Numerical Insights into Solutions of Lasso and Group Lasso

Final Project, STAT 527

Songqian Chen

Start: May 4, 2016

Very nice - job! And clear, well organized writeup!

Contents

Preface

I am not good at writing, but I tried my best this time:

The original title for this work is ***"A Story about Lasso"***, where I initially planed to write a summary of all kinds of Lasso, like elastic net, group Lasso, sparse group Lasso, graphic Lasso, fused Lasso, adaptive Lasso and so on, into just one article. That is why the chapter 2 is named as **"Steps Further"** but not **"A Step Further"**. At that time I thought Lasso is just something with ℓ_1-norm penalty, and therefore introducing all kinds of Lasso will not be that hard since each extended form just changes the penalty term a little bit. However, in retrospect, I find I was too naive: **the world of Lasso is much more complicated and larger than what I have ever imagined**. Meanwhile, it is also **fantastic**!

Among all calculations and proof, there are two things in chapter 1 that impressed me most. The first one is the concise **closed form solution of Lasso estimators under orthonormal cases** (10). Although we know the geometric interpretation for why Lasso is able to eliminate variables is straightforward (diamond-shape constrain domain), the numerical solution still impressed me with its concise and powerful illustration of **the threshold condition for a variable to be eliminated** and **why the estimates are unbiased**. Actually I have also tried to figure out the solution under general case, however, it is until I studied the duality and KKT conditions in the following chapter that I confirmed there should only be iterative solutions but not a closed one under general case. Then the second thing is the **Bayesian insight** for Lasso. In order to understand this Bayesian view comprehensively, **I learnt by myself everything about Bayesian regression**, found out clearly the **difference between estimation and prediction** in the framework of Bayesian regression, and finally realized that **Lasso estimates are maximum a posteriori (MAP) estimators with Laplace priors** (14) if we try to interpret Lasso from a Bayesian perspective.

I love everything I did and proved in chapter 2, although it was a long journey. At the beginning I just wanted to figure out **why group Lasso can eliminate a group of variables together**. The original paper (Yuan and Lin, 2006[10]) illustrated this point by providing a necessary and sufficient condition of the optimal solution, and it said it is according to **Karush-Kuhn-Tucker (KKT) conditions**. So I went to study the KKT conditions, which then quickly led me to the **optimization problem with duality** of which KKT conditions tends to solve. Hence I came to study the **duality**. Materials varies and some of them are a little confusing, but I finally realize that **KKT conditions is neither necessary nor sufficient condition for a optimization problem, both of these two direction need extra conditions**. However, further study convinced me that **when the optimization problem is convex, and Slater's condition is also satisfied, we would have KKT conditions become a necessary and sufficient condition for the optimal solution**. That is really a fantastic conclusion! Now although we have KKT conditions, the optimization problem is still intractable since the penalty is not differential. Therefore the concept of **subgradient** is needed and with this concept we **extended the stationarity condition so that it can deal with non-differentiable cases** (29). Lastly, **solving the solutions for group Lasso using KKT conditions is really tricky while interesting** (43)! The most interesting trick will be the one (45) when simplifying this solution under within-group orthonormal cases (43).

Finally, I really appreciate the help from these two books:

Statistical learning with sparsity: the lasso and generalizations
(Hastie, Tibshirani and Wainwright, 2015[19])
Convex optimization
(Boyd and Vandenberghe, 2004[8])

This story will never end! Further ideas, standardized group Lasso (Simon, 2012[14]) and sparse group Lasso (Simon, 2013[16]), having been raised to remove the within-group orthonormality and extend sparsity inside each group respectively, are very worthy of discussion as section 2.2 and 2.3. Also the protoLasso (Reid and Tibshirani, 2015[18]) by combining ideas of cluster prototypes and group Lasso is interesting because I just studied their hierarchical clustering via minimax linkage as a project for another course recently! In a nutshell, **I am looking forward to a next version of this document**!

I am really happy that you so broadly these lasso ideas!

Songqian Chen
June, 2016

Chapter 0. At The Beginning

This story begins with our original linear regression model. Imagine we have samples $(\mathbf{x}_i, y_i)$ with $i \in \{1, 2, ..., n\}$, and $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$ are p-dimensional vectors. By fitting the parametric model $y_i = \mathbf{x}_i\beta + \epsilon_i$ with $\epsilon_i \overset{iid}{\sim} (0, \sigma^2)$, our goal is to obtain the best estimates of β subject to minimizing the loss function. That is,

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \alpha - \mathbf{x}_i^T\beta)^2$$

where α is the intercept term. Without loss of generality, we can assume $\bar{y} = 0$ so that we can omit the intercept $\hat{\alpha}$ since $\hat{\alpha} = \bar{y}$. Then we will have

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T\beta)^2. \tag{1}$$

The estimation $\hat{\beta}$ we obtain under this criterion is called ordinary least squares (OLS) estimates, and as is known to most of us, $\hat{\beta}_{OLS} = (X^TX)^{-1}X^Ty$, where $y = (y_1, y_2, ..., y_n)^T$ and X is the design matrix (without **1** as its first column) for our samples. However, as Tibshirani (1996[5], p267) stated, ***"data analysts are often not satisfied with the OLS estimates"***. The first reason is **OLS estimates often have low bias but large variance**, which may cause over-fitting and thus affect the prediction accuracy of the model. While the second reason is **interpretation**:*"with a large number of predictors, we often would like to determine* ***a smaller subset*** *that exhibits the strongest effects"*, Tibshirani said.

