Math 680 - Assignment #2

Kevin McGregor

October 15th, 2015

Question 1

We're trying to minimize $\|Y - X\beta\|^2$ with respect to β. We know that the least squares criterion is a convex function, therefore the minimum we find will be a global minimum. The function can be rewritten as:

$$\begin{aligned} \|Y - X\beta\|^2 &= (Y - X\beta)^\top (Y - X\beta) \\ &= Y^\top Y - (X\beta)^\top Y - Y^\top X\beta + (X\beta)^\top X\beta \\ &= Y^\top Y - 2(X\beta)^\top Y + \beta^\top X^\top X\beta. \end{aligned}$$

Then, differentiating with respect to β gives:

$$\nabla_\beta \|Y - X\beta\|^2 = -2X^\top Y + 2X^\top X\beta.$$

So, if we plug in the centered least squares estimate for β into this equation, we know it will be a global minimum for the least squares criterion. As it's in a similar form as ordinary least squares, the solution for centered least squares works out to be: $\hat{\beta}_{-1} = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \tilde{Y}$ and $\hat{\beta}_1 = \bar{Y} - \bar{x}^\top \hat{\beta}_{-1}$. Then, we have:

$$\begin{aligned} X(\hat{\beta}_1, \hat{\beta}_{-1}^\top)^\top &= 1_n(\bar{Y} - \bar{x}^\top \hat{\beta}_{-1}) + X_{-1}(\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \tilde{Y} \\ &= 1_n \bar{Y} - 1_n \bar{x}^\top \hat{\beta}_{-1} + X_{-1}(\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \tilde{Y} \\ &= 1_n \bar{Y} + (-1_n \bar{x}^\top + X_{-1})(\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \tilde{Y} \\ &= 1_n \bar{Y} + \tilde{X}(\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top (Y - 1_n \bar{Y}) \\ &= (I_n - \tilde{X}(\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top) 1_n \bar{Y} + \tilde{X}(\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top Y \end{aligned}$$

and then we have:

$$\begin{aligned} X^\top X(\hat{\beta}_1, \hat{\beta}_{-1}^\top)^\top &= X^\top (I_n - \tilde{X}(\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top) 1_n \bar{Y} + X^\top \tilde{X}(\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top Y \\ &= X^\top 1_n \bar{Y} + X^\top \tilde{X}(\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top Y \\ &= X^\top Y \end{aligned}$$

Therefore, $(\hat{\beta}_1, \hat{\beta}_{-1}^\top)^\top$ makes $\nabla_\beta = 0$, and so it is a global minimizer.

Question 2

(a)

$$\begin{aligned} \|\tilde{Y} - \tilde{X}\beta\|^2 + \lambda\|\beta\|^2 &= (\tilde{Y} - \tilde{X}\beta)^\top (\tilde{Y} - \tilde{X}\beta) + \lambda\beta^\top \beta \\ &= \tilde{Y}^\top \tilde{Y} - 2\beta^\top \tilde{X}^\top \tilde{Y} + \beta^\top \tilde{X}^\top \tilde{X}\beta + \lambda\beta^\top \beta \\ \nabla f(\beta) &= -2\tilde{X}^\top \tilde{Y} + 2\tilde{X}^\top \tilde{X}\beta + 2\lambda\beta \end{aligned}$$

(b)

The function f is quadratic in β, and the matrix in the quadratic term is $\tilde{X}^\top \tilde{X} + \lambda I$, which is positive definite. Therefore f is strictly convex.

(c)

Since f is strictly convex, there is only one global minimizer of f.

(d)

See R code in "a2_q2d.R" and "centeredRidge_func.R"

(e)

The solution to ridge regression is found by setting $\nabla f(\beta) = 0$, which gives us $\hat{\beta}_{-1}^{(\lambda)} = (\tilde{X}^\top \tilde{X} + \lambda I)^{-1} \tilde{X}^\top \tilde{Y}$. So we have the expected value:

$$\begin{aligned} E[\hat{\beta}_{-1}^{(\lambda)}] &= E[(\tilde{X}^\top \tilde{X} + \lambda I)^{-1} \tilde{X}^\top \tilde{Y}] \\ &= (\tilde{X}^\top \tilde{X} + \lambda I)^{-1} \tilde{X}^\top E[\tilde{Y}] \\ &= (\tilde{X}^\top \tilde{X} + \lambda I)^{-1} \tilde{X}^\top \tilde{X} \beta_{*-1}, \end{aligned}$$

