STAT406- Methods of Statistical Learning Lecture 7

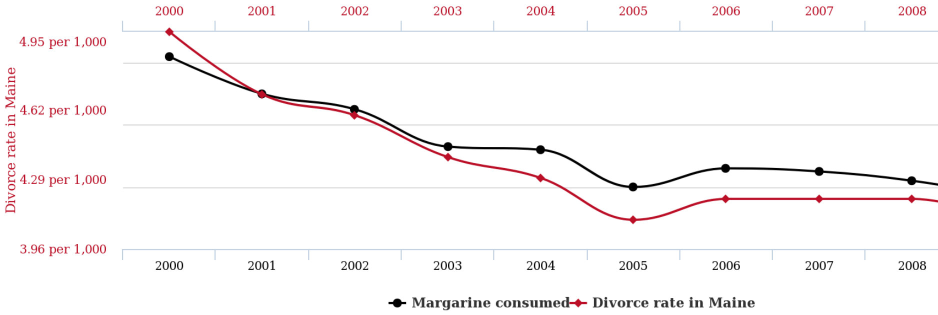
Matias Salibian-Barrera

UBC - Sep / Dec 2019

Divorce rate in Maine

correlates with

Per capita consumption of margarine



Correlation: 99.26%

http://www.tylervigen.com/spurious-correlations

Model / feature selection - LASSO

- Another regularized method is given by LASSO

$$\min_{\alpha,\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \alpha - \boldsymbol{\beta}' \mathbf{x}_i \right)^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right|$$

$$\min_{\alpha,\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \alpha - \boldsymbol{\beta}' \mathbf{x}_i \right)^2 + \lambda \, \|\boldsymbol{\beta}\|_1$$

for some $\lambda > 0$

Model / feature selection - LASSO

- The above is equivalent to

$$\min_{\alpha,\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \alpha - \boldsymbol{\beta}' \mathbf{x}_i \right)^2$$

subject to

$$\sum_{j=1}^{p} \left| \beta_j \right| \le K$$

for some $K > 0$

LASSO

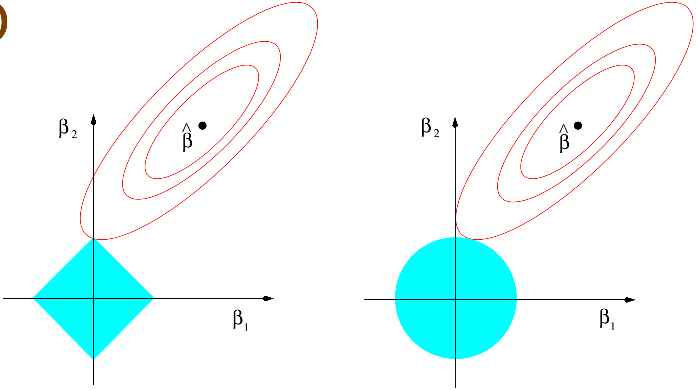
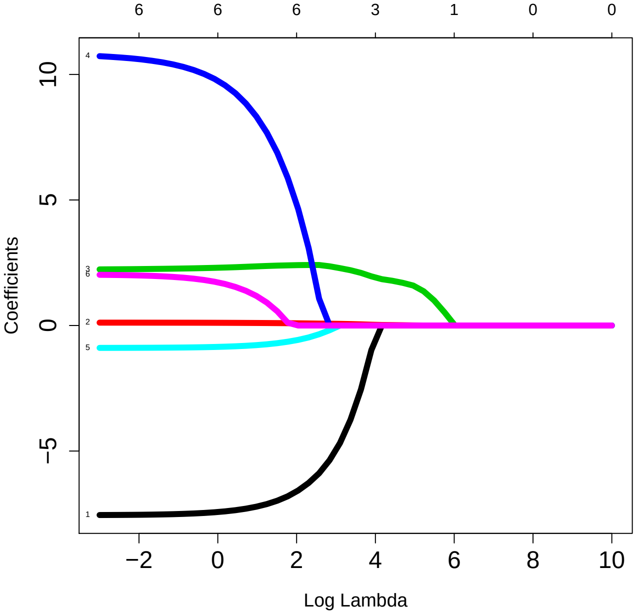


