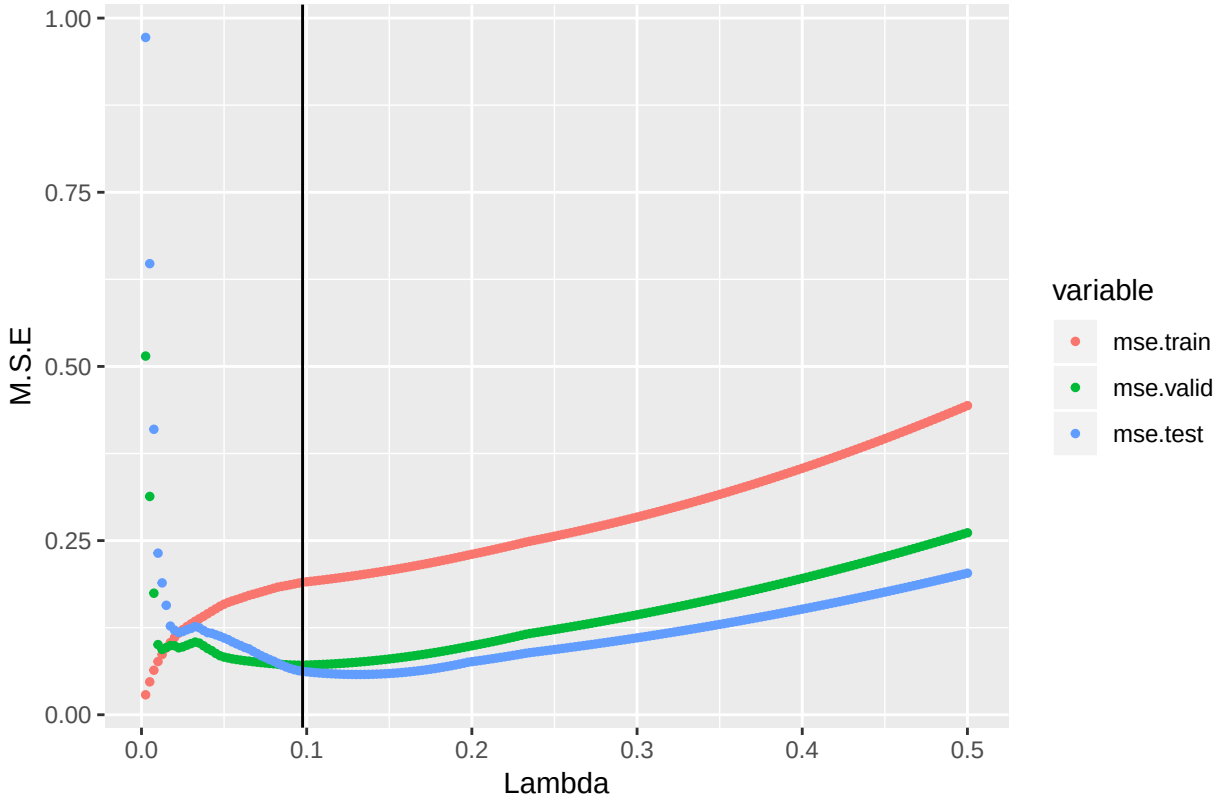
The explanatory variables in this model are level I(1) variables, level I(0) variables, first-differenced I(2) variables and lags of the dependent variables.



Scatter plot of two sets of coefficients of LASSO when lambda equals to zero and 1E−5



