# STAT 154 Lab 6: Beyond ordinary Least Squares

Yuansi Chen and Raaz Dwivedi

March 4, 2019

## 1 Weighted Least Squares

Weighted least squares (WLS) is a generalization of ordinary least squares in which the errors covariance matrix can be a diagonal matrix different to an identity matrix. Given the weights $w_i$ for each data point, WLS is solved via minimization of the following weighted sum of squares.

$$\min_{\beta} \sum_{i=1}^{n} w_i \left(y_i - x_i^\top \beta\right)^2, \tag{1}$$

where the weights $w_i$ denote some importance score that is put on the $i$-th data point.

We now show that we can derive WLS in a linear model when each data point has non-identical error variance. Fix $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ as $n$ samples (non-random). The responses $y_i \in \mathbb{R}$ is generated in as follows:

$$y_i = x_i^\top \beta^* + z_i, \tag{2}$$

where $z_i \sim \mathcal{N}(0, \sigma_i^2)$ are independent Gaussian random variables (independent of $x_i$ as well) and $\beta^* \in \mathbb{R}^d$ is a fixed but unknown regression coefficient. Note that we assume that the variances $\left\{\sigma_i^2\right\}$ are known but they are no longer assumed identical as in the standard linear model.

1. **What is the likelihood of the $i$-th data point $p(y_i|x_i;\beta)$?**
2. **Show that computing the maximum likelihood estimator of the above model** (2) **is equivalent to the weighted least squares problem. Also write the final solution of the WLS in matrix form.**
3. **ESL book Ex. 2.6**. Consider a regression problem with inputs $x_i$ and outputs $y_i$, and a parameterized model $f_\theta(x)$ to be fit by least squares. Show that if there are observations with tied or identical values of $x$, then the fit can be obtained from a reduced weighted least squares problem.
   *Clarification:* Here "reduced" means less number of data points.

## 2 Ridge regression

1. Show that ridge regression with $\mathbf{X} \in \mathbb{R}^{n\times d}, \mathbf{y} \in \mathbb{R}^n$ can be seen as an OLS problem with augmented data set $\tilde{\mathbf{X}} \in \mathbb{R}^{n+d\times d}, \tilde{\mathbf{y}} \in \mathbb{R}^{n+d}$.
2. What happens in ridge regression when two feature columns $\mathbf{X}_{\cdot j}$ and $\mathbf{X}_{\cdot k}$ are identical?
3. Show the bias-variance trade-off of ridge regression.
4. (Bayesian interpretation) Assume $y_i \sim \mathcal{N}(x_i^\top \beta^*, \sigma^2)$, $i = 1, 2, \ldots, n$. Additionally, we assume a prior on $\beta^*$ that each $\beta_j^*$ is distribution $\mathcal{N}(0, \tau^2)$. What is the posterior? What is the maximum posterior estimator?
5. Show that $\left\|\hat{\beta}_{\text{ridge}}\right\|_2$ increases as its tuning parameter $\lambda$ goes to 0.

# 3 glmnet package

Glmnet is a package that fits a generalized linear model via penalized maximum likelihood. The regularization path is computed for the lasso or elasticnet penalty at a grid of values for the regularization parameter lambda. The algorithm is fast, and can exploit sparsity in the input matrix $\mathbf{X}$.

**glmnet** solves the following minimization problem under option *family="gaussian"*.

$$\min_{\beta} \frac{1}{n} \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda \left[ (1-\alpha) \|\beta\|_2^2 /2 + \alpha \|\beta\|_1 \right].$$

Try ridge and LASSO with **glmnet** in the following simulation settings.True model $\mathbf{y} = \mathbf{X}\beta^* + \boldsymbol{\epsilon}$. Plot the training MSE v.s. CV MSE as a function of regularization parameter $\lambda$.

1. $\beta^* = (10, 10, 5, 5, \underbrace{1, \ldots, 1}_{10}, 0, \ldots, 0)^\top$, $d = 50 < n = 100$. $\mathbf{X}$ with entries i.i.d normal. $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbb{I})$.
2. $\beta^* = (\underbrace{1, \ldots, 1}_{15}, 0, \ldots, 0)^\top$, $d = 5000 > n = 1000$. $\mathbf{X}$ with entries i.i.d normal. $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbb{I})$.
3. $\beta^* = (\underbrace{1, \ldots, 1}_{1500}, 0, \ldots, 0)^\top$, $d = 5000 > n = 1000$. $\mathbf{X}$ with entries i.i.d normal. $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbb{I})$.
4. $\beta^* = (10, 10, 5, 5, \underbrace{1, \ldots, 1}_{10}, 0, \ldots, 0)^\top$, $d = 50 < n = 100$. $Cov(\mathbf{X})_{ij} = (0.7)^{|i-j|}$. $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbb{I})$.