and the variance:

$$\begin{aligned} var[\hat{\beta}_{-1}^{(\lambda)}] &= var[(\tilde{X}^\top \tilde{X} + \lambda I)^{-1} \tilde{X}^\top \tilde{Y}] \\ &= [(\tilde{X}^\top \tilde{X} + \lambda I)^{-1} \tilde{X}^\top] var[\tilde{Y}] [(\tilde{X}^\top \tilde{X} + \lambda I)^{-1} \tilde{X}^\top]^\top \\ &= \sigma_*^2 [(\tilde{X}^\top \tilde{X} + \lambda I)^{-1} \tilde{X}^\top] [(\tilde{X}^\top \tilde{X} + \lambda I)^{-1} \tilde{X}^\top]^\top, \end{aligned}$$

since $var[\tilde{Y}] = \sigma_*^2 I$.

In the simulation study, we independently generate Y 1000 times. The R code for this can be seen in the file "a2_q2d.R". Figure 1 plots the theoretical vs. observed expectation and variance (i.e. the individual elements of each vector or matrix). Since most points fall close to the diagonal, the observed expectation and variance were close to the theoretical values.

Question 3

See R code in "ridge_crossval.R".

Question 4

The simulation setup and top-level function (ridgeSim) are in the file "a2_q4.R". The function ridgeSim also calls functions from the files "centeredRidge_func.R" and "ridge_crossval.R". It calculates the singular value decomposition only once, then selects the proper rows for each iteration of the cross-validation step. In the simulations with $p = 1000$, the ordinary least squares fit failed, and unfortunatly I did not have enough time to redo this. Results can be found in Table 1 and Table 2.

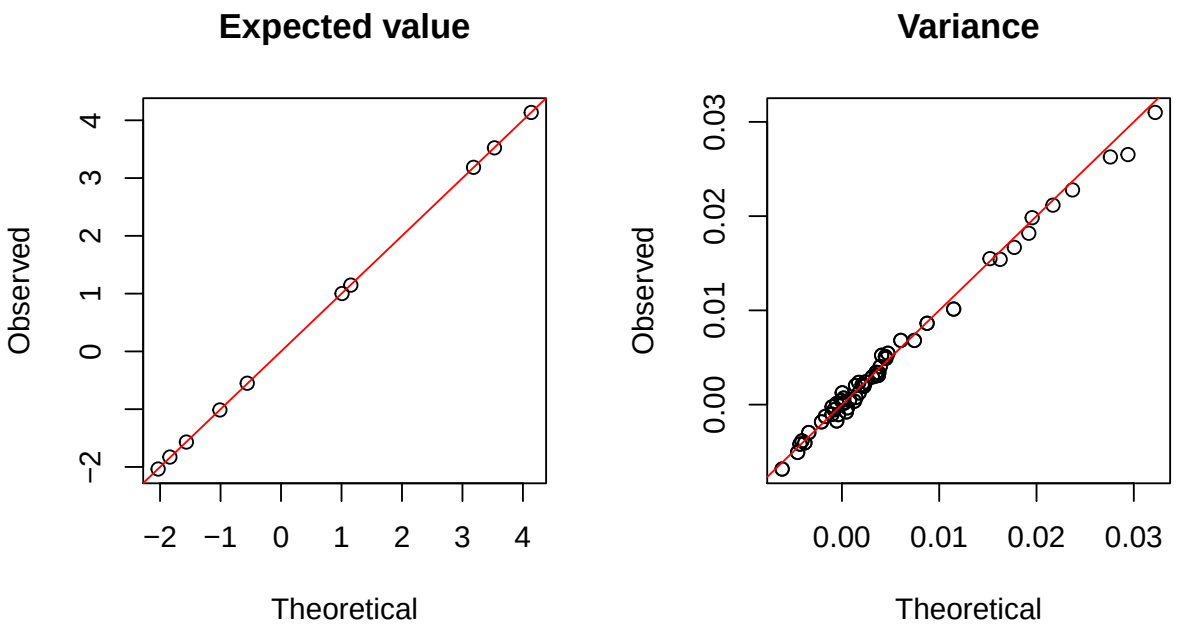

Figure 1: Observed vs. Theoretical Expectation and Variance for the centered ridge estimator