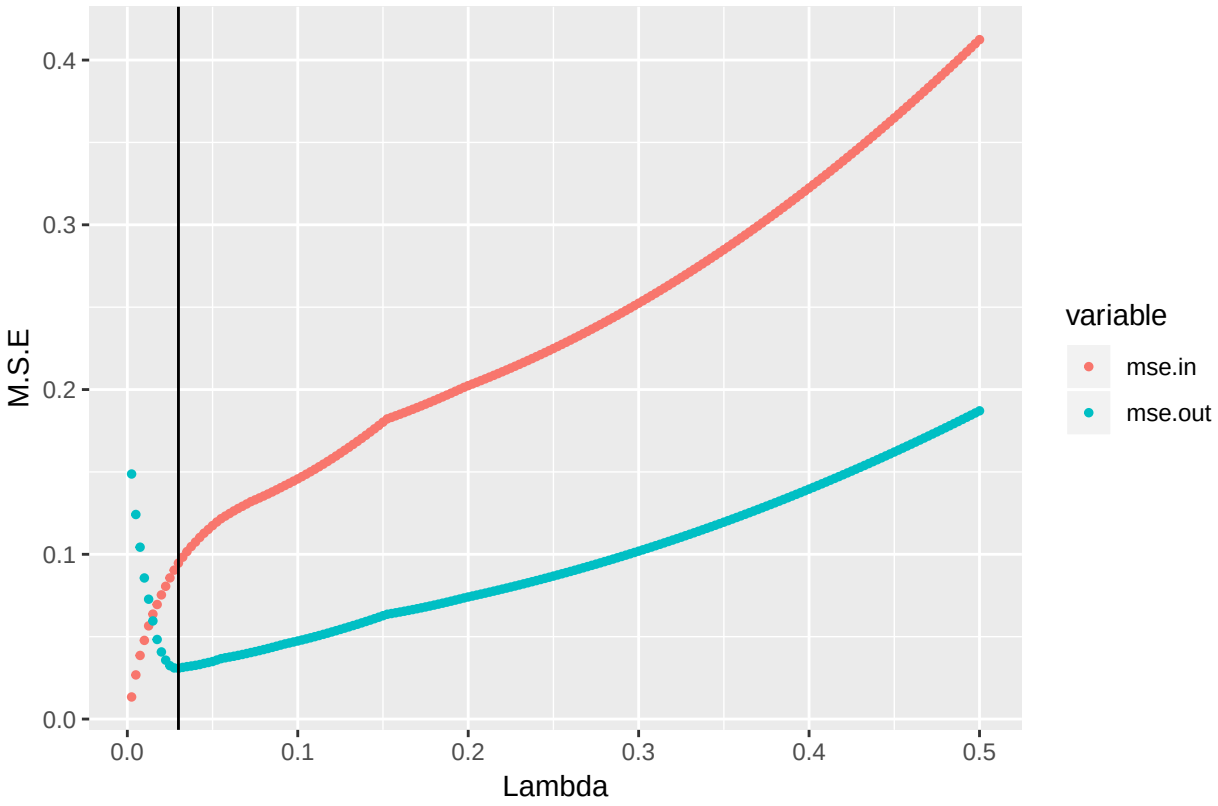
Path of M.S.E in two datasets against lambda, LASSO 1



