Data Mining: 36-462/36-662
Homework 4 Solutions

Jack Rae

April 15, 2013

Problem 1 [30]

(a) [10]

Recall the least squares criterion

$$S(\beta) = ||y - X\beta||_2^2 \tag{1}$$

If we let $\hat{\beta}$ be the minimizer of (1), and $\tilde{\beta} = \hat{\beta} + c \cdot v$; where $v \in \mathbb{R}^{n \times p}$ satisfies $Xv = 0$ and $c \in \mathbb{R}$, then we see

$$y - X\tilde{B} = y - X(\hat{\beta} + c \cdot v) = y - X\hat{\beta} - cXv = y - X\hat{\beta}$$

which implies $S(\tilde{\beta}) = S(\hat{\beta})$. Thus $\tilde{\beta}$ minimizes (1) $\square$

(b) [5]

If the columns of X are linearly independent then $Xv = 0 \Rightarrow v = \mathbf{0}$, the zero vector.

(c) [15]

(i)

As $p > n$, the columns of X are necessarily linearly dependent, thus by definition $\exists v \neq 0$ such that $Xv = 0$.

(ii)

So from (a) we conclude, since $v \neq 0$ we can construct infinite regression estimates

$$\tilde{\beta}_c = \hat{\beta} + c \cdot v \quad c \in \mathbb{R}$$

that minimize (1) .

(iii)

Let $d = \hat{\beta}_i$ equal the ith component of $\hat{\beta}$. We note that $\hat{\beta}$ and $\tilde{\beta}_{-2d} = \hat{\beta} - 2d \cdot v$ are both solutions to (1) yet in $\tilde{\beta}_{-2d}$ the sign of the ith coefficient has flipped!

(iv)

This is undesirable, for one thing we cannot interpret the coefficients if their sign is not unique. We gain no intuition of whether the ith component contributes positively or negatively to the response variable y.

Problem 2 [30]

(a) [5]

We will denote $\hat{\beta}^{ridge}$ the minimizer of the ridge regression criterion and $\hat{\beta}'$ the minimizer of the sum of squares criterion for the modified matrices $\tilde{X}, \tilde{Y}$. We see the objective function that $\hat{\beta}'$ minimizes

$$\begin{aligned}
||\tilde{Y} - \tilde{X}\beta||_2^2 &= \sum_{i=1}^{n+p} (\tilde{y}_i - \tilde{x}_i\beta)^2 \\
&= \sum_{i=1}^{n} (\tilde{y}_i - \tilde{x}_i\beta)^2 + \sum_{i=n+1}^{n+p} (\tilde{y}_i - \tilde{x}_i\beta)^2 \\
&= \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \sum_{i=n+1}^{n+p} (0 - \sqrt{\lambda} I_i\beta)^2 \\
&= ||Y - X\beta||^2 + \lambda ||\beta||_2^2
\end{aligned}$$

is equal to the objective function $\hat{\beta}^{ridge}$ minimizes. Both optimize over all $\beta \in \mathbb{R}^p$, hence $\hat{\beta}^{ridge} = \hat{\beta}'$.

(b) [5]

Quick argument: the columns of $\sqrt{\lambda}I$ are linearly independent, which implies the columns of $\tilde{X}$ are.

More detailed argument: We note that for an arbitrary column i, $X_{(i+n),i} = 1$ however $X_{(i+n),j} = 0$ for all columns $j \neq i$, by definition of the identity matrix. Thus the $(i+n)$th element of the ith vector (1) cannot be expressed as a linear combination of the $(i+n)$th row of the other column vectors (0) which implies it is linearly independent of the other vectors. We can apply that argument for arbitrary $i \in \{1, 2, \ldots, n\}$ thus proving the columns are linearly independent and thus $\tilde{X}$ has full column-rank.

Once we are assured that $\tilde{X}$ has LI columns, we can argue $\hat{\beta}^{ridge}$ is unique as,

$$\tilde{X} \text{ full column rank} \Rightarrow \tilde{\beta} \text{ is unique} \overset{(a)}{\Rightarrow} \hat{\beta}^{ridge} \text{ is unique.}$$

(c) [5]

$\hat{\beta}^{ridge}$ can be easily computed from vector calculus on the objective function, or from inserting $\tilde{X}$ into the solution of the least squares formula,

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

which we see is relevant from (a). We obtain,

$$\hat{\beta}^{ridge} = (X^T X + \lambda \mathbf{I})^{-1} X^T Y$$

so $\alpha^T \hat{\beta}^{ridge} = \alpha^T (X^T X + \lambda \mathbf{I})^{-1} X^T Y = LY$ which demonstrate's it's linear in Y.