Table 1: Losses for the various simulation scenarios when $p = 50$

(a) $\theta = 0.5$

Param	Type	Average loss	SE loss
β	OLS	0.7614	0.2025
	$K = 5$	0.7416	0.1885
	$K = 10$	0.7400	0.1872
	$K = n$	0.7275	0.1840
$\tilde{X}\beta$	OLS	25.0465	4.9994
	$K = 5$	26.9697	5.9619
	$K = 10$	26.9092	5.9876
	$K = n$	26.6653	5.8784

(b) $\theta = 0.9$

Param	Type	Average loss	SE loss
β	OLS	4.7030	1.4635
	$K = 5$	4.4790	1.2592
	$K = 10$	4.4578	1.2733
	$K = n$	4.4965	1.2758
$\tilde{X}\beta$	OLS	24.5898	5.1006
	$K = 5$	43.8779	9.4545
	$K = 10$	43.7444	9.3660
	$K = n$	43.6703	8.9286

Table 2: Losses for the various simulation scenarios when $p = 1000$

(a) $\theta = 0.5$

Param	Type	Average loss	SE loss
β	OLS	NA	NA
	$K = 5$	434.20	3.68
	$K = 10$	433.46	2.19
	$K = n$	446.70	1.11
$\tilde{X}\beta$	OLS	NA	NA
	$K = 5$	2199.03	2984.08
	$K = 10$	1587.96	1721.32
	$K = n$	12537.21	931.73

(b) $\theta = 0.9$

Param	Type	Average loss	SE loss
β	OLS	NA	NA
	$K = 5$	462.20	5.25
	$K = 10$	467.30	2.79
	$K = n$	468.09	0.46
$\tilde{X}\beta$	OLS	NA	NA
	$K = 5$	5538.02	2978.66
	$K = 10$	8437.47	1581.00
	$K = n$	8886.17	65.73

Question 5

(a)

The objective function can be rewritten as:

$$\begin{aligned} g(\beta, \sigma^2) &= \frac{n}{2}\log(\sigma^2) + \frac{1}{2\sigma^2}\|\tilde{Y} - \tilde{X}\beta\|^2 + \frac{\lambda}{2}\|\beta\|^2 \\ &= \frac{n}{2}\log(\sigma^2) + \frac{1}{2\sigma^2}(\tilde{Y}^\top\tilde{Y} - 2\beta^\top\tilde{X}^\top\tilde{Y} + \beta^\top\tilde{X}^\top\tilde{X}\beta) + \frac{\lambda}{2}\beta^\top\beta, \end{aligned}$$

then the partials are calculated as:

$$\nabla_\beta g(\beta, \sigma^2) = \frac{1}{2\sigma^2}\left(-2\tilde{X}^\top\tilde{Y} + 2\tilde{X}^\top\tilde{X}\beta\right) + \lambda\beta$$

$$\nabla_{\sigma^2} g(\beta, \sigma^2) = \frac{n}{2\sigma^2} - \frac{(\tilde{Y}^\top\tilde{Y} - 2\beta^\top\tilde{X}^\top\tilde{Y} + \beta^\top\tilde{X}^\top\tilde{X}\beta)}{(\sigma^2)^2}.$$

(b)

The second derivative with respect to σ^2 is given by:

$$\nabla^2_{\sigma^2} g(\beta, \sigma^2) = -\frac{n}{2(\sigma^2)^2} + \frac{2(\tilde{Y}^\top\tilde{Y} - 2\beta^\top\tilde{X}^\top\tilde{Y} + \beta^\top\tilde{X}^\top\tilde{X}\beta)}{(\sigma^2)^3},$$

Since this could be negative, we know that g cannot be convex.

(c)

If there exists some β such that $\tilde{Y} - \tilde{X}\beta = 0$, then the term $\|\tilde{Y} - \tilde{X}\beta\|^2$ disappears. As σ^2 tends to infinity, $\log(\sigma^2)$ also increases to infinity, and as σ^2 tends to zero, $\log(\sigma^2)$ decreases to minus infinity. Therefore, since σ^2 has no global minimum, the function g has no global minimum.