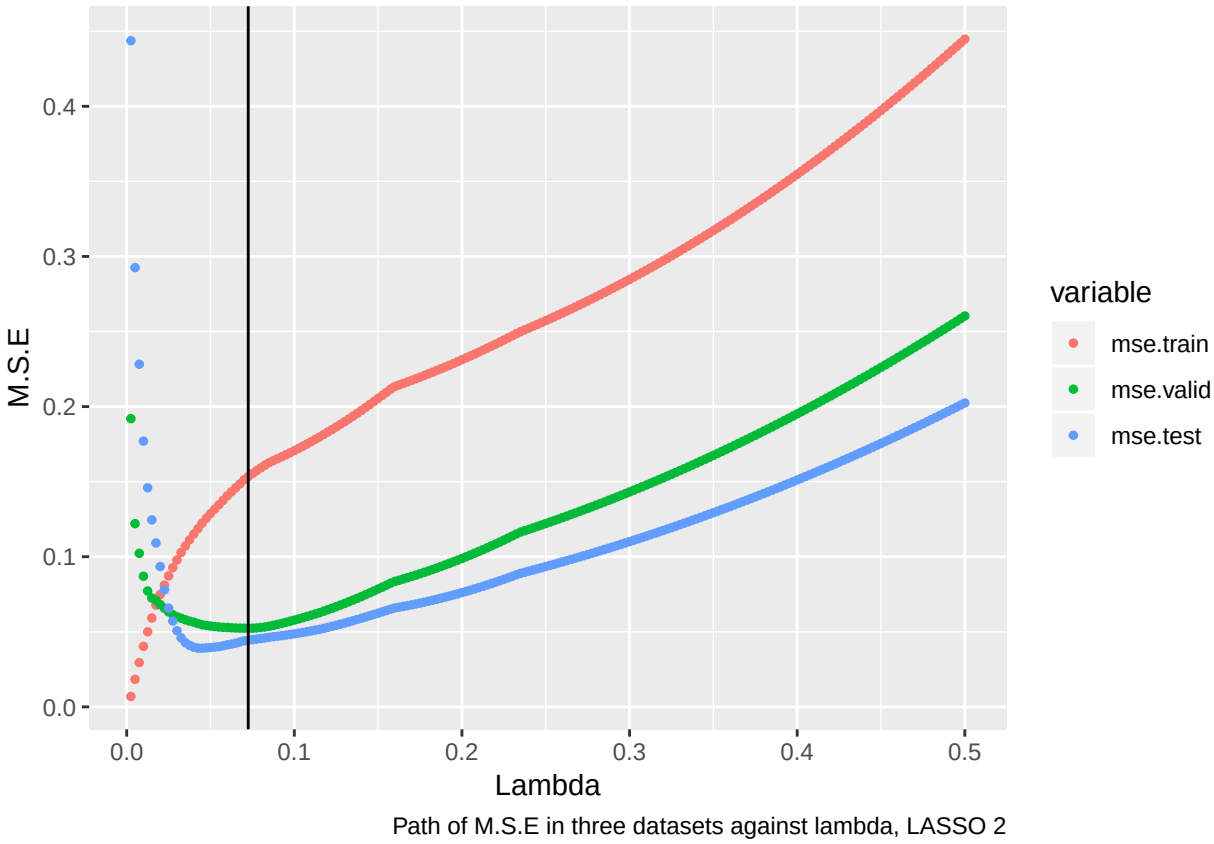
Path of M.S.E in three datasets against lambda, LASSO 1

- LASSO 2

The explanatory variables in this model are first-differenced I(1) variables, level I(0) variables, twice-differenced I(2) variables and lags of the dependent variables.



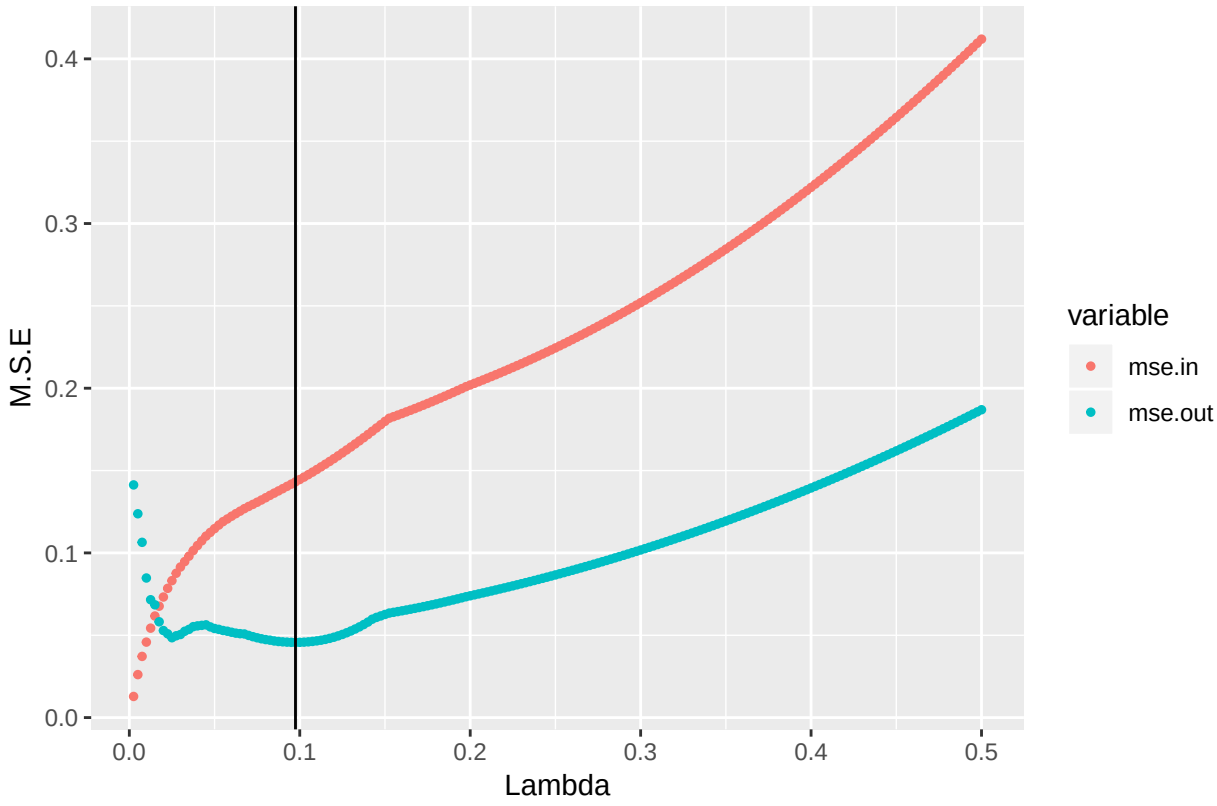
Path of M.S.E in two datasets against lambda, LASSO 2



