

ADDITIONAL TOPICS ON THE LASSO

-STATISTICAL MACHINE LEARNING-

Lecturer: Darren Homrighausen, PhD

ℓ_1 -REGULARIZED REGRESSION

REMINDER Known as

- ‘lasso’
- ‘basis pursuit’

The estimator satisfies

$$\hat{\beta}_{\text{lasso},t} = \underset{\|\beta\|_1 \leq t}{\operatorname{argmin}} \|\mathbb{Y} - \mathbb{X}\beta\|_2^2$$

In its corresponding Lagrangian dual form:

$$\hat{\beta}_{\text{lasso},\lambda} = \underset{\beta}{\operatorname{argmin}} \|\mathbb{Y} - \mathbb{X}\beta\|_2^2 + \lambda \|\beta\|_1$$

SOME ADDITIONAL TOPICS

1. GRIDS AND CROSS-VALIDATION
2. SPARSE MATRICES: In some cases, most of entries in \mathbb{X} are zero and hence we can store/manipulate \mathbb{X} much cheaper using **sparse matrices**
3. ELASTIC NET: For use when covariates are highly **related** to each other
4. REFITTED LASSO: A proposal for **reducing** the lasso bias
5. SCALED SPARSE REGRESSION

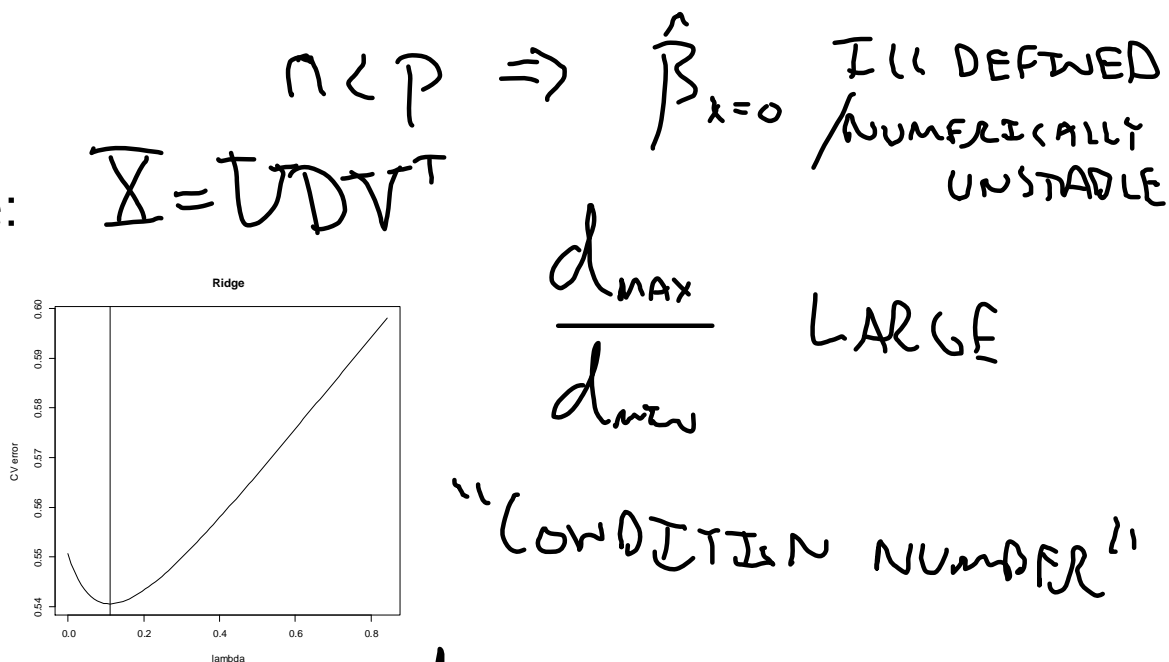
Grids and cross-validation

SOME COMMENTS ABOUT GLMNET

(NOTE: This section's examples are in terms of Ridge regression. There are the same problems with lasso and elastic net. I just am picking one for simplicity)