In fact when truth is linear, bias of OLS = 0!

Shrinking, also called regularization or penalization nowadays, thus was proposed to improve the prediction accuracy and prevent models from over-fitting by penalizing the large estimated coefficients. ***ridge regression***, raised by Hoerl (1970[2]), had been well developed in playing such an important role by putting an l_2 penalty to the coefficient:

$$\hat{\beta}_{ridge} = \arg\min_{\beta} \Big\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T\beta)^2 + \lambda||\beta||_2^2 \Big\}. \tag{2}$$

This method worked well in respect of shrinking coefficients since it balances the tradeoff between loss function and penalty. However, although it has constraints on coefficients, ridge regression rarely "push" any coefficient to 0, and thus do not choose a "smaller useful subset" that can improve interpretation.

How to put shrinkage and variables selection together? Breiman raised his idea ***non-negative garrote*** (Breiman, 1995[4]) in 1995:

$$minimize \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} c_j x_{ij} \hat{\beta}_j)^2 \quad subject\ to\ c_j \geq 0, \quad \sum_{j=1}^{p} c_j \leq t, \tag{3}$$

where $\hat{\beta}_j$ are usual least squares estimates. This method is the first try to combine variables shrinking and subset selection: *"As the garrote is drawn tighter by decreasing s, more of the $\{c_j\}$ become zero and the remaining nonzero $\tilde{\beta}_j(s)$ are shrunken"*, Breiman stated. Also, in 1993, Frank and Friedman (1993[3]) generalize the ridge regression to the ***bridge regression***, which uses a penalty $\lambda \sum |\beta_j|^\gamma$ with both λ and γ estimated from the data.

Inspired by these ideas, Tibshirani (Tibshirani, 1996[5]) finally gave birth to a powerful technique by perfectly combining the least squares stage and subset selection stage into one in 1996, and this famous child is named ***"Lasso"***, of which the beauty is well-known to all the world nowadays.

I really like your style! :)

Chapter 1. What Is Lasso?

Section 1.1 Definition

Similar to Breiman's idea of garotte and the penalty idea from ridge regression, ***Lasso***, short for ***Least Absolute Shrinkage and Selection Operator***, is defined as a technique minimizing least square loss function subject to a constrain to the absolute sum of coefficients (without loss of generality we still assume $\bar{y} = 0$ to omit the intercept $\hat{\alpha}$):

$$\hat{\beta}_{Lasso} = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 \quad subject\ to \sum_{j=1}^{p} |\beta_j| \leq t, \tag{4}$$

where $t \geq 0$ is a tuning parameter. Moreover, **we would always standardize each predictor, that is, $\sum_{i=1}^{n} x_{ij}^2 = 1$, so that the solution will not depend on the scale of data**. Then rewrite the above definition into Lagrangian form, we would have an equivalent definition that is more well-known:

$$\hat{\beta}_{Lasso} = \arg\min_{\beta} \Big\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \Big\}. \tag{5}$$

Further, by defining the standard l_p norm as $||\mathbf{x}||_p = (\sum_{i=1}^{p} |x_i|^p)^{\frac{1}{p}}$, we can finally have the objective function for lasso written in the norm form:

$$\hat{\beta}_{Lasso} = \arg\min_{\beta} \Big\{ ||y - X\beta||_2^2 + \lambda ||\beta||_1 \Big\}. \tag{6}$$

Section 1.2 Numeric Insight in Orthonormal Predictors Case

(I) Closed Form Solution for Lasso Estimates

To get a deeper numeric insight into Lasso, let we assume a simple case where all the predictors, or covariates, are orthonormal. That is, we would have $X^T X = I$, where X is the design matrix. In this case, the OLS estimates become

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y = X^T y.$$

Now let we solve the equation (6) to get the closed form solutions of Lasso estimates:

$$\begin{aligned} & minimize \quad \Big\{ ||y - X\beta||_2^2 + \lambda ||\beta||_1 \Big\} \\ \Rightarrow \quad & minimize \quad \Big\{ (y - X\beta)^T (y - X\beta) + \lambda ||\beta||_1 \Big\} \\ \Rightarrow \quad & minimize \quad \Big\{ y^T y - 2\beta^T X^T y + \beta^T X^T X \beta + \lambda ||\beta||_1 \Big\}. \end{aligned}$$

Since $y^T y$ is a constant, $X^T y = \hat{\beta}_{OLS}$ and $X^T X = I$, the above objection thus is equivalent with solving

$$\begin{aligned} & minimize \quad \Big\{ -2\beta^T \hat{\beta}_{OLS} + \beta^T \beta + \lambda ||\beta||_1 \Big\} \\ \Rightarrow \quad & minimize \quad \sum_{i=1}^{p} \Big\{ -2\hat{\beta}_i^{OLS} \beta_i + \beta_i^2 + \lambda |\beta_i| \Big\}, \end{aligned}$$

where $\hat{\beta}_i^{OLS}$ and β_i is the i-th element of $\hat{\beta}_{OLS}$ and β respectively. Notice that β_i have no interaction term between each other, therefore minimizing our whole objective function can be divided into minimizing p separated ones, of which only contain β_i only. That is, for each β_i, now we only need to minimize

$$-2\hat{\beta}_i^{OLS}\beta_i + \beta_i^2 + \lambda|\beta_i|. \tag{7}$$

Now obtaining Lasso estimates β_i^{Lasso} has been simplified into solving objective functions (7) for each β_i. As we can see from the equation, when $\hat{\beta}_i^{OLS} > 0$, we must have $\beta_i > 0$ since we want to minimize the function while β_i^2 and $|\beta_i|$ are both symmetric term. Likewise, when $\hat{\beta}_i^{OLS} \leq 0$, we would also have $\beta_i \leq 0$. Then we should discuss two cases:

- If $\hat{\beta}_i^{OLS} > 0$,
  In this situation, we would have the objective function (7) equals to $-2\hat{\beta}_i^{OLS}\beta_i + \beta_i^2 + \lambda\beta_i$. By taking derivative of the objective function on β_i and setting it to 0, we would have

  $$-2\hat{\beta}_i^{OLS} + 2\beta_i + \lambda\big|_{\beta_i=\hat{\beta}_i^{Lasso}} = 0,$$

  which leads to the estimates $\hat{\beta}_i^{OLS} - \frac{\lambda}{2}$. However, these estimates only hold when $\beta_i > 0$. Therefore we should have $\beta_i^{Lasso} > 0$, which finally give us the solution

  $$\beta_i^{Lasso} = (\hat{\beta}_i^{OLS} - \frac{\lambda}{2})_+,$$

  where $(x)_+ = \max(0, x)$. Finally, by denoting $\text{sgn}(\cdot)$ as the sign function that $\text{sgn}(x) = \begin{cases} 1, & x > 0 \\ -1, & x \leq 0 \end{cases}$, and also since $\text{sgn}(\hat{\beta}_i^{OLS}) > 0$, we can rewrite these estimates as

  $$\beta_i^{Lasso} = \text{sgn}(\hat{\beta}_i^{OLS})(|\hat{\beta}_i^{OLS}| - \frac{\lambda}{2})_+. \tag{8}$$

- If $\hat{\beta}_i^{OLS} \leq 0$,
  When $\hat{\beta}_i^{OLS} \leq 0$, the objective function (7) equals to $-2\hat{\beta}_i^{OLS}\beta_i + \beta_i^2 - \lambda\beta_i$. By applying the same method in the above case, we have

  $$-2\hat{\beta}_i^{OLS} + 2\beta_i - \lambda\big|_{\beta_i=\hat{\beta}_i^{Lasso}} = 0.$$

  And also with $\beta_i \leq 0$ in this case, we thus would have

  $$\beta_i^{Lasso} = -(-\hat{\beta}_i^{OLS} - \frac{\lambda}{2})_+.$$

  Finally, since $\text{sgn}(\hat{\beta}_i^{OLS}) = -1$ and $\hat{\beta}_i^{OLS} < 0$, we can also write the solution in a form similar to that in the above case:

  $$\beta_i^{Lasso} = \text{sgn}(\hat{\beta}_i^{OLS})(|\hat{\beta}_i^{OLS}| - \frac{\lambda}{2})_+. \tag{9}$$

Notice that estimates (8) and (9) are actually same. Therefore we can conclude that the **closed form solutions of Lasso estimates under orthonormal cases** is

$$\beta_i^{Lasso} = \text{sgn}(\hat{\beta}_i^{OLS})(|\hat{\beta}_i^{OLS}| - \frac{\lambda}{2})_+. \tag{10}$$

(II) Closed Form Solutions for Ridge and Garrote Estimates

In order to numerically compare the distinct between Lasso regression and ridge and garrote one, we also derive the closed form solutions for ridge and garrote estimates under orthogonal cases. We would firstly begin with ridge estimates.

For ridge regression, we aim to minimize the objective function:

$$\begin{aligned} & minimize \quad \Big\{||y - X\beta||_2^2 + \lambda||\beta||_2^2\Big\} \\ \Rightarrow \quad & minimize \quad \Big\{(y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta\Big\} \\ \Rightarrow \quad & minimize \quad \Big\{y^Ty - 2\beta^TX^Ty + \beta^TX^TX\beta + \lambda\beta^T\beta\Big\}. \end{aligned}$$

Also since y^Ty is a constant, $X^Ty = \hat{\beta}_{OLS}$ and $X^TX = I$, we thus have the equivalent objection as

$$\begin{aligned} & minimize \quad \Big\{y^Ty - 2\beta^TX^Ty + \beta^TX^TX\beta + \lambda\beta^T\beta\Big\} \\ \Rightarrow \quad & minimize \quad \sum_{i=1}^{p}\Big\{ - 2\hat{\beta}_i^{OLS}\beta_i + \beta_i^2 + \lambda\beta_i^2\Big\}. \end{aligned}$$