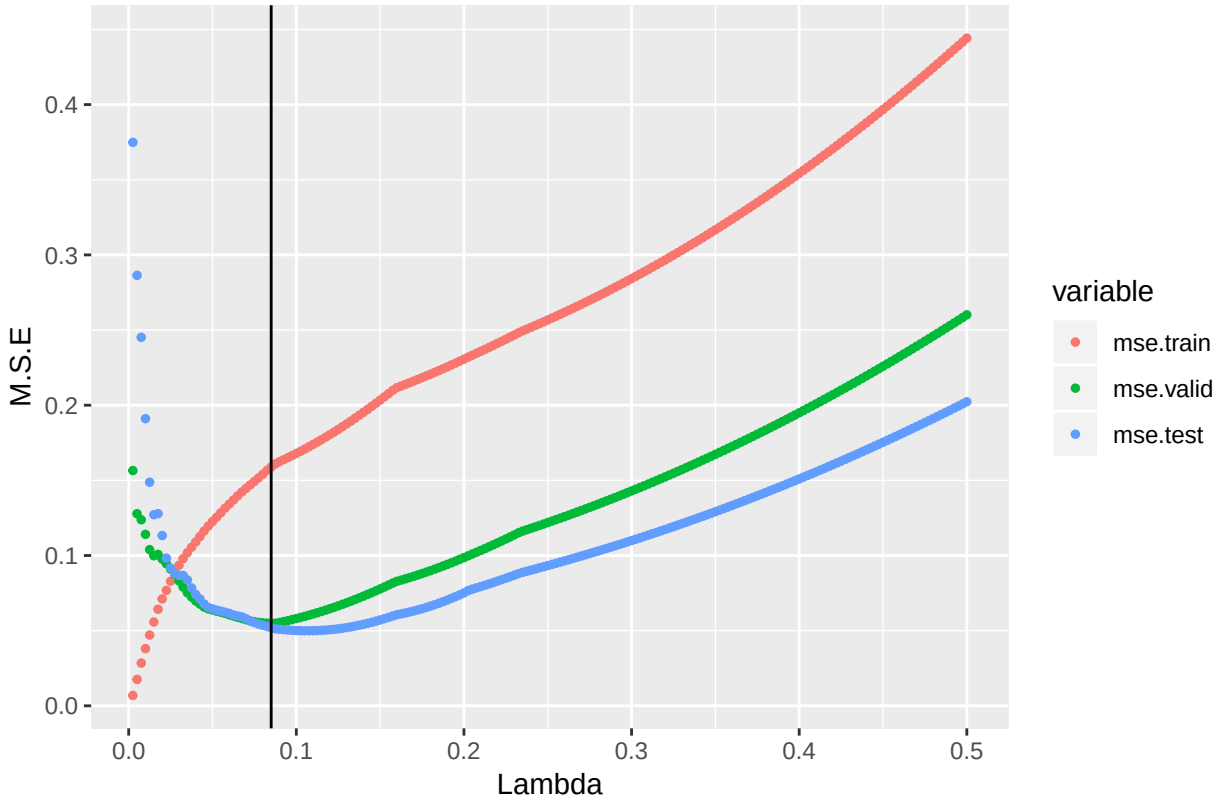
Path of M.S.E in three datasets against lambda, LASSO 2

- LASSO 3

The explanatory variables in this model are first-differenced I(1) variables, level I(0) variables, twice-differenced I(2) variables,level I(1) variables, first-differenced I(2) variables and lags of the dependent variables.