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

Credit data - glmnet output



Credit data - glmnet output

```
a <- glmnet(x=xm, y=yc, lambda=lambdas,
    family='gaussian', alpha=1, intercept=FALSE)

> coef(a, s=1)
7 x 1 sparse Matrix of class "dgCMatrix"
                    1
(Intercept)  .
Income      -7.4285710
Limit        0.1078894
Rating       2.3006418
Cards        9.7499618
Age         -0.8515917
Education    1.7182477
```

Credit data - glmnet output

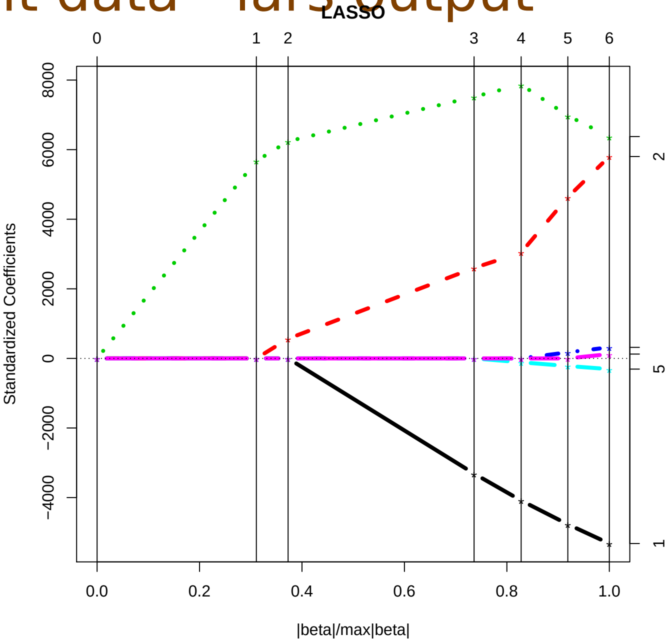
```
> coef(a, s=exp(4))
7 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept)   .
Income       -0.63094341
Limit         0.02749778
Rating        1.91772580
Cards         .
Age           .
Education     .
```

Credit data - another implementation

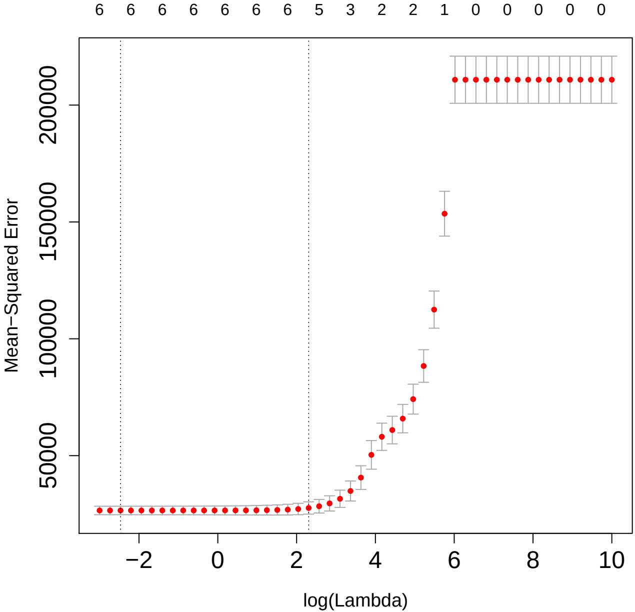
```
> library(lars)
> b <- lars(x=xm, y=yc, type='lasso', intercept=FALSE)
> coef(b)
        Income      Limit   Rating     Cards        Age Education
[1,]  0.000000 0.00000000 0.000000  0.000000  0.0000000  0.000000
[2,]  0.000000 0.00000000 1.835963  0.000000  0.0000000  0.000000
[3,]  0.000000 0.01226464 2.018929  0.000000  0.0000000  0.000000
[4,] -4.703898 0.05638653 2.433088  0.000000  0.0000000  0.000000
[5,] -5.802948 0.06600083 2.545810  0.000000 -0.3234748  0.000000
[6,] -6.772905 0.10049065 2.257218  6.369873 -0.6349138  0.000000
[7,] -7.558037 0.12585115 2.063101 11.591558 -0.8923978  1.998283
> b

Call:
lars(x = xm, y = yc, type = "lasso", intercept = FALSE)
R-squared: 0.878
Sequence of LASSO moves:
     Rating Limit Income Age Cards Education
Var       3     2      1   5     4         6
Step      1     2      3   4     5         6
```

