

# MATH 680 Homework 2 Fall 2015

October 10, 2015

This homework is due on Tuesday, Oct 13 at 11:59pm. Provide both pdf, R files. Make an individual R file with proper comments for each sub-problem.

This homework will explore methods to fit the linear regression cases model. Let  $y = (y_1, \dots, y_n)'$  be the measured responses for the  $n$  cases and let  $X \in \mathbb{R}^{n \times p}$  be the nonrandom design matrix, where  $x_{i1} = 1$  for  $i = 1, \dots, n$ . The linear regression cases model assumes that  $y$  is realization of

$$Y = X\beta_* + \epsilon, \quad (1)$$

where  $\beta_* = (\beta_{*1}, \dots, \beta_{*p})' \in \mathbb{R}^p$  is the unknown regression coefficient vector and  $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$  has  $\epsilon_1, \dots, \epsilon_n$  iid with an unspecified distribution having mean zero and unknown variance  $\sigma_*^2$ .

Define  $X_{-1}$  as the matrix with its first column removed. Let  $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$  and  $\bar{x}' = n^{-1} 1_n' X_{-1} = (n^{-1} \sum_{i=1}^n x_{i2}, \dots, n^{-1} \sum_{i=1}^n x_{ip})$ . Let  $\tilde{Y} = (Y_1 - \bar{Y}, \dots, Y_n - \bar{Y})'$  and  $\tilde{X} = X_{-1} - 1_n \bar{x}'$ .

1. Prove Proposition 1 in the lecture note 2 on least squares and matrix decompositions.
2. Suppose we want to compute the ridge penalized least squares estimator of  $\beta_*$  defined by

$$\begin{aligned} \hat{\beta}_{-1}^{(\lambda)} &= \arg \min_{\beta \in \mathbb{R}^{p-1}} \|\tilde{Y} - \tilde{X}\beta\|^2 + \lambda \|\beta\|^2, \\ \hat{\beta}_1^{(\lambda)} &= \bar{Y} - \bar{x}' \hat{\beta}_{-1}^{(\lambda)}. \end{aligned} \quad (2)$$

where  $\lambda \geq 0$  is a tuning parameter. The estimator is  $\hat{\beta}^{(\lambda)} = (\hat{\beta}_1^{(\lambda)}, \hat{\beta}_{-1}^{(\lambda)})'$ .

- (a) Let  $f$  be the objective function in (2). Derive an expression for  $\nabla f(\beta)$ .
- (b) Suppose that  $\lambda > 0$ . Is  $f$  strictly convex? Explain.
- (c) Suppose that  $\lambda > 0$ . Are there multiple global minimizers of (2)? Explain.
- (d) Given a realization of  $Y$ , write an R function that computes the corresponding realization of  $\hat{\beta}^{(\lambda)} = (\hat{\beta}_1^{(\lambda)}, \hat{\beta}_{-1}^{(\lambda)})'$ . This function should have three arguments:

- **X**, the design matrix in  $\mathbb{R}^{n \times p}$
- **y**, the measured response vector in  $\mathbb{R}^n$
- **lam**, this is  $\lambda$

and return a list with two elements:

- **b1**, this is the realization of  $\hat{\beta}_1^{(\lambda)} \in \mathbb{R}$

- **b**, this is the realization of  $\hat{\beta}_{-1}^{(\lambda)} \in \mathbb{R}^{p-1}$

Take advantage of **the singular value decomposition**.

- (e) Derive expressions for  $E(\hat{\beta}_{-1}^{(\lambda)})$  and  $Var(\hat{\beta}_{-1}^{(\lambda)})$  in terms of  $\tilde{X}$ ,  $\beta_*$ ,  $\sigma_*^2$  and  $\lambda$ . Pick an  $n$ ,  $p$ ,  $X$ ,  $\sigma_*$ ,  $\beta_*$  and  $\lambda$  and perform a simulation study to compute simulated estimates of  $E(\hat{\beta}_{-1}^{(\lambda)})$  and  $Var(\hat{\beta}_{-1}^{(\lambda)})$ , where, in each independent replication,  $y$  is generated from (1) with  $\epsilon_1, \dots, \epsilon_n$  iid  $N(0, \sigma_*^2)$ . Compare these estimates to their formulas.
3. Write an R function that performs  $K$ -fold cross-validation, minimizing prediction error, to select the tuning parameter  $\lambda$  for  $\hat{\beta}^{(\lambda)}$  defined in problem 2. Here  $K \in \{2, \dots, n\}$ . This function should have four arguments:

- **X**, the design matrix in  $\mathbb{R}^{n \times p}$
- **y**, the measured response vector in  $\mathbb{R}^n$
- **lam.vec**, the vector of candidate values for  $\lambda$  over which the cross-validation procedure searches for the best.
- **K**, the number of folds to use.

