Solutions

February 14, 2023

Answer Q.1. We start from the naive linear algebra approach to get the coefficient of regression. Let ϵ be residual, our goal is to minimize $L = \epsilon^T\epsilon$. Then

$$\begin{aligned}L &= (y - x\beta)^T(y - x\beta)\\ &= (y^T\beta^T x^T)(y - x\beta)\\ &= y^T y - y^T x\beta - \beta^T x^T y + \beta^T x^T x\beta\end{aligned}$$

Take first order derivative and set it to 0 gives us

$$\frac{\partial L}{\partial \beta} = -x^T y - x^T y + 2x^T x\beta = 0$$

So we get $\beta = (x^T x)^{-1}x^T y$ assuming that $x^T x$ is non-singular.
Next we use QR decomposition, where Q is an orthogonal matrix and R is a upper triangular matrix. A special case is that when our model matrix x is a square matrix, we can perform QR decomposition on x directly ie, x = QR. Then

$$\begin{aligned}(QR)^T(QR)\beta &= (QR)^T y\\ R^T Q^T QR\beta &= R^T Q^T y\\ (R_T)^{-1}R^T R\beta &= (R_T)^{-1}R^T Q^T y\\ R\beta &= Q^T y\\ \beta &= R^{-1}Q^T y\end{aligned}$$

If x is not a square matrix, then we can perform QR decomposition on $x^T x$ because $x^T x$ is a symmetric square matrix. Then we have

$$\begin{aligned}(x^T x)^{-1}\beta &= x^T y\\ (QR)^{-1}\beta &= x^T y\\ \beta &= QRx^T y\end{aligned}$$

The third way is to use SVD. We still want to solve the normal equation $x^T x\beta = x^T y$ and now we have $x = UDV^T$ where U and V are orthogonal matrix if we only consider real numbers and D is a diagonal matrix whose entries are square root of eigenvalues of x. Then

$$\begin{aligned}(UDV^T)^T UDV^T\beta &= (UDV^T)^T y\\ VDU^T UDV^T\beta &= VDU^T y\\ V^T VD^2V^T\beta &= V^T VDU^T y\\ D^{-2}D^2V^T\beta &= D^{-2}DU^T y\\ VV^T\beta &= VD^{-1}U^T y\\ \beta &= VD^{-1}U^T y\end{aligned}$$

The final method we use is Cholesky decomposition which can be performed on symmetric matrix. We know that $x^T x$ is symmetric and let $x^T x = LL^T$ where L is a lower triangular matrix. If we assume that $x^T x$ is non singular, then

$$\begin{aligned}(x^T x)^{-1}\beta &= x^T y \\ (LL^T)^{-1}\beta &= x^T y \\ \beta &= LL^T x^T y\end{aligned}$$

From the geom_point figure we can see that there is a positive relationship between time and number of observations as well as features. But when we smooth them, the pattern is very strange. There is a sudden decrease when the log scale changes from 6 to 7, ie: when the number of observations change from 403 to 1100. I'm not sure whether this is caused by calculation steps of function "microbenchmark".