Since the above equation is derivable, thus an obvious and straight forward way to find the estimates is to take the partial derivatives on each β_i and set them to 0:

$$-2\hat{\beta}_i^{OLS} + 2(1+\lambda)\beta_i\big|_{\beta_i=\beta_i^{ridge}} = 0,$$

by which we can obtain **the closed form solutions of Ridge estimates under orthonormal cases** as

$$\beta_i^{ridge} = \frac{\hat{\beta}_i^{OLS}}{1+\lambda}. \tag{11}$$

When we comes to garrote regression, as we define in equation (3), our objection is to find optimized c_i such that

$$minimize \ \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}c_jx_{ij}\hat{\beta}_j^{OLS})^2 \quad subject\ to\ c_j \geq 0, \quad \sum_{j=1}^{p}c_j \leq t,$$

which is equivalent with the Lagrangian form that

$$minimize \ \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}c_jx_{ij}\hat{\beta}_j^{OLS})^2 + \lambda\sum_{j=1}^{p}c_j \quad with\ c_j \geq 0.$$

Now we denote a $p \times p$ matrix C as a diagonal matrix with diagonal elements $(c_1, c_2, ..., c_p)$:

$$C = \begin{pmatrix} c_1 & 0 & 0 & \cdots & 0 \\ 0 & c_2 & 0 & \cdots & 0 \\ 0 & 0 & c_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & c_p \end{pmatrix},$$

and then we can rewrite the above objective function into the matrix form:

$$\begin{aligned} & minimize\ (y - XC\hat{\beta}_{OLS})^T(y - XC\hat{\beta}_{OLS}) + \lambda\sum_{j=1}^{p}c_j \quad with\ c_j \geq 0 \\ \Rightarrow \quad & minimize\ y^Ty - 2\hat{\beta}_{OLS}^TC^TX^Ty + \hat{\beta}_{OLS}^TC^TX^TXC\hat{\beta}_{OLS} + \lambda\sum_{j=1}^{p}c_j \quad with\ c_j \geq 0 \\ \Rightarrow \quad & minimize\ -2\hat{\beta}_{OLS}^TC^T\hat{\beta}_{OLS} + \hat{\beta}_{OLS}^TC^TC\hat{\beta}_{OLS} + \lambda\sum_{j=1}^{p}c_j \quad with\ c_j \geq 0, \end{aligned}$$

which is actually the same as

$$minimize \ -2\sum_{j=1}^{p} c_j(\hat{\beta}_j^{OLS})^2 + \lambda\sum_{j=1}^{p} c_j^2(\hat{\beta}_j^{OLS})^2 + \sum_{j=1}^{p} c_j \quad with\ c_j \geq 0.$$

By taking derivatives on each of c_j and setting them to 0,

$$-2(\hat{\beta}_j^{OLS})^2 + 2c_j(\hat{\beta}_j^{OLS})^2 + \lambda\big|_{c_j=\hat{c}_j^{garrote}} = 0,$$

we would have the solutions $1 - \frac{\lambda}{2(\hat{\beta}_j^{OLS})^2}$. Also take into consideration constrains that $c_j \geq 0$, we thus obtain that

$$\hat{c}_j^{garrote} = \big(1 - \frac{\lambda}{2(\hat{\beta}_j^{OLS})^2}\big)_+.$$

Now, since the garrote estimates are defined as $\hat{\beta}_j^{garrote} = \hat{c}_j^{garrote}\hat{\beta}_j^{OLS}$, thus we finally have **the closed form solution of garrote estimates uner orthonormal cases** as

$$\hat{\beta}_j^{garrote} = \hat{c}_j^{garrote}\hat{\beta}_j^{OLS} = \hat{\beta}_j^{OLS}\big(1 - \frac{\lambda}{2(\hat{\beta}_j^{OLS})^2}\big)_+. \tag{12}$$

(III) Comparison and Insight

Now that we have derived the closed form solutions under orthonormal cases for Lasso and other methods, we can go on to make a comparison between these estimates and thus obtain a further insight:

- For **ridge estimates** with closed form solution (11)

  $$\beta_i^{ridge} = \frac{\hat{\beta}_i^{OLS}}{1+\lambda},$$

  we see that ridge regression technique is actually shrinking the estimated coefficients by a scale $\frac{1}{1+\lambda}$ in orthonormal cases. Since the OLS estimates $\hat{\beta}_i^{OLS}$ are unbiased for β_i, ridge estimates are thus **biased** with bias $\frac{\lambda}{1+\lambda}\beta_i$. However, by sacrificing some bias, ridge estimates have lower variances than OLS ones and thus are able to make a more precise and stable prediction (Hoerl, 1970[2]). On the other hand, as increasing λ, β_i^{ridge} will become smaller and smaller but **will never be pushed to 0**. Therefore, ridge regression is not able to select a subset of variable as we want.

- Then **garrote estimates** was raised with closed form solution (12)

  $$\hat{\beta}_i^{garrote} = \hat{\beta}_i^{OLS}\big(1 - \frac{\lambda}{2(\hat{\beta}_i^{OLS})^2}\big)_+,$$

  aiming to implement subset selection. As can be seen from the solution, variable selection is actually result from the "truncated positive" part $\big(1 - \frac{\lambda}{2(\hat{\beta}_i^{OLS})^2}\big)_+$. That is, for each $\hat{\beta}_i^{garrote}$, we **would eliminate it to 0 if** $|\hat{\beta}_i^{OLS}| \leq \frac{\sqrt{\lambda}}{\sqrt{2}}$. We can notice that garrote estimates are also **biased estimates** with bias $E[\frac{\lambda}{2\hat{\beta}_i^{OLS}}]$