This function should return a list with four objects:

- **b1**, this is the realization of  $\hat{\beta}_1^{(\hat{\lambda})}$ , the estimate of the intercept at the selected value for the tuning parameter.
  - **b**, this is the realization of  $\hat{\beta}_{-1}^{(\hat{\lambda})}$ , the estimate of the regression coefficients at the selected value for the tuning parameter.
  - **best.lam**, this is the selected value for the tuning parameter, i.e. the element of **lam.vec** with the minimum squared prediction error. (accumulated over the  $K$  folds)
  - **cv.error**, this is the vector with the same number of elements as **lam.vec**. The  $j$ th element has the squared prediction error totaled over the  $K$  folds for the estimate with tuning parameter equal to the  $j$ th element of **lam.vec**.
4. In this problem, you will perform a simulation study to compare the ordinary least squares estimator  $\hat{\beta}^{(0)}$  to  $\hat{\beta}^{(\hat{\lambda}_K)}$ , where  $\hat{\lambda}_K$  is the selected tuning parameter from  $K$ -fold cross validation (minimizing prediction error) searching in  $\{10^{-8}, 10^{-7.5}, \dots, 10^{7.5}, 10^8\}$ .

Given values for  $n$ ,  $p$  and  $\theta$  (defined below), set  $\sigma_*^2 = 1/2$  and generate the entries of  $\beta_* \in \mathbb{R}^p$  as independent draws from  $N(0, \sigma_*^2)$ . Set  $X = (1_n, X_{-1})$ , where the rows of  $X_{-1}$  are independent draws from  $N_{p-1}(0, \Sigma)$ , where  $\Sigma$  has  $(j, k)$ th entry  $\theta^{|j-k|}$  for  $j, k \in \{1, \dots, p-1\}$ . This  $X$  and  $\beta_*$  will only be generated once (so they will be the same in all replications).

For each of 200 independent replications, do the following: (i) Generate  $y$  from (1) with  $\epsilon_1, \dots, \epsilon_n$  iid  $N(0, \sigma_*^2)$ . (ii) Compute realizations of the ordinary least squares estimator  $\hat{\beta}^{(0)}$ ,  $\hat{\beta}^{(\hat{\lambda}_5)}$ ,  $\hat{\beta}^{(\hat{\lambda}_{10})}$  and  $\hat{\beta}^{(\hat{\lambda}_n)}$ . (iii) For each of the four estimates, record the value of two losses: the first loss is  $\|b - \beta_*\|^2$  and the second loss is  $\|Xb - X\beta_*\|^2$ , where  $b$  represents the estimate of  $\beta_*$ .

After completing these 200 replications, report the average value of the two losses for each

of the four estimators: these 8 averages are simulated point estimates of 8 expected values:  $E\|\hat{\beta}^{(0)} - \beta_*\|^2$ ,  $E\|\hat{\beta}^{(\hat{\lambda}_5)} - \beta_*\|^2$ ,  $E\|\hat{\beta}^{(\hat{\lambda}_{10})} - \beta_*\|^2$ ,  $E\|\hat{\beta}^{(\hat{\lambda}_n)} - \beta_*\|^2$ ,  $E\|X\hat{\beta}^{(0)} - X\beta_*\|^2$ ,  $E\|X\hat{\beta}^{(\hat{\lambda}_5)} - X\beta_*\|^2$ ,  $E\|X\hat{\beta}^{(\hat{\lambda}_{10})} - X\beta_*\|^2$ ,  $E\|X\hat{\beta}^{(\hat{\lambda}_n)} - X\beta_*\|^2$ . Report an estimated standard error for each of these eight point estimates.

- (a) Perform this study with  $n = 100$ ,  $p = 50$ , and  $\theta = 0.5$ .
  - (b) Perform this study with  $n = 100$ ,  $p = 50$ , and  $\theta = 0.9$ .
  - (c) Perform this study with  $n = 100$ ,  $p = 1000$ , and  $\theta = 0.5$ .
  - (d) Perform this study with  $n = 100$ ,  $p = 1000$ , and  $\theta = 0.9$ .
5. Suppose we want to compute the ridge penalized Normal likelihood estimator of  $(\beta_*, \sigma_*^2)$  defined by

$$\begin{aligned} (\hat{\beta}_{-1}^{(\lambda, ML)}, \hat{\sigma}^{2(\lambda, ML)}) &= \arg \min_{(\beta, \sigma^2) \in \mathbb{R}^{p-1} \times \mathbb{R}_+} \left\{ \frac{n}{2} \log(\sigma^2) + \frac{1}{2\sigma^2} \|\tilde{Y} - \tilde{X}\beta\|^2 + \frac{\lambda}{2} \|\beta\|^2 \right\} \quad (3) \\ \hat{\beta}_1^{(\lambda, ML)} &= \bar{Y} - \bar{x}' \hat{\beta}_{-1}^{(\lambda, ML)}, \end{aligned}$$

where  $\lambda \geq 0$  is a tuning parameter.