Path of M.S.E in two datasets against lambda, LASSO 3

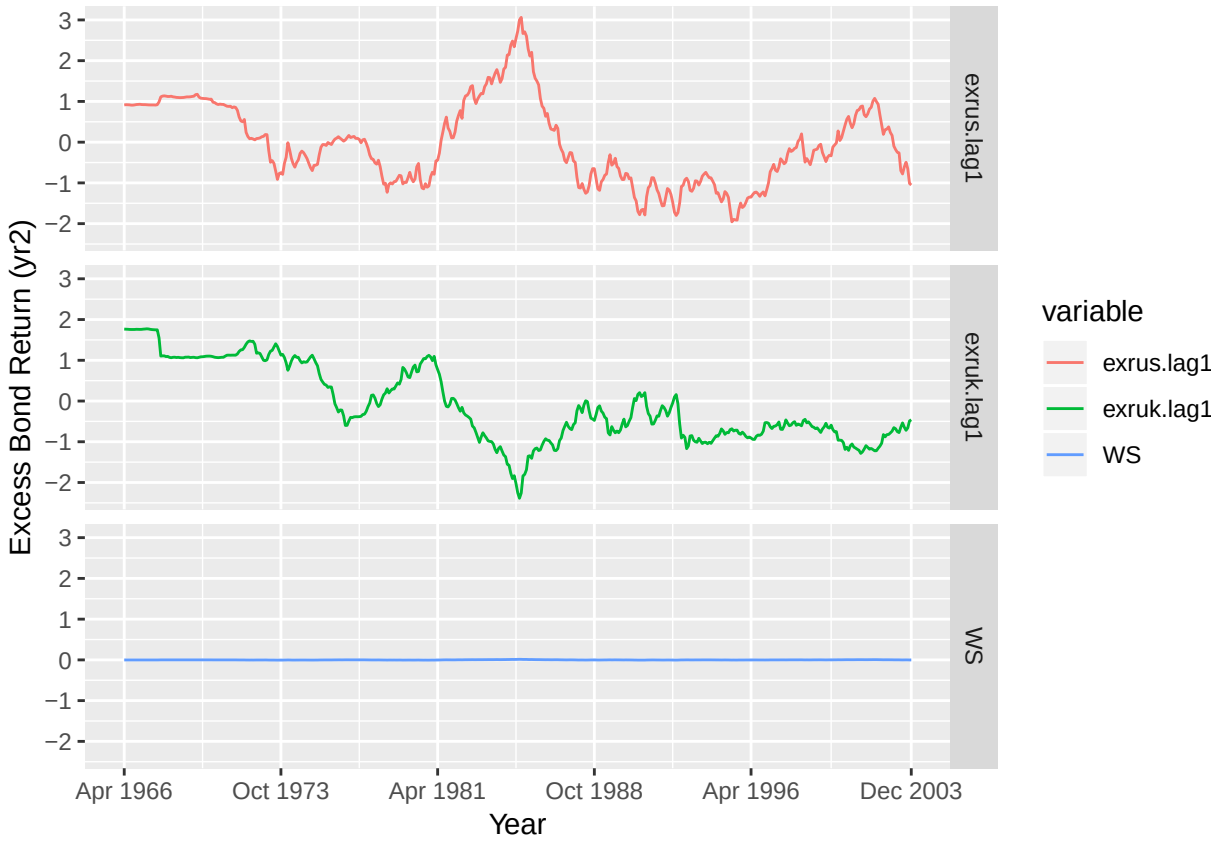


Path of M.S.E in three datasets against lambda, LASSO 3

Table 1: Variables selected by the LASSO with data driven lambda (valid) for yr2

variable	LASSO 1	LASSO 2	LASSO 3
ces015.lag1	-0.06302	NA	-0.081
pmcp.lag1	-0.08144	-0.07776	-0.06601
D.gmdcn.lag5	-3e-06	NA	NA
fclbmc.lag12	0.00039	NA	NA
ytem.lag1	0.7669	0.8223	0.795
ytem.lag5	NA	0.009655	NA
D.ypr.lag1	NA	-0.003463	-0.000973
D.a0m051.lag1	NA	-0.01647	-0.004059
pmcp.lag5	NA	-0.01831	-0.001411
D.fygt10.lag5	NA	-0.003416	-0.001061
D.fygt1.lag12	NA	0.001242	1.4e-05
D.fygt5.lag12	NA	0.04613	0.04731
D.fygt10.lag12	NA	0.04782	0.03405
D.fyff.lag13	NA	-0.004052	NA
exrus.lag1	NA	NA	0.002564
exruk.lag1	NA	NA	-0.002343

Graphs of the two sets of potential co-integrating vectors.



Model Confidence Set

```
##
## Model L1_valid eliminated 2019-01-25 12:28:24
## Model L1_valid_all eliminated 2019-01-25 12:28:37
## ##############################################################################################################
## Superior Set Model created   :
##              Rank_M        v_M  MCS_M Rank_R        v_R  MCS_R       Loss
## L1_test           9  1.1290306 0.7820     11  3.2915306 0.0136 0.04929423
## L2_test           2 -3.5395993 1.0000      1 -0.6686474 1.0000 0.03080323
## L2_valid          5 -0.1096763 1.0000      9  3.1910667 0.0174 0.04444512
## L3_test           7  0.3792867 0.9980      8  3.0336109 0.0290 0.04567235
## L3_valid         10  1.3434476 0.6248      6  2.7912930 0.0594 0.05180697
## L1_test_all      12  1.9470483 0.2336     12  3.6343405 0.0050 0.04960958
## L2_test_all       1 -3.8169600 1.0000      2  0.6686474 0.9976 0.03209796
## L2_valid_all      8  0.6866800 0.9654     10  3.2292734 0.0156 