- For **Lasso estimates** we got the closed form solution (10):

  $$\beta_i^{Lasso} = \mathrm{sgn}(\hat{\beta}_i^{OLS})(|\hat{\beta}_i^{OLS}| - \frac{\lambda}{2})_+.$$

  From the solution, we see that Lasso estimates β_i^{Lasso} are **biased estimates** with bias $\frac{\lambda}{2}$ since the OLS estimates $\hat{\beta}_i^{OLS}$ are unbiased for β_i. Increasing the parameter λ for penalty will **shrink** β_i^{Lasso}

with small absolute values, $|\hat{\beta}_i^{OLS}| \leq \frac{\lambda}{2}$,**to 0**, which reviews the subset selection aspect of lasso. Actually, under the orthonormal cases we see a great similarity between lasso and garrote estimates according to their closed form solutions with respect to the shrinkage idea and threshold. However, **it may be a advantage for lasso than garrote ones that lasso estimates have constant bias while that of garrote estimates depend on** $\hat{\beta}_i^{OLS}$, which is hard to measure and bound.

Section 1.3 Bayesian Insight

(I) Likelihood Function and Maximum Likelihood Technique

In classical treatment of regression with model $y_i = \mathbf{x}_i\beta + \epsilon_i$ and $\epsilon_i \overset{iid}{\sim} (0, \sigma^2)$, we aim to seek a points estimate of the unknown parameter vector $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ by minimizing the least square loss function:

$$minimize \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\beta)^2,$$

which leads to the **ordinary least square estimate** $\hat{\beta}_{OLS}$ for β as we mention in above sections. Now we add an assumption that our model noise ϵ_i are actually from normal distributions with known σ^2:

$$\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2).$$

Then we have $y_i = \mathbf{x}_i^T\beta + \epsilon_i \sim \mathcal{N}(\mathbf{x}_i^T\beta, \sigma^2)$. Given observed data $\mathbf{x}$, we call the probability to obtain the observed outcomes y under specific parameters the **likelihood function**, denoted by $L(\cdot)$, which actually is a function of parameters. In our case, we would have our likelihood function as (there are n samples and the dimension for each sample is p)

$$L(\beta) = p(y|\beta, X) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\big\{ - \frac{\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\beta)^2}{2\sigma^2}\big\}.$$

By maximize the likelihood function, we would have the **maximum likelihood estimates** (MLE) $\hat{\beta}_{MLE}$, which also can be obtained by minimize the negative log-likelihood function $\{-\ln L(\beta)\}$ since $\ln(\cdot)$ is a monotonic increasing function on $(0, \infty)$. Thus we have

$$\begin{aligned}
\hat{\beta}_{MLE} &= \arg\max_{\beta} L(\beta) = \arg\min_{\beta} \big\{ - \ln L(\beta)\big\} \\
&= \arg\min_{\beta} \Big\{ \frac{n}{2}\ln(2\pi) + \frac{n}{2}\ln(\sigma^2) + \frac{\sum_{i=1}^{n}(y_i - \mathbf{x}_i\beta)^2}{2\sigma^2} \Big\} \\
&= \arg\min_{\beta} \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\beta)^2.
\end{aligned}$$

Compare the above equation with the definition of OLS in equation (1), we can find out **they are actually the same**! That is, we would have $\hat{\beta}_{OLS} = \hat{\beta}_{MLE}$ under the normal noise assumption, which mean from now on we can **discuss linear regressions under the framework of probability**.

(II) Bayesian Regression

Now we make a step farther, introducing a more interesting statistical view, **Bayesian view**, to our linear model (Bishop, 2003[7]). Instead of viewing β as constants in classical frequentist's view, Bayesian view regard β as variables, and characterize the uncertainty of β through a probability distribution $p(\beta)$.

Before taking into account the data we observed, we would have a initial belief of the probability density of β which we call **prior distribution** $p(\beta|\lambda)$, where λ is a **hyperparameter** that I would like to interpret

as "the parameter of the distribution of another parameter". Then **the prior belief will be modified by our data through the likelihood function** $L(\beta) = p(y|X, \beta, \lambda)$, where by "modify" we mean using Baye's rule to produce a new modified distribution for β:

$$p(\beta|y, X, \lambda) \propto L(\beta)p(\beta|\lambda),$$

which is called **posterior distribution** of β.

Once we have the distribution of β, we could say that we have already obtain the model we want. Therefore, above is almost the all basic idea behind **bayesian regression model**, but there is still one exciting part missing: prediction! Actually, when it comes to prediction, two concepts will always be confusing to most people who firstly meet Bayesian view and thus they need to be clarified (although this topic may not relate too much to the plotline of our story):

- **posterior mode**: Just like the idea of finding a point that maximizes the likelihood function, a point that maximizes the posterior probability density of β will be named **posterior mode**:

  $$\beta_{mode} = \arg\max_{\beta} p(\beta|y, X, \lambda).$$

  However, an important thing that must be remarked is that in Bayesian view, β_i **is not longer a constant but instead a variable**. Therefore, we now **cannot directly construct an "estimate" for β_i from the posterior mode** as we did in MLE, and instead we can just estimate the $E[\beta]$ by the posterior mean which is the mean of posterior distribution. Hence, **in the framework of Bayesian regression, posterior mode of β will be useful only when we want to estimate $E[\beta]$ and the posterior distribution is symmetric so that posterior mean equals to posterior mode**.