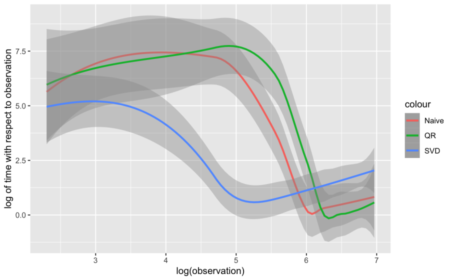

Figure 1: Log time with respect to log observation

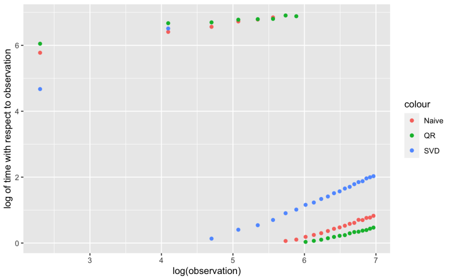

Figure 2: Log time with respect to log observation

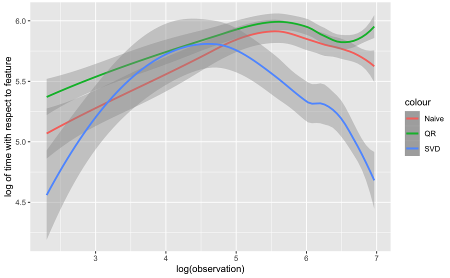

Figure 3: Log time with respect to log feature

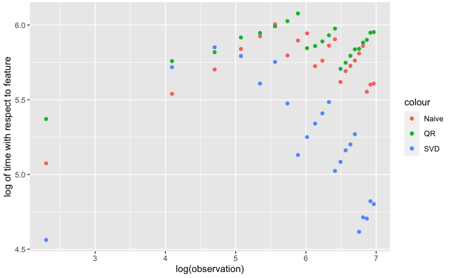

Figure 4: Log time with respect to log feature

Answer Q.2. Please refer to the R markdown file.

Answer Q.3. By Bayesian's rule, we know that posterior $= \frac{prior*likelihood}{evidence}$ ie, $p(\beta|y) = \frac{p(y|\beta)p(\beta)}{p(y)}$. And according to law of total probability, $p(y) = \int p(y|\beta)p(\beta)d\beta$ which has nothing to do when we take derivative of loglikelihood. So we will just call it C for simplicity. We know that $p(y|\beta) \sim N(x\beta, \sigma^2 I), p(\beta) \sim N(0, \tau I)$ and the density of multivariate normal random variable is

$$f(x) = (2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

Then we can write $p(\beta|y)$ as

$$p(\beta|y) = \frac{1}{C}(2\pi)^{-\frac{n}{2}}\sigma^{-n}e^{\frac{(y-x\beta)^T(y-x\beta)}{2\sigma^2}}(2\pi)^{-\frac{p}{2}}\tau^{-\frac{p}{2}}e^{-\frac{\beta^T\beta}{2\tau}}$$

Take log and put all constant into one term then we can get the loglikelihood as

$$log(p(\beta|y)) = K - \frac{(y-x\beta)^T(y-x\beta)}{2\sigma^2} - \frac{\beta^T\beta}{2\tau}$$

Take derivative of this loglikelihood with respect to β then we get

$$\frac{\partial log(p(\beta|y))}{\partial\beta} = -\frac{-2x^Ty + 2x^Tx\beta}{2\sigma^2} - \frac{2\beta}{2\tau}$$

Set this derivative to 0 and solve for β we can get

$$\hat{\beta} = (x^Tx + \frac{\sigma^2}{\tau}I)^{-1}x^Ty$$

Compare this to formula (3.34) in ESL and we can see that if we set $\lambda = \frac{\sigma^2}{\tau}$, then it's just ridge regression solution. In order to prove that this solution is the mean of posterior distribution. Recall that the product of two multivariate normal random variables gives us another multivariate normal random variable. So if we compare the quadratic term then we can see that the variance covariance of β is $\Sigma^{-1} = \frac{x^Tx}{\sigma^2} + \frac{I}{\tau}$. And linear terms are $-2\beta^T\Sigma^{-1}\beta$ and $\frac{-2\beta^Tx^Ty}{\sigma^2}$. Thus we equate these two and solve for β, we get

$$\begin{aligned}\beta &= \frac{\Sigma x^Ty}{\sigma^2}\\ &= \frac{(\frac{x^Tx}{\sigma^2} + \frac{I}{\tau})^{-1}x^Ty}{\sigma^2}\\ &= (x^Tx + \frac{\sigma^2}{\tau}I)^{-1}x^Ty\\ &= \hat{\beta}\end{aligned}$$

Answer Q.4. Recall that the ridge regression solution is $\beta_{ridge} = (x^Tx + \lambda I)^{-1}x^Ty$ and we use SVD, ie : $X = UDV^T$ where U and V are orthogonal matrix if we only take real numbers into consideration and D is a diagonal matrix. So we have $UU^T = I$and $VV^T = I$. Then β can be expressed as

$$\begin{aligned}\beta_{ridge} &= (VDU^TUDV^T + \lambda I)^{-1}VDU^Ty\\ &= (VD^2V^T + \lambda VV^T)^{-1}VDU^Ty\\ &= (V(D^2 + \lambda I)V^T)^{-1}VDU^Ty\\ &= V(D^2 + \lambda I)^{-1}DU^Ty\end{aligned}$$