Some further details

- Note that in this figure:



many solutions have **almost** the same CV error

In fact, since CV is a risk estimate, it is **random**

- The lower end point of the grid is somewhat arbitrary chosen

SOME COMMENTS ABOUT GLMNET

The way that `glmnet` works is to

1. form a `grid` of λ values,
2. find the cross-validation error for each ridge solution on that grid
3. compute the minimum cross-validated λ : $\hat{\lambda}$
4. report $\hat{\beta}_{\text{ridge}, \hat{\lambda}}$ as the final solution

The important piece is that the final solution `depends` on which grid we choose

The function `cv.glmnet` comes with a `plotting` function

```
ridge.cv = cv.glmnet(x=X,y=Y,alpha=0)
plot(ridge.cv)
```

SOME COMMENTS ABOUT GLMNET

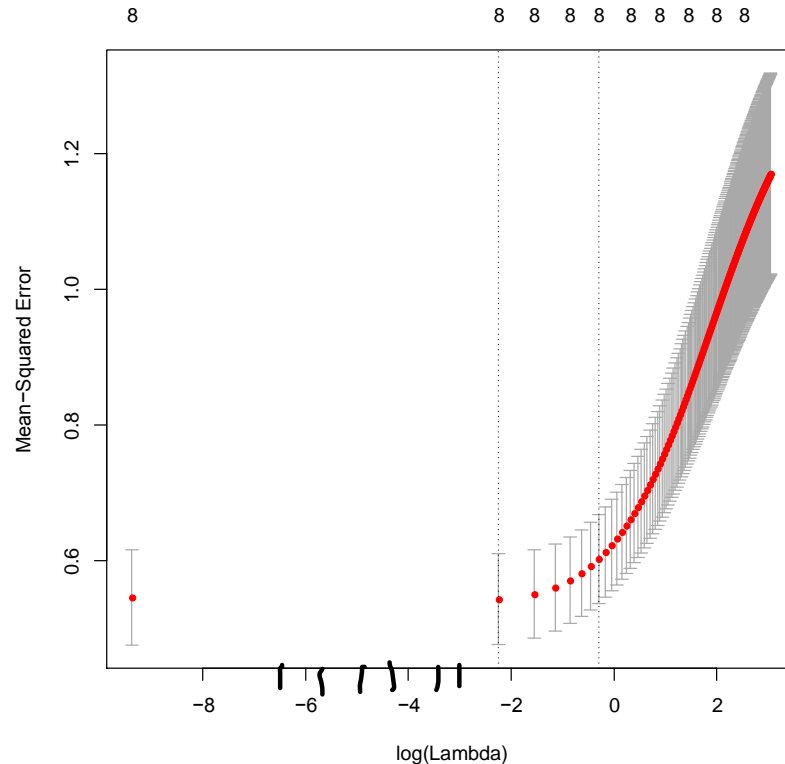
LASSO:

λ LARGER

\Rightarrow MORE MODEL
SELECTION

($\|\hat{\beta}_{\lambda}\|_1$ smaller)

AIC/CV OVERSELECT
FEATURES



- The left-most dotted, vertical line occurs at the CV minimum
- The right-most dotted, vertical line is the
 - ▶ largest value of λ ...
 - ▶ such that the error is within one standard-error of the minimum

(the so called **one-standard-error** rule)