0.04695511
## L3_test_all       6  0.2819309 0.9996      7  2.9539224 0.0594 0.04548748
## L3_valid_all     11  1.4615826 0.5398      5  2.7775475 0.3302 0.05350068
## AR                4 -0.2500794 1.0000      3  1.9895810 0.4032 0.04356516
## AR_all            3 -0.2521589 1.0000      4  2.1099687 0.3302 0.04361781
## p-value  :
## [1] 0.2336
##
## ##############################################################################################################

##
## ------------------------------------------
## -          Superior Set of Models        -
```

```
## ----------------------------------------
##              Rank_M        v_M  MCS_M Rank_R        v_R  MCS_R       Loss
## L1_test           9  1.1290306 0.7820     11  3.2915306 0.0136 0.04929423
## L2_test           2 -3.5395993 1.0000      1 -0.6686474 1.0000 0.03080323
## L2_valid          5 -0.1096763 1.0000      9  3.1910667 0.0174 0.04444512
## L3_test           7  0.3792867 0.9980      8  3.0336109 0.0290 0.04567235
## L3_valid         10  1.3434476 0.6248      6  2.7912930 0.0594 0.05180697
## L1_test_all      12  1.9470483 0.2336     12  3.6343405 0.0050 0.04960958
## L2_test_all       1 -3.8169600 1.0000      2  0.6686474 0.9976 0.03209796
## L2_valid_all      8  0.6866800 0.9654     10  3.2292734 0.0156 0.04695511
## L3_test_all       6  0.2819309 0.9996      7  2.9539224 0.0594 0.04548748
## L3_valid_all     11  1.4615826 0.5398      5  2.7775475 0.3302 0.05350068
## AR                4 -0.2500794 1.0000      3  1.9895810 0.4032 0.04356516
## AR_all            3 -0.2521589 1.0000      4  2.1099687 0.3302 0.04361781
##
## Details
## ----------------------------------------
##
## Number of eliminated models  :   2
## Statistic    :   Tmax
## Elapsed Time :   Time difference of 40.46658 secs
```