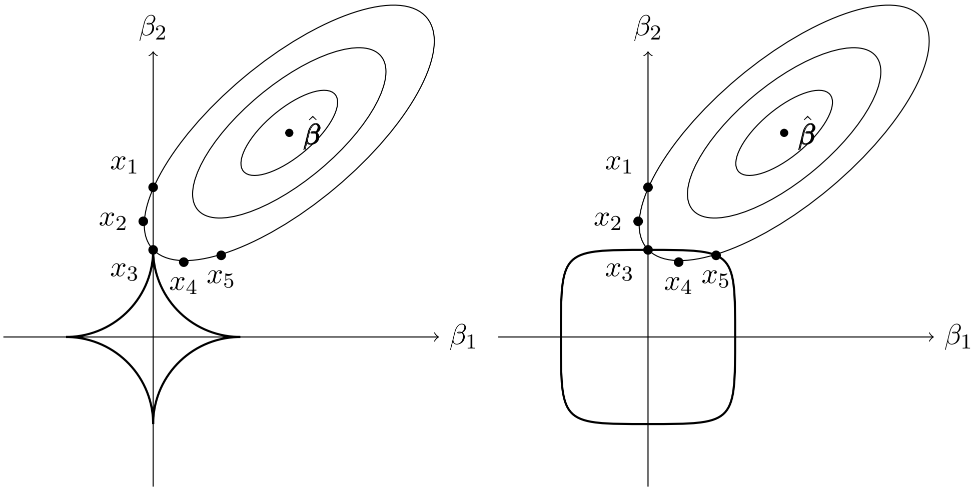
Applied Statistical Methods
Statistics GU3105 — Fall 2021

Homework 3

The following problems are due on Tuesday, October 26, 11:59pm.

1. ℓ_1 **Regression**



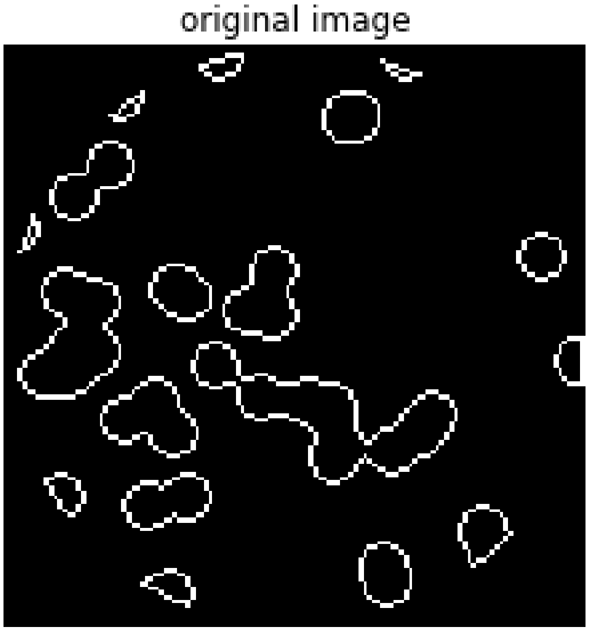
The figures show the cost function components of the ℓ_q-regression problems with $q = 0.5$ (left) and $q = 4$ (right).

(a) Does one/none/both of the cost functions encourage sparse estimates? If so, which one? Explain your answer.

(b) Which of the points $x_1, \ldots, x_5$ would achieve the smallest cost under the ℓ_q-constrained least squares cost function? For each of the two cases, name the respective point and give a brief explanation for your answer.

(c) Write down the loss function with $q = 0.5$ and $q = 4$ respectively. Which one do you think is easier to solve numerically and why?

2. **Lasso vs ridge regression, compressed sensing** In this coding exercise, we will see how to use lasso and ridge regression methods to reconstruct an image from a set of corrupted data. The dataset contains 2 txt files, `hw3_Q2_X.txt` and `hw3_Q2_Y.txt`. The original image is shown as the following:



The dimension of the figure is 128 × 128 = 16384, but we only have 2304 of them available. In class, this is identified as the "high-dimensional" problem: we are using 2304 samples to predict the value of 16384. We will experiment with different values of $\lambda \in \{0.000001, 0.0001, 0.01, 0.1, 1\}$.

Please refer to the R Hints for ridge regression, lasso regression, and plotting the coefficient vector as an image. You will need to install **raster** and **glmnet** if you haven't done so already. The **glmnet** function allows you to fit Lasso and Ridge regressions by tweaking the parameter **alpha** for Elastic Net. To conduct the experiments, you will also need to tweak the parameter **lambda** according to the values mentioned above.

For your own sake, please do not modify or remove any of the remaining parameters, unless you are 100% certain about what you are doing!

Listing 1: R Hints

```
library(raster)
library(glmnet)

# Lasso Regression (alpha = 1)
lasso_reg <- glmnet(X_mat, Y_lst, family = "gaussian", intercept = T,
                  alpha=1, lambda = 1e-6, lambda.min.ratio = 1e-8,
                  standardize = F, nlambda = 1)
# Reshape Lasso Coefficients to a 128x128 matrix
beta_mat <- matrix(lasso_reg$beta[,1], nrow=128, byrow=T)
# Visualize the Matrix in Grey Scale
image(t(apply(beta_mat, 2, rev)), col=grey(seq(0, 1, length=256)))

# Ridge Regression (alpha = 0)
ridge_reg <- glmnet(X_mat, Y_lst, family = "gaussian", intercept = T,
                  alpha=0, lambda = 1e-6, lambda.min.ratio = 1e-8,
                  standardize = F, nlambda = 1)
# Reshape Ridge Coefficients to a 128x128 matrix
beta_mat <- matrix(ridge_reg$beta[,1], nrow=128, byrow=T)
# Visualize the Matrix in Grey Scale
image(t(apply(beta_mat, 2, rev)), col=grey(seq(0, 1, length=256)))
```

(a) (10 Points) Using the 5 penalty values of λ mentioned above to fit the ridge regression model. You will have the coefficient $\beta \in \mathbb{R}^{16384}$. Reshape β into 128×128 matrix, and plot the matrix as in the original image. Which coefficient yields the best result?

(b) (10 Points) Using the 5 penalty values of λ mentioned above to fit the lasso model. You will have the coefficient $\beta \in \mathbb{R}^{16384}$. Reshape β into 128×128 matrix, and plot the matrix as in the original image. Which coefficient yields the best result?

(c) (5 Points) Comparing the best results from lasso and ridge regression respectively, what do you find?

Project Question Series 2

Linear Regression Models

Your instructor used to and will travel frequently to attend different conferences. In order to arrive at the conference meeting on time, she always needs to take the flight delays into consideration. Your project is to help her get an estimation of the delayed time.

1. Use the methods discussed in HW session 2 to clean the data. **Make sure you are using the cleaned data for all subsequent problems.**

2. Set $X =$ (`dep_time, dep_delay, arr_time`) as the predictors and $Y =$ `arr_delay` as the response variable, formulate a linear regression model to help your instructor to predict the travel time of future trips.

3. Apply the regression analysis to get parameter estimation for your proposed models and translate the $Y = X\beta$ equation into plain English so that your instructor can tell her grandma.

4. Validate the assumptions behind the linear regression model or diagnose the residuals.

5. Try to include more information and formulate another linear regression model.

6. Which model is the best among your analysis, and give the reason for your choice.