SOME COMMENTS ABOUT GLMNET

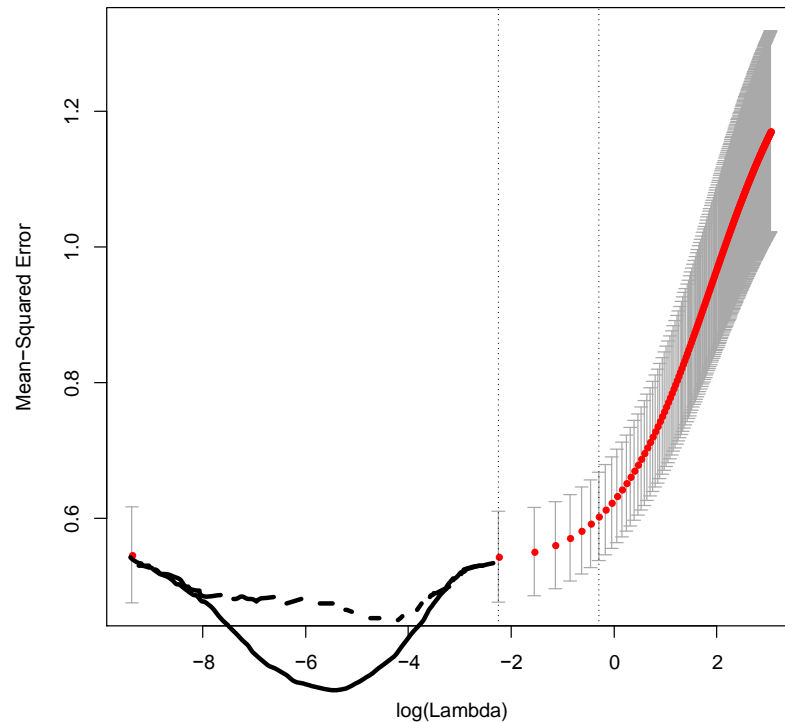
Though `glmnet` automatically allocates a grid, it isn't necessary any good

Sometimes...

- the grid values are too far apart near the minimum
- the grid doesn't allow small/large enough λ values

SOME COMMENTS ABOUT GLMNET

glmnet plots "df" for each $\hat{\beta}_\lambda$
 TOTALLY WRONG FOR RIDGE, actually $\text{tr}(X(X^T X + \lambda I)^{-1} X^T)$
 KINDA WRONG FOR LASSO $\text{df} \neq \# \text{ OF NONZERO ENTRIES in } \hat{\beta}_\lambda$



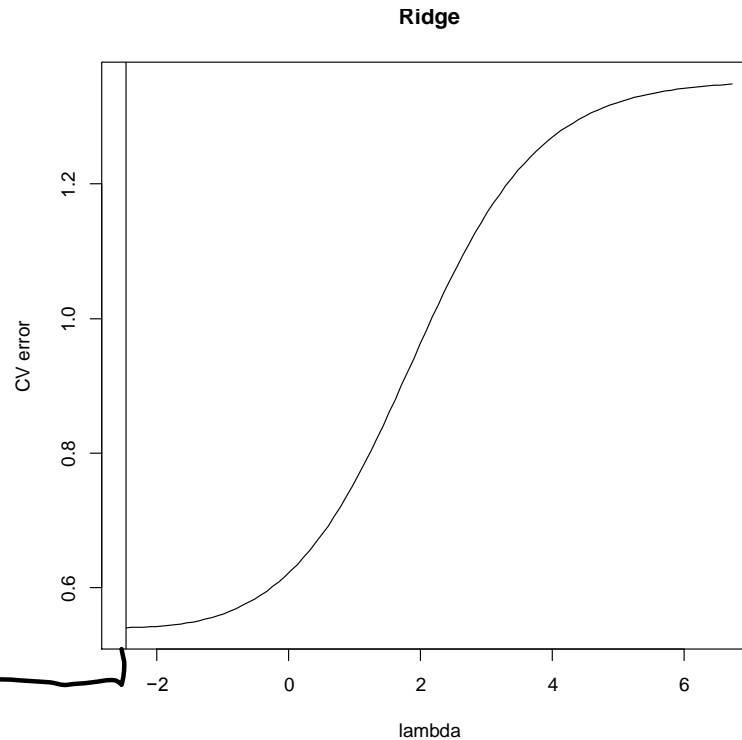
$$\text{df} = \mathbb{E} \text{rank } \underline{X} \underline{g}(\lambda)$$