- **Bayesian regression prediction**: Instead of using estimates of $E[\beta]$ to make prediction as frequentists, since β is a variable and we have already obtain the posterior distribution of it, in Bayesian treatments **we would make prediction by integrating all possibilities of** β. That is, given a new observation $\mathbf{x}_{new}$, we would like to find a y that maximizing this integration:

  $$\hat{y} = \arg\max_{y} \int p(y|\beta, \mathbf{x}_{new})p(\beta|y, X, \lambda),$$

  where $p(y|\beta, \mathbf{x}_{new})$ is the likelihood function with new data and $p(\beta|y, X, \lambda)$ is the posterior distribution of β.

In a nutshell, **estimation of β and prediction of $\hat{y}$ are two separate tasks in the framework of Bayesian regression**. Since in this article we would focus more on the estimation aspect for Lasso and ridge regression, therefore posterior mode may seem to be useful in our upcoming bayesian insight. In face, as we will prove in the following part, **the estimates $\hat{\beta}_{Lasso}$ and $\hat{\beta}_{ridge}$ for Lasso and ridge regression are both posterior modes under the bayesian regression framework with different prior distributions for β.**

(III) Lasso and Ridge Regression in Bayesian View

Now we give β a specific prior distribution in exponential form:

$$p(\beta|\lambda) \propto \exp\Big\{ -\frac{\lambda}{2\sigma^2}\Omega(\beta)\Big\},$$

where λ is the hyperparameter and $\Omega(\beta)$ is a function of β. As in the above parts, we assume normality for model noise ϵ such that the likelihood function will still be

$$L(\beta) = p(y|\beta, X) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\Big\{ -\frac{\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\beta)}{2\sigma^2}\Big\}.$$

The bayesian connection is very interesting! Generally Bayesians prefer posterior means to modes – will using the mean affect the Bayesian interp of Lasso? ridge? (hint. It won't for ridge!)

Therefore we can calculate the posterior mode of β as

$$\begin{aligned}\beta_{mode} &= \arg\max_{\beta}\{p(\beta|y,X,\lambda)\} = \arg\min_{\beta}\{-\ln\big(p(\beta|y,X,\lambda)\big)\}\\ &= \arg\min_{\beta}\{-\ln\big(L(\beta)\big) - \ln\big(p(y|\beta,X)\big)\}\\ &= \arg\min_{\beta}\{\frac{n}{2}\ln(2\pi) + \frac{n}{2}\ln(\sigma^2) + \frac{\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\beta)^2}{2\sigma^2} + \frac{\lambda}{2\sigma^2}\Omega(\beta)\}\\ &= \arg\min_{\beta}\{\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\beta)^2 + \lambda\Omega(\beta)\}.\end{aligned}$$

Surprised! The above result for posterior mode is quite similar with the definition of ridge regression (2) and that of Lasso estimates (6). This motivate us to think Lasso and ridge regression in a Bayesian view:

- **Ridge regression**: From the definition of ridge regression estimates (2), we would have $\Omega(\beta)$ in the above equation as

  $$\Omega(\beta) = ||\beta||_2^2 = \sum_{i=1}^{p}\beta_i^2.$$

  That is, for β in ridge regression model, we would assume that β has a prior distribution in the form

  $$p(\beta|\lambda) \propto \exp\{-\frac{\lambda}{2\sigma^2}\sum_{i=1}^{p}\beta_i^2\},$$

  which is obviously the kernel of **a multivariate normal distribution**. Therefore, for ridge regression, it actually gives a normal distribution for each β_i as it prior distribution:

  $$p(\beta_i|\lambda) \overset{iid}{\sim} \mathcal{N}\big(0, \frac{\sigma^2}{\lambda}\big). \tag{13}$$

- **Lasso regression**: Then according to the definition of Lasso regression estimates (6), $\Omega(\beta)$ can be written as

  $$\Omega(\beta) = ||\beta||_1 = \sum_{i=1}^{p}|\beta_i|.$$

  Therefore in Lasso regression, we would essentially assume that β has a prior distribution in the form

  $$p(\beta|\lambda) \propto \exp\big\{-\frac{\lambda}{2\sigma^2}\sum_{i=1}^{p}|\beta_i|\big\},$$

  which is the kernel of **a Laplace distribution (double exponential distribution)**. Therefore, for Lasso regression, it actually gives a Laplace distribution $Laplace(0, \frac{2\sigma^2}{\lambda})$ for each β_i as it prior distribution. That is,

  $$p(\beta_i|\lambda) = \frac{\lambda}{4\sigma^2}\exp\Big\{-\frac{\lambda}{2\sigma^2}|\beta_i|\Big\}, \quad \beta_i \in (0,\infty). \tag{14}$$

After successfully combine the ridge and Lasso regression into the Bayesian framework, we now can find out that **the estimates $\hat{\beta}_{ridge}$ and $\hat{\beta}_{Lasso}$ are both posterior modes under the Bayesian regression framework with normal and Laplace prior distribution respectively**. Then the comparison between ridge and Lasso regression will be clearer if we plot the figure of these two prior distributions (Figure 1).