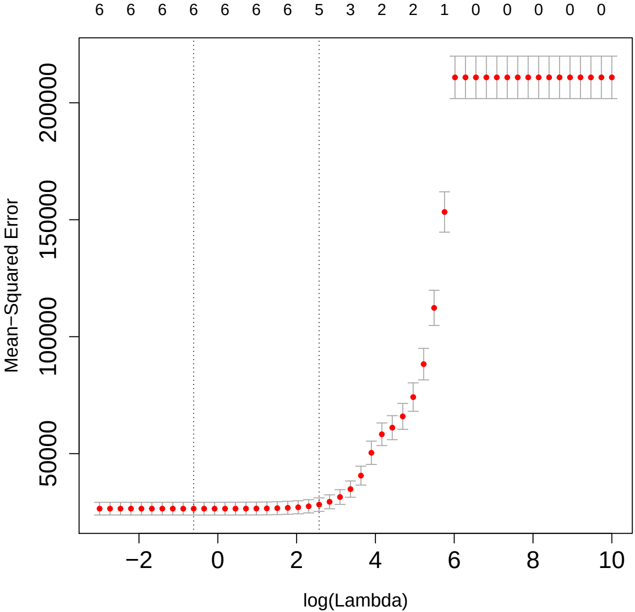
Credit data - lars output



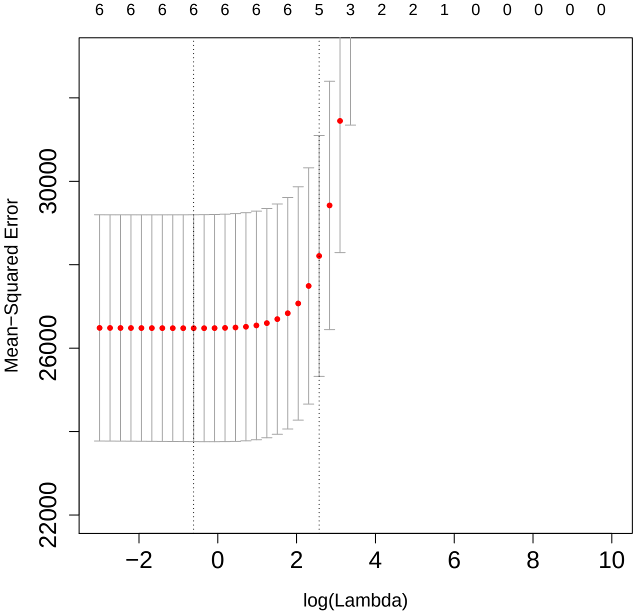
Credit data - CV - glmnet



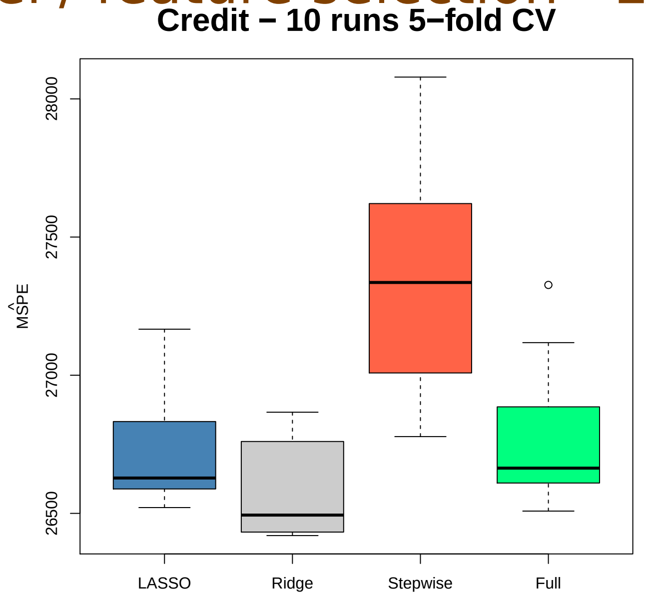
Credit data - CV - another run



Credit data - CV - zoom



Model / feature selection - LASSO



Model / feature selection - LASSO

- Worse estimated MSPE than Ridge Regression in this case
- It provides a sequence of explanatory variables, an ordered set of models
- Much like stepwise, but with better MSPE in this case

Model / feature selection - LASSO

- Why does it work? It is the convex proxy for the "nuclear norm"
- Also generates infinitely many estimates, but there's a clever algorithm
- Inference?

Model / feature selection - LASSO