- (a) Let  $g$  be the objective function in (3). Derive an expression for  $\nabla_{\beta} g(\beta, \sigma^2)$  and  $\nabla_{\sigma^2} g(\beta, \sigma^2)$ .
  - (b) Is  $g$  convex? Explain.
  - (c) Prove that if there exists a  $\bar{\beta}$  such that  $\tilde{Y} - \tilde{X}\bar{\beta} = 0$  and  $\lambda > 0$ , then a global minimizer for the optimization problem in (3) does not exist.
  - (d) For the case that  $\tilde{Y} - \tilde{X}\bar{\beta} \neq 0$  for all  $\bar{\beta} \in \mathbb{R}^{p-1}$ , propose an algorithm to compute the solution to (3).
  - (e) Write an R function to test your algorithm. Use the same arguments as you used for ridge penalized least squares function.
  - (f) Randomly generate data to test your function. Does the corresponding realization of  $\nabla g(\hat{\beta}_{-1}^{(\lambda, ML)}, \hat{\sigma}^{2(\lambda, ML)})$  approximately equal vector of zeros?
6. Consider the **fused ridge regression** firstly introduced in [1]. The estimator of  $\beta_*$  defined by

$$\begin{aligned} \hat{\beta}_{-1}^{(\lambda_1, \lambda_2)} &= \arg \min_{\beta \in \mathbb{R}^{p-1}} \left\{ \frac{1}{2} \|\tilde{Y} - \tilde{X}\beta\|^2 + \frac{\lambda_1}{2} \|\beta\|^2 + \frac{\lambda_2}{2} \sum_{j=2}^p (\beta_j - \beta_{j-1})^2 \right\} \quad (4) \\ \hat{\beta}_1^{(\lambda_1, \lambda_2)} &= \bar{Y} - \bar{x}' \hat{\beta}_{-1}^{(\lambda_1, \lambda_2)}, \end{aligned}$$

where  $\lambda_1, \lambda_2 \geq 0$  are tuning parameters. The first penalty  $\frac{\lambda_1}{2} \|\beta\|^2$  controls the overall size of the coefficient. The second penalty  $\frac{\lambda_2}{2} \sum_{j=2}^p (\beta_j - \beta_{j-1})^2$  controls the size of coefficient differences. Figure 1 shows the idea of the fused ridge problem.

- (a) Prove that the objective function in (4) is strictly convex when  $\lambda_1 > 0$ .

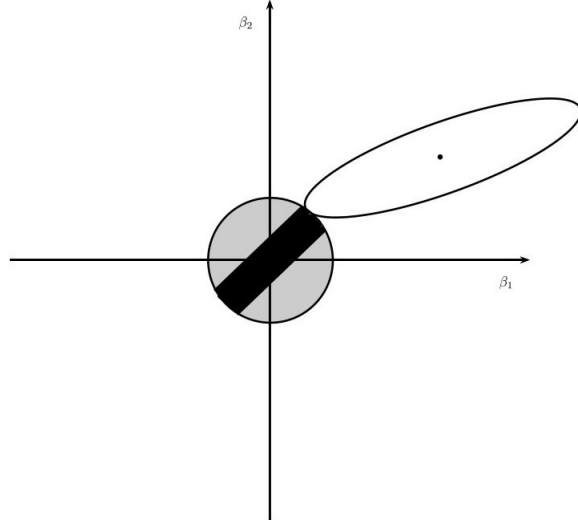


Figure 1: Schematic diagram of the fused ridge, for the case  $N > p = 2$ : we seek the first time that the contours of the sum-of-squares loss function with respect to  $(\beta_1, \beta_2)$  (ellipses) satisfy  $\sum_j \beta_j^2 \leq s_1$  (gray zone) and  $\sum_j (\beta_j - \beta_{j-1})^2 \leq s_2$  (black zone), for  $s_1, s_2 \geq 0$ .

- (b) Derive expressions for  $E(\hat{\beta}_{-1}^{(\lambda_1, \lambda_2)})$  and  $Var(\hat{\beta}_{-1}^{(\lambda_1, \lambda_2)})$  in terms of  $\tilde{X}$ ,  $\beta_*$ ,  $\sigma_*^2$ ,  $\lambda_1$  and  $\lambda_2$ . Pick an  $n$ ,  $p$ ,  $X$ ,  $\sigma_*$ ,  $\beta_*$ ,  $\lambda_1$  and  $\lambda_2$  and perform a simulation study to compute simulated estimates of  $E(\hat{\beta}_{-1}^{(\lambda_1, \lambda_2)})$  and  $Var(\hat{\beta}_{-1}^{(\lambda_1, \lambda_2)})$ , where, in each independent replication,  $y$  is generated from (1) with  $\epsilon_1, \dots, \epsilon_n$  iid  $N(0, \sigma_*^2)$ . Compare these estimates to their formulas. Compare these estimates to their formulas.

## References

- [1] Stephanie R Land and Jerome H Friedman. Variable fusion: A new adaptive signal regression method. Technical Report, Department of Statistics, Carnegie Mellon University Pittsburgh, 1997.