Now consider the square of norm , then we have

$$\begin{aligned}||\beta_{ridge}||^2 &= \beta_{ridge}^T\beta_{ridge}\\ &= y^TUD(D^2+\lambda I)^{-1}V^TV(D^2+\lambda I)^{-1}DU^Ty\\ &= y^TUD(D^2+\lambda I)^{-2}DU^Ty\\ &= (U^Ty)^T[D(D^2+\lambda I)^{-2}D](U^Ty)\\ &= q^TAq\end{aligned}$$

We can see from the expression that squared norm is just a quadratic form where A is a diagonal matrix with entries $\frac{d_j^2}{d_j^2+\lambda}$ along the diagonal according to formula (3.47) in ESL. Thus $||\beta_{ridge}||^2 = \sum\frac{d_j^2k^2}{d_j^2+\lambda}$ where k is just a scalar. We can see as λ approaches 0 , the squared norm increases, so does norm of β_{ridge}. This property holds for LASSO. The difference between LASSO and ridge regression is a matter of L_1 and L_2 norm, respectively. They can be written as $\hat{\beta} = argmin\{\sum(y_i - \beta_0 - \sum\beta_jx_{ij})^2 + \lambda\sum|\beta_j|^q\}$. When λ approaches 0, the term $\sum|\beta_j|^q$ must increases to get the property of penalty and make $\lambda\sum|\beta_j|^q$ constant.

Answer Q.5. We make a new model matrix X_{new} and β_{new} as the following

$$x_{new} = \left[\begin{array}{c|c} x & x^* \end{array}\right]$$

$$\beta_{new} = \left[\begin{array}{c}\beta\\ \hline \beta^*\end{array}\right]$$

Then we have $x_{new}\beta_{new} = x\beta + x^*\beta^* = x\beta + x\beta^*$. The LASSO problem is that we want to find β such that

$$\begin{aligned}\beta &= argmin\{||y - x_{new}\beta_{new}||_2^2 + \lambda(\sum_{j=1}^p|\beta_j| + |\beta_j^*|)\}\\ &= argmin\{||y - x(\beta+\beta^*)||_2^2 + \lambda(\sum_{j=1}^p|\beta_j+\beta_j^*|) + \lambda(\sum_{j=1}^p|\beta_j| + |\beta_j^*| - |\beta_j+\beta_j^*|)\}\end{aligned}$$

Notice that in this expression, the first two term is the original LASSO problem since $\beta + \beta^*$ is a dummy variable which can be replaced by another $\tilde{\beta}$. The triangle inequality implies that $|\beta_j + \beta_j^*| \leq |\beta_j| + |\beta_j^*|$ which means the third term is non negative. We conclude that the set of solutions of β_j and β_j^* is all β_j and β_j^* such that $\beta_j + \beta_j^* = a$ and they have the same sign. The reason is that when they have the same sign, the third term has 0 so it has no contribution on the optimization.

Answer Q.6. In this question, we use the square norm for simplicity since for any vector x, $||x||^2 =< x, x >= x^Tx$. Suppose originally we have p features and n observations ie, $y_1 \in R^n$ and x_1 is a n*(p+1) model matrix. Then consider

$$X_2 = \left[\begin{array}{c}X_1\\ \hline \tau I_k\end{array}\right]$$

$$y_2 = \left[\frac{y_1}{O_k}\right]$$

Then X_2 is a (n+k)*(p+1) matrix and $y_2 \in R^{n+k}$. We can write $y_2 - x_2\beta$ as

$$y_2 - x_2\beta = \left[\frac{y_1 - x_1\beta}{-\tau\beta}\right]$$

Then the squared L_2 norm is

$$||y_2 - x_2\beta||_2^2 = ||y_1 - x_1\beta||_2^2 + \tau^2||\beta||_2^2$$

The LASSO problem is that we want to find $\hat{\beta}$ such that

$$\hat{\beta} = argmin(||y_2 - x_2\beta||_2^2 + \delta||\beta||_1)$$

If we replace the first term, which is RSS, then we get

$$\hat{\beta} = argmin(||y_1 - x_1\beta||_2^2 + \tau^2||\beta||_2^2 + \delta||\beta||_1)$$

Compare this to the original elastic-net optimization,

$$argmin(||y_1 - x_1\beta||_2^2 + \lambda\alpha||\beta|_2^2 + (1-\alpha)||\beta||_1)$$

We only need to set $\tau = \sqrt{\lambda\alpha}$ and $\delta = 1 - \alpha$.