- When covariates are correlated, LASSO will typically pick any one of them, and ignore the rest
- Ridge Regression, on the other hand, combines the coefficients of correlated covariates, but doesn't provide sparse models

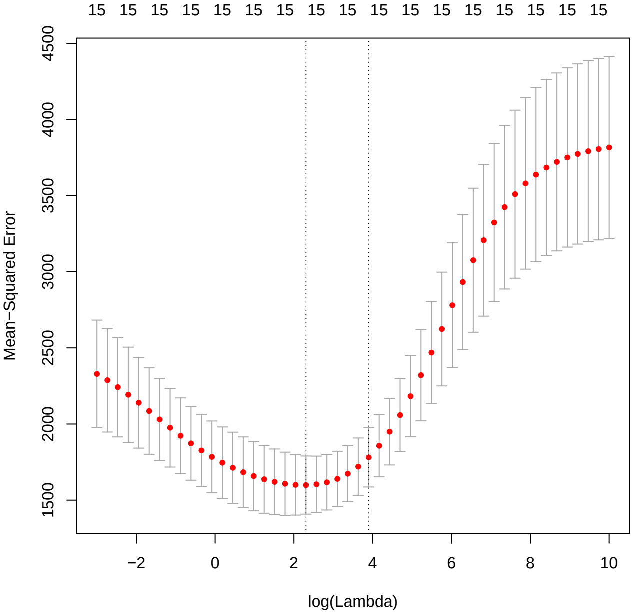
Ridge vs. LASSO

- Compare Ridge and LASSO on the air pollution data

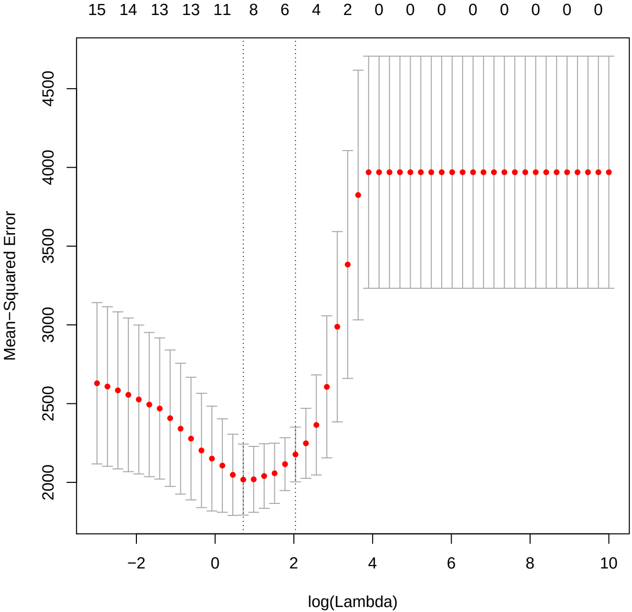
Air pollution example

```
airp <- read.table('..-30861_CSV-1.csv',
    header=TRUE, sep=',')
y <- as.vector(airp$MORT)
xm <- as.matrix(airp[, names(airp) != 'MORT'])
# Ridge
set.seed(123)
air.l2 <- cv.glmnet(x=xm, y=y, lambda=lambdas,
    nfolds=5, alpha=0, family='gaussian',
    intercept=TRUE)
# LASSO
set.seed(23)
air.l1 <- cv.glmnet(x=xm, y=y, lambda=lambdas,
    nfolds=5, alpha=1, family='gaussian',
    intercept=TRUE)
```