EXAMPLE: $\hat{\beta}_\lambda = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}$
 $\underline{g}(\lambda) = \{2, 3\}$

$$\hookrightarrow \hat{\text{df}} = \text{rank } \underline{X} \underline{g}(\lambda)$$

Example of a **bad** minimum: Grid values too far apart

SOME COMMENTS ABOUT GLMNET



OR, USE LARS

$$\|\sum \beta - Y\|_2^2 + \lambda \|\beta\|_1$$

COMPUTE COEFFICIENT
PATH EXACTLY
(NO GRID)

LASSO:
 $\lambda > \|\sum Y\|_\infty$
 $\rightarrow \hat{\beta}_\lambda = 0$

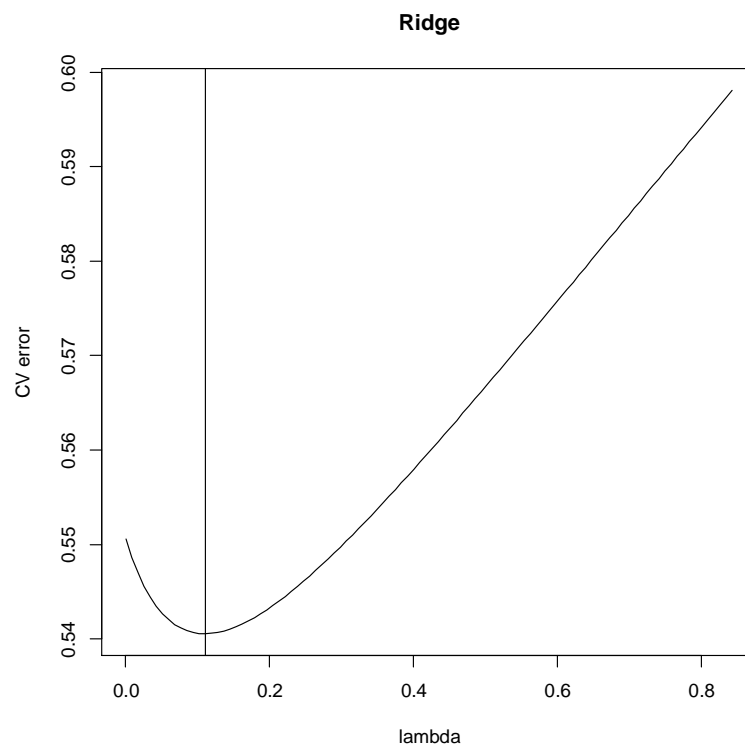
Example of a **bad** minimum: Grid values too large

How to fix it:

```
ridge.cv = cv.glmnet(x=X,y=Y,alpha=0)
min.lambda = min(ridge.cv$lambda)
lambda.new = seq(min.lambda,min.lambda*.001,length=100)
ridge.cv = cv.glmnet(x=X,y=Y,alpha=0,lambda=lambda.new)
lambda.hat = ridge.cv$lambda[which.min(ridge.cv$cvm)]
```

SOME COMMENTS ABOUT GLMNET

New minimum, after moving λ grid **smaller**:



Sparse matrices

SPARSE MATRICES

```
load("../data/hiv.rda")
X = hiv.train$x
> X[5:12,1:10]
```

	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10
[1,]	0	0	0	0	0	0	0	0	0	0
[2,]	0	0	0	0	0	0	0	0	0	0
[3,]	0	0	0	0	0	0	0	0	0	0
[4,]	0	0	0	0	0	0	0	0	0	0
[5,]	0	0	0	0	0	0	0	1	0	0
[6,]	0	0	0	0	0	0	0	0	0	0
[7,]	1	0	0	0	0	0	0	0	0	0
[8,]	0	0	0	0	0	0	0	0	0	0

Many zero entries!