(d) [5]

We have seen the

$$MSE(\alpha^T \hat{\beta}^{ridge}) < MSE(c^T \hat{\beta})$$

As $\alpha^T \hat{\beta}^{ridge} = LY$ is linear, if we suppose it is unbiased then we know

$$MSE(\alpha^T \hat{\beta}^{ridge}) \geq MSE(c^T \hat{\beta})$$

due to $c^T \hat{\beta}$ being a *Best Linear Unbiased Estimator*. This arrives at a contradiction, thus we conclude $\alpha^T \hat{\beta}^{ridge}$ is a biased estimator.

(e) [5]

Using the fact that $U^T U = I$,

$$X^T X = VDU^T UDV^T = VD^2 V^T.$$

Using the fact that $VV^T = I$,[1]

$$X^T X + \lambda I = V(D^2 + \lambda I)V^T.$$

Therefore $(X^T X + \lambda I)^{-1} = V(D^2 + \lambda I)^{-1} V^T$, and

$$\hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T Y = V(D^2 + \lambda I)^{-1} V^T VDU^T Y = V(D^2 + \lambda I)^{-1} DU^T Y.$$

(f) [5]

Thus

$$\begin{aligned}\mathbb{E}(\alpha^T \hat{\beta}^{ridge}) &= \alpha^T V(D^2 + \lambda I)^{-1} DU^T \mathbb{E}(Y) \\ &= \alpha^T V(D^2 + \lambda I)^{-1} DU^T UDV^T \beta \\ &= \alpha^T V(D^2 + \lambda I)^{-1} D^2 V^T \beta.\end{aligned}$$

As $\lambda > 0$ we conclude it is a biased estimator. $\square$

Problem 3 [40]

(a) [20]

The missing code snippets are displayed below,

```
folds = vector(mode="list",length=K)
for (k in 1:(K-1)) {
folds[[k]] = i.mix[((k-1)*d+1):(k*d)]
}
folds[[K]] = i.mix[((K-1)*d+1):n]

a.rid = glmnet(x.tr,y.tr,lambda=lam.rid,alpha=0)
a.las = glmnet(x.tr,y.tr,lambda=lam.las,alpha=1)
```

[1] This assumes linear independence of the columns of X; for linearly dependent columns, a similar but slightly more complicated argument applies.

```
cv.rid = colMeans(e.rid)
cv.las = colMeans(e.las)

pe.rid = matrix(0,K,nlam)
pe.las = matrix(0,K,nlam)
for (k in 1:K) {
i.val = folds[[k]]
pe.rid[k,] = colMeans(e.rid[i.val,])
pe.las[k,] = colMeans(e.las[i.val,])
}
se.rid = apply(pe.rid,2,sd)/sqrt(K)
se.las = apply(pe.las,2,sd)/sqrt(K)

i1.rid = which.min(cv.rid)
i2.rid = max(which(cv.rid<=cv.rid[i1.rid]+se.rid[i1.rid]))

i1.las = which.min(cv.las)
i2.las = max(which(cv.las<=cv.las[i1.las]+se.las[i1.las]))
```

(b) [10]

The resulting cross validation error plots are shown in Figure 1. For ridge regression, the usual rule chooses $\lambda = 4$ and the s.e. rule chooses $\lambda = 15$. For the lasso, the usual rule chooses $\lambda = 0.057$ and the s.e. rule chooses $\lambda = 0.144$.