# 4 LASSO vs ridge picture

Recall the famous picture purporting to explain why $\ell_1$ regularization leads to sparsity, while $\ell_2$ regularization does not. Figure 3.11 in ESL book.

In this problem we'll show that the level sets of the training loss are indeed **ellipsoids** centered at the training loss minimizer $\hat{\beta}$. The mean squared error (MSE) on the training set in the linear regression problem with $\mathbf{X} \in \mathbb{R}^{n \times d}, \mathbf{y} \in \mathbb{R}^n$ $(n > d)$ is

$$\begin{aligned} R(\beta) &= \frac{1}{n} \sum_{i=1}^{n} \left( x_i^\top \beta - y_i \right)^2 \\ &= \frac{1}{n} \|\mathbf{X}\beta - \mathbf{y}\|_2^2 . \end{aligned}$$

Recall that $\hat{\beta}_{\text{OLS}} = \left( \mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{y}$. The LASSO estimator $\hat{\beta}_{\text{LASSO}}$ is the minimizer of the following minimization problem

$$\min_{\beta} \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda_{\text{LASSO}} \|\beta\|_1 . \tag{3}$$

The ridge estimator $\hat{\beta}_{\text{ridge}}$ is the minimizer of the following minimization problem

$$\min_{\beta} \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda_{\text{ridge}} \|\beta\|_2^2 . \tag{4}$$

1. [Optional] Show that for any $\lambda_{\text{LASSO}} > 0$, there exists $\mu_{\text{LASSO}}$ such that the LASSO problem (3) is equivalent to the following minimization problem,

$$\begin{aligned} &\min_{\beta} \|\mathbf{X}\beta - \mathbf{y}\|_2^2 \\ &\text{such that } \|\beta\|_1 \leq \mu_{\text{LASSO}}. \end{aligned}$$

   Show that similar result hold for ridge. You might solve it without introducing the Lagrange multiplier.

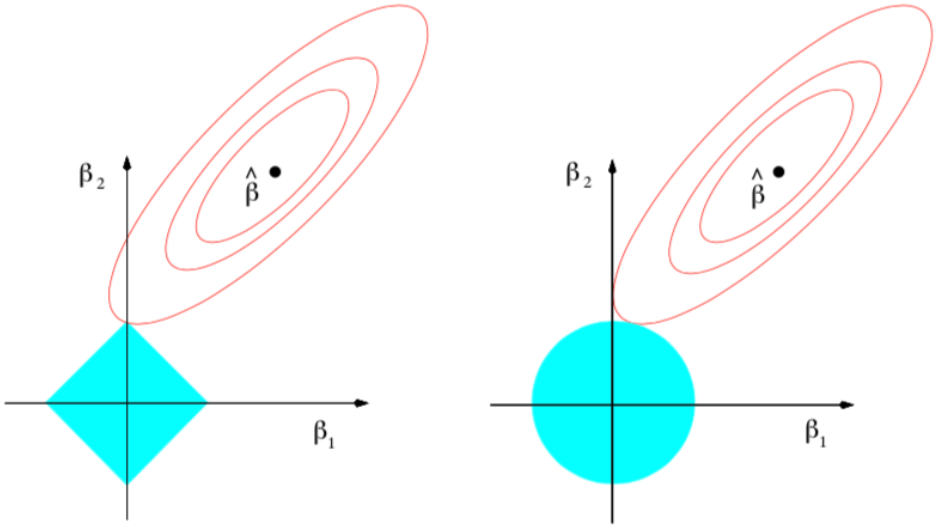


**FIGURE 3.11.** *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

2. Show that for any $\beta \in \mathbb{R}^d$, we have

$$R(\beta) = \frac{1}{n}\left(\beta - \hat{\beta}_{\text{OLS}}\right)^\top \mathbf{X}^\top \mathbf{X}\left(\beta - \hat{\beta}_{\text{OLS}}\right) + R(\hat{\beta}_{\text{OLS}}).$$

3. Give an expression for the set of $\beta$ for which the empirical risk exceeds $R(\hat{\beta}_{\text{OLS}})$ by an amount $c > 0$. If $\mathbf{X}$ is full rank, then $\mathbf{X}^\top \mathbf{X}$ is positive definite, and this set is an ellipsoid – what is its center?