As we can see in Figure 1, the Laplace prior distribution for Lasso regression is more concentrated than the normal prior distribution for ridge regression. More importantly, although the "peaks" for Lasso and ridge regressions are both at $x = 0$, the one for Lasso regression is sharper than that for ridge regression. Therefore, after modified by data through likelihood function, the posterior mode, or the "peak" of posterior distribution, for Lasso regression is still more likely to remain at the point $x = 0$ while that will be hard for ridge regression to do the same thing.

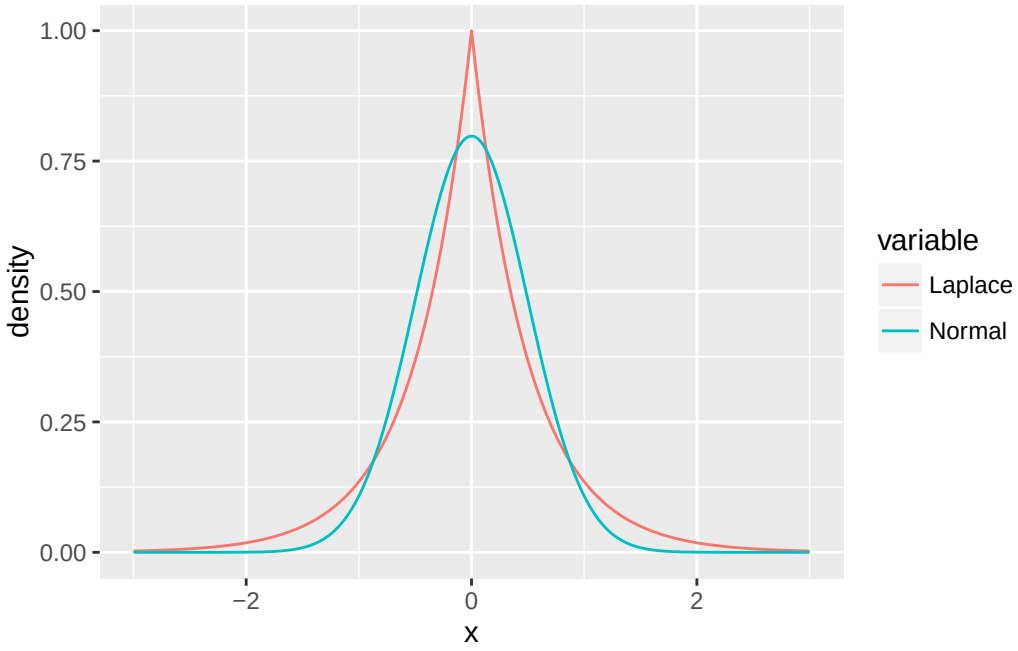

Figure 1: Prior Distributions

Chapter 2. Steps Further

Section 2.1 Group Lasso

(I) Motivation

Consider following situations:

- In the framework of linear regression, now assume some of our predictors are categorical variables. In this case, a traditional method for tackling categorical variables is to code their levels into a set of dummy variables or contrasts. Then when we plan to use Lasso regression for variable selection, **it may happen that for a categorical predictor, some of its levels are remained while some are removed from our model. This is no reasonable since we would want to include or exclude these variables in a "group" together,rather than individual coefficients**. In this case, original Lasso technique is not enough and we need to improve it to realize this desired function.
- Another case is for the genomic data data, where dimension of our data will always be large and variable selection technique is urgently needed. Lasso regression thus seems to be a good choice to deal with these data. However, since there will always be some genes lie in a same biological pathway, and the investigators are often more interesting in which pathways are actually affect the outcome expression than whether particular individual genes are, original Lasso still cannot deal with this case. Improvement is needed!

One common observation for the above two situations is that there is a **"group structure"** underlying the data that we want to deal with when implementing Lasso regression. Traditional Lasso technique tend to choose one in a group of highly correlated variables. However these highly correlated "grouped" variables should be selected or discarded as a group in most cases. Therefore, instead of penalizing coefficients individually, we want a improved Lasso idea that gives penalty with respect to the concept of group. This, obviously, raises the basic idea behind ***Group Lasso***. In a word,

- Basic Idea :***Group Lasso is designed to tackle data with "group structure" under the framework of Lasso regression***.

That is the basic idea and motivation for Group Lasso.

(II) Definition and Insight

As our default setting above, we have n observations (y, X) with y is centered (subtracting the mean) and X is a $n \times p$ sample matrix standardized with respect to each predictor. Now we take a step further, we assume there is a **"group structure"** underlying each X. That is, there are J groups of predictors among all p predictors for X, and thus X can be rewritten as

$$X = (X_1, X_2, ..., X_J)^T,$$

where X_j is a $n \times p_j$ submatrix of the original sample matrix X with p_j variables in the j-th group and $\sum_{j=1}^{J} p_j = p$. Now we can define the objective function for ***group Lasso*** as

$$\hat{\beta}_{group} = \arg\min_{\beta} \left\{ ||y - \sum_{j=1}^{J} X_j \beta_j||_2^2 + \lambda \sum_{j=1}^{J} \sqrt{p_j} ||\beta_j||_2 \right\}, \tag{15}$$

where $\beta = (\beta_1, \beta_2, ..., \beta_J)^T$ and **each β_j is a p_j-dimensional vector**. Several things should be noticed for this definition:

- When $p_1 = p_2 = ... = p_J = 1$, that is, we only have one variable in each group, we would have the penalty in group Lasso (15) $\lambda \sum_{j=1}^{J} \sqrt{p_j} ||\beta_j||_2$ equals to $\lambda \sum_{j=1}^{J} |\beta_j|$, which is exactly the penalty term for original Lasso! That is, the original Lasso is a special case of this more general form, group Lasso.
- With geometrically thinking, we could see that with this $||\cdot||_2$ norm, the constrain domain of coefficients is diamond-shape between different groups while spherical inside each group. By this shape, the estimated coefficients will be more likely to touch the "edge" of those "diamond", and thus coefficients of variables in a same group are more likely to become zero or non-zero simultaneously.
- It is an important note that in the original group lasso paper (Yuan and Lin, 2006[10]), within-group orthonormality is assumed. This assumption was made to simplify the solutions and computation for group Lasso, which we will show in the following study.