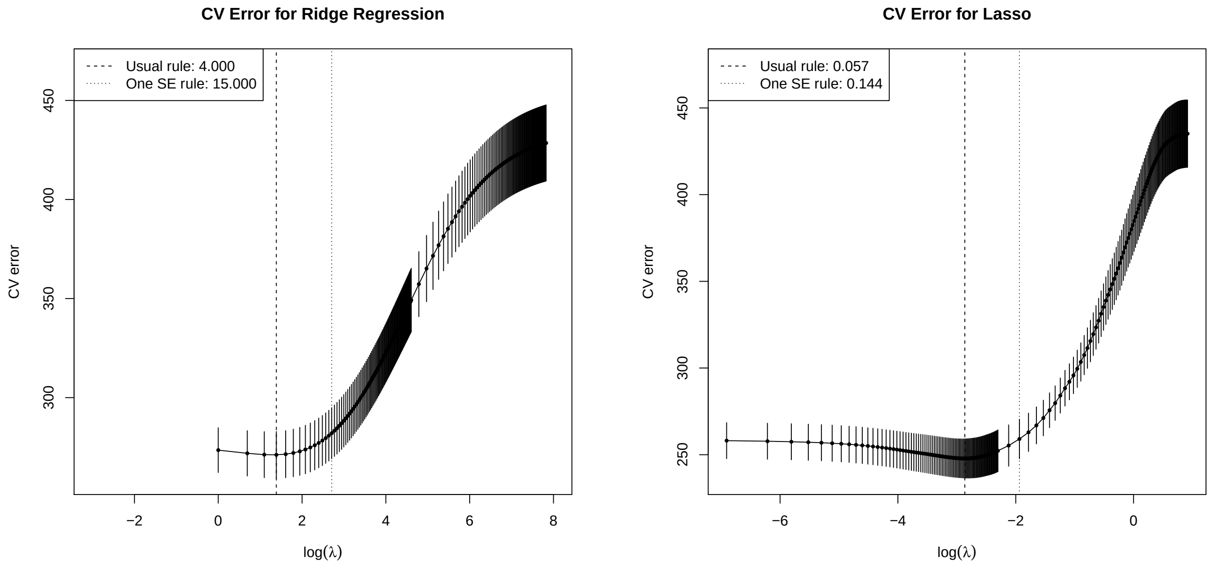
The lasso has smaller minimum cross-validation error than ridge regression, circa 250 vs 300.

(c) [5]

We plot the resulting four reconstructed images in Figure 2. We see that the ridge regression images are much noisier. There appears to be less noise in the images where the s.e. rule was used to pick lambda. The lasso regression using the s.e. rule matches `bstar` the closest.

(d) [5]

We calculate the following sum of square errors,



(a) (b)

Figure 1:

```
> sum((bstar-rid.beta[,i1.rid])^2)
[1] 226.3899
> sum((bstar-rid.beta[,i2.rid])^2)
[1] 245.3781
> sum((bstar-las.beta[,i1.las])^2)
[1] 148.6028
> sum((bstar-las.beta[,i2.las])^2)
[1] 167.567
```

and observe that the lasso using standard cross validation results in the smallest square error.

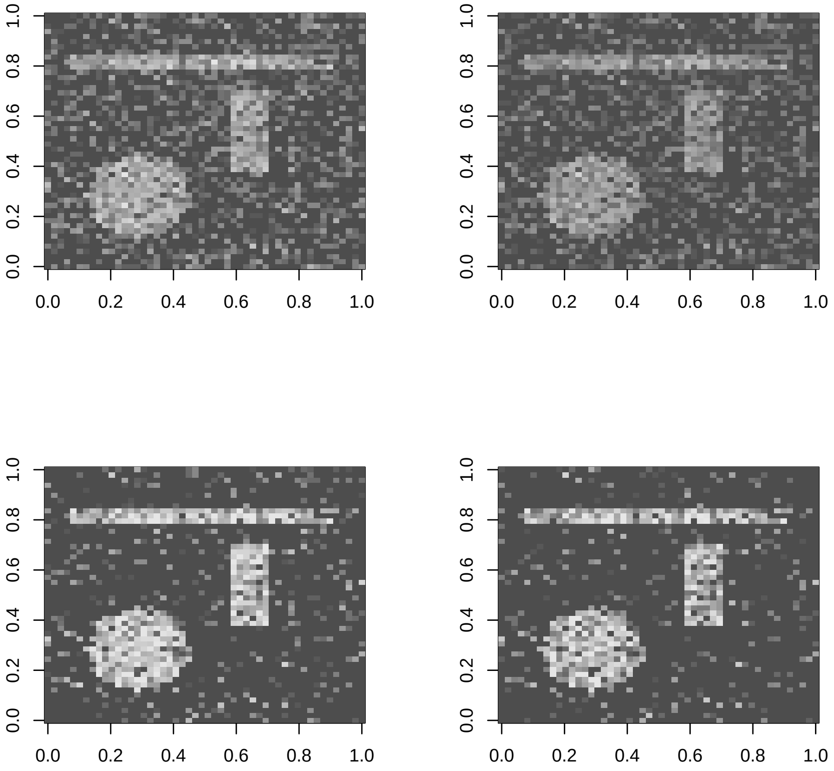


Figure 2: Reading from left to right, top to bottom; ridge regression with 'usual' lambda, ridge regression with 's.e. method' lambda, lasso with 'usual' lambda, lasso with 's.e. method' lambda.