(d)

Using Newton's method, the algorithm is as follows:

1. Initialize $\beta_{(k)}$ and $\sigma^2_{(k)}$, and set $k = 1$.
2. Update only β: $\beta' = \beta_{(k)} - \nabla^2_\beta g(\beta_{(k)}, \sigma^2_{(k)})^{-1}\nabla_\beta g(\beta_{(k)}, \sigma^2_{(k)})$
3. Now update only σ^2, but using the new value of β: $(\sigma^2)' = \sigma^2_{(k)} - \nabla^2_{\sigma^2} g(\beta', \sigma^2_{(k)})^{-1}\nabla_{\sigma^2} g(\beta', \sigma^2_{(k)})$
4. Set $k = k + 1$, $\beta_{(k)} = \beta'$, and $\sigma^2_{(k)} = (\sigma^2)'$
5. If $\|\beta_{(k)} - \beta_{(k-1)}\|^2 < \epsilon$ and $(\sigma^2_{(k)} - \sigma^2_{(k-1)})^2 < \delta$, then stop. Otherwise, go back to step 2.

(e)

See R code in "normRidge.R"

(f)

The code for the simulation can be found in the file "a2_q5f.R". The summary of the gradient with respect to β after this run is found in Table 3. The table shows that all the components of the gradient are close to zero, as required to find the minimum. Additionally, the derivative with respect to σ^2 is zero.

Component of $\beta_{-1}^{(\lambda,ML)}$	Value
1	-7.27e-08
2	-7.43e-08
3	-6.66e-08
4	-6.52e-08
5	-6.30e-08
6	-7.14e-08
7	-7.00e-08
8	-7.13e-08
9	-7.10e-08
10	-6.75e-08

Table 3: The observed gradient for β found by normal likelihood ridge regression

Question 6

(a)

$$\begin{aligned} g(\beta) &= \frac{1}{2}\|\tilde{Y} - \tilde{X}\beta\|^2 + \frac{\lambda_1}{2}\|\beta\|^2 + \frac{\lambda_2}{2}\sum_{j=2}^{p}(\beta_j - \beta_{j-1})^2 \\ &= \frac{1}{2}\left(\tilde{Y}^\top\tilde{Y} - 2\tilde{X}^\top\beta^\top\tilde{Y} + \beta^\top\tilde{X}^\top\tilde{X}\beta\right) + \frac{\lambda_1}{2}\beta^\top\beta + \frac{\lambda_2}{2}\left(\sum_{j=2}^{p}\beta_j^2 - 2\sum_{j=2}^{p}\beta_j\beta_{j-1} + \sum_{j=2}^{p}\beta_{j-1}^2\right) \end{aligned}$$

Differentiating with respect to β gives us:

$$\nabla_\beta g(\beta) = -\tilde{X}^\top\tilde{Y} + \tilde{X}^\top\tilde{X}\beta + \lambda_1\beta + \lambda_2 M\beta,$$

where M is a $p \times p$ matrix, written as:

$$M = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & 0 & -1 & 2 & -1 & \dots & 0 \\ \vdots & & & & & \ddots & \vdots \\ 0 & & \dots & & 0 & -1 & 1 \end{bmatrix}$$

Calculating the second derivative results in:

$$\nabla_\beta^2 g(\beta) = \tilde{X}^\top\tilde{X} + \lambda_1 I + \lambda_2 M.$$

When $\lambda_1 > 0$ and $\lambda_2 > 0$, the matrix $\nabla_\beta^2 g(\beta)$ is positive definite. Therefore, the function $g(\beta)$ is strictly convex.

(b)

Setting the gradient to zero gives us the solution $\hat{\beta}_{-1}^{\lambda_1,\lambda_2} = (\tilde{X}^\top\tilde{X} + \lambda_1 I + \lambda_2 M)^{-1}\tilde{X}^\top\tilde{Y}$. The expressions for the expectation and variance of $\hat{\beta}_{-1}^{\lambda_1,\lambda_2}$ are:

$$\begin{aligned} E(\hat{\beta}_{-1}^{\lambda_1,\lambda_2}) &= E\left[(\tilde{X}^\top\tilde{X} + \lambda_1 I + \lambda_2 M)^{-1}\tilde{X}^\top\tilde{Y}\right] \\ &= (\tilde{X}^\top\tilde{X} + \lambda_1 I + \lambda_2 M)^{-1}\tilde{X}^\top E[\tilde{Y}] \\ &= (\tilde{X}^\top\tilde{X} + \lambda_1 I + \lambda_2 M)^{-1}\tilde{X}^\top\tilde{X}\beta_* \end{aligned}$$