SPARSE MATRICES

All numbers in **R** take up the same **space**
(Space in this context means RAM aka memory)

```
> print(object.size(0),units='auto')  
48 bytes  
> print(object.size(pi),units='auto')  
48 bytes
```

IDEA: If we can tell **R** in advance which entries are zero, **it doesn't need to save that number**

SPARSE MATRICES

This can be accomplished in several ways in R

One is with the **Matrix** package

```
library('Matrix')
```

```
Xspar = Matrix(X,sparse=T)
```

SPARSE MATRICES

Let's take a look at the space difference

```
> print(object.size(X),units='auto')  
1.1 Mb  
> print(object.size(Xspar),units='auto')  
140.7 Kb
```

Pretty substantial! Only 12.1% as large

SPARSE MATRICES

Lastly, we can create sparse matrices without having the original matrix \mathbb{X} ever in memory

This is usually done with three vectors of the same length:

- A vector with row numbers
- A vector with column numbers
- A vector with the entry value

```
i = c(1,2,2)
```

```
j = c(2,2,3)
```

```
val = c(pi,1.01,100)
```

```
sparseMat = sparseMatrix(i = i, j = j, x = val,dims=c(4,4))
```

```
regularMat = as(Matrix(sparseMat,sparse=F),'dgeMatrix')
```

SPARSE MATRICES

```
> print(sparseMat)
4 x 4 sparse Matrix of class "dgCMatrix"
```

```
[1,] . 3.141593 . .
[2,] . 1.010000 100 .
[3,] . . . .
[4,] . . . .
```

```
> print(regularMat)
4 x 4 Matrix of class "dgeMatrix"
```

```
      [,1]      [,2] [,3] [,4]
[1,]      0 3.141593      0      0
[2,]      0 1.010000    100      0
[3,]      0 0.000000      0      0
[4,]      0 0.000000      0      0
```

"DENSE" \Rightarrow "NOT SPARSE"

SPARSE MATRICES

Sparse matrices 'act' like regular (**dense**) matrices

They just only keep track of which entries are non zero and perform the operation on these entries

For our purposes, **glmnet** (and other methods) automatically check to see if \mathbf{X} is a sparse matrix object

This can be a substantial speed/storage savings for large, sparse matrices

Elastic net

ELASTIC NET

The ridge solution is always **unique** and does well when the covariates are highly related to each other:

$$\hat{\beta}_{ridge,\lambda} = \underset{\beta}{\operatorname{argmin}} ||\mathbb{Y} - \mathbb{X}\beta||_2^2 + \lambda ||\beta||_2^2 = (\mathbb{X}^\top \mathbb{X} + \lambda I)^{-1} \mathbb{X}^\top \mathbb{Y}$$

The **lasso** solution

$$\hat{\beta}_{lasso,\lambda} = \underset{\beta}{\operatorname{argmin}} ||\mathbb{Y} - \mathbb{X}\beta||_2^2 + \lambda ||\beta||_1$$

isn't necessarily unique, but it can do **model selection**

However, it can do poorly at model selection if the covariates are highly related to each other

ELASTIC NET

The **elastic net** was introduced to combine both of these behaviors

It solves

$$\hat{\beta}_{\alpha,\lambda} = \underset{\beta}{\operatorname{argmin}} \left[||Y - X\beta||_2^2 + \lambda \left((1 - \alpha)||\beta||_2^2 + \alpha||\beta||_1 \right) \right]$$

We can do the elastic net in **R** with **glmnet**

```
alpha = 0.5
```

```
out.elasticNet = glmnet(x = X, y = Y, alpha=alpha)
```

The parameter **alpha** needs to be set

There does not exist any convention for this, but CV can be used
(You have to write this up yourself, though. Usually, people just play around with different values)

Refitted lasso

REFITTED LASSO

Since lasso does both

- regularization
- model selection

it can produce a solution that produces **too much bias**

A common approach is to do the following two steps:

1. choose the λ via the 'one-standard-error rule'
2. refit the (unregularized) least squares solution on the selected covariates

$$\hat{\beta}_{\text{REFIT}} = \text{ARGMIN}_{\beta} \|Y - \sum_{j \in S(\hat{\lambda}_{1SE})} \beta_j x_j\|_2^2$$

REFITTED LASSO

We can do this in R via

```
#Get CV curve
```

```
lasso.cv.glmnet = cv.glmnet(X,Y,alpha=1)
```

```
#Get beta hat with one-standard-error rule
```

```
#      (remove intercept index -> [-1])
```

```
betaHat.temp = coef(lasso.cv.glmnet,s='lambda.1se')[-1]
```

```
# Identify which covariates are nonzero
```

```
selectedCovs = which(abs(betaHat.temp) > 1e-16)
```

```
# Run regular least squares using those covariates
```

```
refitted.lm = lm(Y~.,data=data.frame(X[,selectedCovs]))
```

```
##
```

```
# Output: either predictions or coefficients
```

```
##
```

```
Yhat.refitted = predict(refitted.lm,X_0[,selectedCovs])
```

```
betaHat.refitted = refitted.lm$coefficients
```

$\downarrow \min$
 $\hat{\beta}_{\lambda^{1se}}$
 $J(\lambda)$

$$\min \|Y - \sum_{j \in S(\lambda)} \beta_j\|_2^2$$

REFITTED LASSO

IMPORTANT: Do not attempt to do inference with the reported p-values. These are absolutely not valid!

However, the parameter values are estimates of the effect of that covariate

Scaled-sparse regression

SCALED-SPARSE REGRESSION

Theoretically, the optimal value for λ looks like:

$$\lambda = C \sigma$$

for some constant C .

- If we knew the true σ , we could find an optimal λ
- If we knew the optimal λ , we could find the σ

This speaks to using an **iterative** approach

SCALED-SPARSE REGRESSION

Scaled sparse regression (SSR) jointly estimates the regression coefficients and noise level in a linear model

It alternates between

1. estimating σ via

$$\hat{\sigma} = \sqrt{\frac{1}{n} \left\| Y - \mathbb{X} \hat{\beta}_{\text{lasso}, \lambda} \right\|_2^2}$$

2. setting

$$\lambda = \hat{\sigma} \sqrt{\frac{n}{\log(p)}}$$

SCALED-SPARSE REGRESSION

We can do this in **R** via

```
library(scalreg)
lasso.ssr = scalreg(X = X,y = Y,LSE=F)
> names(lasso.ssr)
[1] "hsigma"          "coefficients"    "residuals"
[4] "fitted.values"   "type"            "call"
```

SCALED-SPARSE REGRESSION

Also, the **LSE** parameter indicates if we want to do refitted lasso

Running

```
lasso.ssr = scalreg(X = X, y = Y, LSE=T)
```

Creates an object **lse**

```
> names(lasso.ssr)
```

```
[1] "hsigma" "coefficients" "residuals"
```

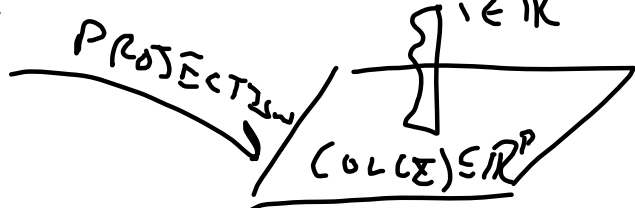
```
[4] "fitted.values" "type" "lse"
```

This object has all the relevant information. For instance predictions

```
Yhat.ssr.refitted = X_0 %*% lasso.ssr$lse$coefficients
```

$$X = \begin{bmatrix} & p \\ & \end{bmatrix}$$

$$X(X^T X)^{-1} X^T$$



$$y = \beta^T X + \epsilon$$
$$\| \beta \|_0 = s$$
$$\text{col}(X) \subseteq \mathbb{R}^p$$