Air pollution - Ridge



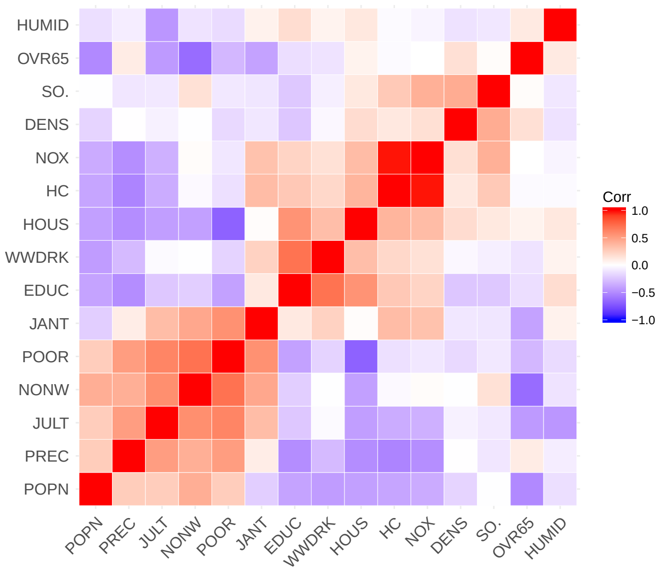
Air pollution - LASSO



Air pollution example

	Ridge	LASSO
(Intercept)	1179.335	1100.355
PREC	1.570	1.503
JANT	-1.109	-1.189
JULT	-1.276	-1.247
OVR65	-2.571	.
POPN	-10.135	.
EDUC	-8.479	-10.510
HOUS	-1.164	-0.503
DENS	0.005	0.004
NONW	3.126	3.979
WWDRK	-0.476	-0.002
POOR	0.576	.
HC	-0.035	.
NOX	0.064	.
SO.	0.240	0.228
HUMID	0.372	.

Air pollution - Correlations



Model / feature selection - LASSO

- Oracle - consistency
- Problem: when $n < p$, LASSO will only choose up to n variables
- When covariates are correlated, LASSO will typically pick any one of them, and ignore the rest
- Ridge Regression, on the other hand, combines the coefficients of correlated covariates, but doesn't provide sparse models

Elastic Net

- Elastic Net is a compromise between the two:

$$\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \boldsymbol{\beta}' \mathbf{x}_i \right)^2 + \lambda \left[\alpha \|\boldsymbol{\beta}\|_1 + \frac{(1-\alpha)}{2} \|\boldsymbol{\beta}\|_2^2 \right]$$

for some $\lambda > 0$ and $0 \leq \alpha \leq 1$.

Elastic Net

- $\alpha = 0$ reduces to Ridge Regression
- $\alpha = 1$ reduces to LASSO
- α needs to be chosen... how would you find a good choice for α?

Air pollution example

- There are correlated covariates
- LASSO solution picks one of each group early on and relegates the rest to the end of the sequence
- Ridge Regression includes all variables always
- EN with $\alpha = 0.10$ gives a nice path of solutions...
- CV? bivariate search, unless α can be chosen beforehand