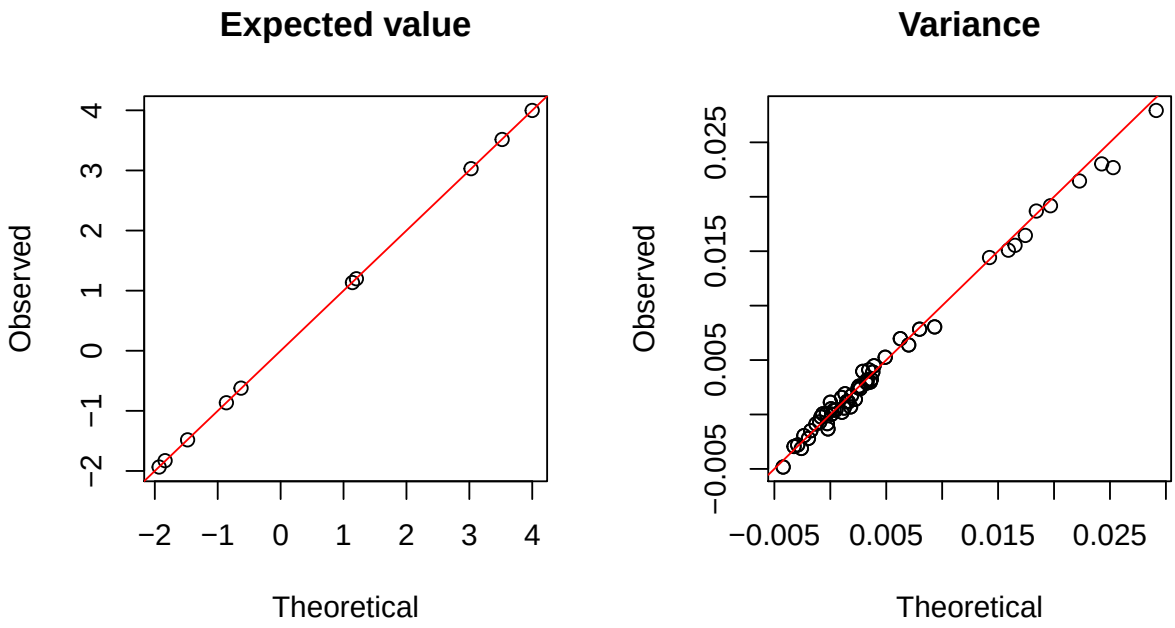

Figure 2: Observed vs. Theoretical Expectation and Variance for the fused ridge estimator

and similarly,

$$\begin{aligned} var(\hat{\beta}_{-1}^{\lambda_1,\lambda_2}) &= var\left[(\tilde{X}^\top\tilde{X}+\lambda_1 I+\lambda_2 M)^{-1}\tilde{X}^\top\tilde{Y}\right] \\ &= \left[(\tilde{X}^\top\tilde{X}+\lambda_1 I+\lambda_2 M)^{-1}\tilde{X}^\top\right] var[\tilde{Y}]\left[(\tilde{X}^\top\tilde{X}+\lambda_1 I+\lambda_2 M)^{-1}\tilde{X}^\top\right]^\top \\ &= \sigma_*^2\left[(\tilde{X}^\top\tilde{X}+\lambda_1 I+\lambda_2 M)^{-1}\tilde{X}^\top\right]\left[(\tilde{X}^\top\tilde{X}+\lambda_1 I+\lambda_2 M)^{-1}\tilde{X}^\top\right]^\top, \end{aligned}$$

since $var[\tilde{Y}] = \sigma_*^2 I$.

The code to do a fused ridge regression is found in "fusedRidge.R", and the code for the simulation study is found in "a2_q6b.R". In Figure 2 we plot the observed elements of the expectation and variance of $\hat{\beta}_{-1}^{\lambda_1,\lambda_2}$ versus the theoretical values. This shows that the observed values are very close to the theoretical values, as all the points lie near the diagonal.