(III) Duality and Karush-Kuhn-Tucker (KKT) Conditions

Before we step further into the computation and solution for group Lasso, we must talk something more about the convex optimization (Boyd and Vandenberghe, 2004[8]). Let us begin with **duality**. Imagine our task is to solve an nonlinear programming problem

$$\begin{aligned} &minimize\ f(x), \\ &subject\ to\ g_i(x) \leq 0, \quad i \in \{1, 2, ..., m\}, \\ &\qquad\qquad h_j(x) = 0, \quad j \in \{1, 2, ..., r\}, \end{aligned} \tag{16}$$

where $x \in \mathcal{D} \subset \mathbb{R}^p$, $\mathcal{D}$ is the intersection of domains of f, g_i and h_j, also f:$\mathbb{R}^p \to \mathbb{R}$ is objective function, g_i:$\mathbb{R}^p \to \mathbb{R}$ and h_i:$\mathbb{R}^p \to \mathbb{R}$ are constrains. We would call finding the x^* such that $f(x^*) = \inf_{x \in \mathcal{D}} f(x)$ under these constrains the **primal problem**, together with x^* as the **primal optimal** (or **primal solution**) and the optimal value $p^* = f(x^*)$ for $f(x)$ the **optimal primal value**.

To solve the above optimization problem we need to rewrite them into the **Lagrangian form** (its equivalence with the above optimization problem is further discussed in by slides made by Gordon and Tibshirani (2012[15])):

$$L(x, u, v) = f(x) + \sum_{i=1}^{m} u_i g_i(x) + \sum_{j=1}^{r} v_j h_j(x), \tag{17}$$

where $u \in \mathbb{R}^m$, $u_i \geq 0,\ i \in \{1, .., m\}$ and $v \in \mathbb{R}^r$. With this Lagrangian form, now we can further define the **Lagrange dual function** (or just **dual function**) $g(u, v)$ as

$$g(u, v) = \inf_{x \in \mathcal{D}} L(x, u, v). \tag{18}$$

That is, **the Lagrangian dual function provide the lower bound for the original Lagrangian form with given u and v**. We then have an important property further this definition:

- Property: For $\forall u \succeq 0$ $(u_i \geq 0, i \in \{1, 2, ..., m\})$ and $\forall v$, we would have

$$g(u, v) \leq p^*, \tag{19}$$

  where p^* is the optimal primal value for the primal problem.

The proof is actually straightforward. Suppose $\tilde{x}$ is an arbitrary feasible point in $\mathcal{D}$, that is, $g_i(\tilde{x}) \leq 0$ and $h_j(\tilde{x}) = 0$. Also with $u \succeq 0$, we thus have

$$\sum_{i=1}^{m} u_i g_i(\tilde{x}) + \sum_{j=1}^{r} v_j h_j(\tilde{x}) \leq 0.$$

And this also means

$$L(\tilde{x}, u, v) = f(\tilde{x}) + \sum_{i=1}^{m} u_i g_i(\tilde{x}) + \sum_{j=1}^{r} v_j h_j(\tilde{x}) \leq f(\tilde{x}).$$

Therefore we have

$$g(u, v) = \inf_{x \in \mathcal{D}} L(x, u, v) \leq L(\tilde{x}, u, v) \leq f(\tilde{x}),$$

for $\forall \tilde{x}$ that are feasible for the constrains g_i and h_j. Since this inequality hold for arbitrary feasible $\tilde{x}$ and the primal optimal x^* must be among these feasible points, therefore the property equation (19) is proven.

This property means **for $\forall(u, v)$ with $u \succeq 0$, we are always providing an lower bound for the optimal primal value p^* of the primal problem**. It is inspiring! Then a natural follow-up question is: Can we choose some points (u, v) that provide the best lower bound for p^*. That is the motivation for **Lagrange dual problem**:

$$\begin{aligned} &\textit{maximize } g(u, v), \\ &\textit{subject to } u \succeq 0. \end{aligned} \tag{20}$$

Sometimes we would just call (20) the **dual problem**. Similarly, we define the point (u^*, v^*) with $u^* \succeq 0$ that maximize $g(u, v)$ as the **dual optimal**, or **dual solution**. And the optimal value $d^* = g(u^*, v^*)$ for $g(u, v)$ as the **optimal dual value**. Moreover, since we always have $g(u, v) \leq p^*$ for $\forall(u, v)$ with $u \succeq 0$, we thus know that

$$d^* \leq p^*. \tag{21}$$

At this end of duality part, we define the **duality gap** as the difference between optimal primal value p^* and optimal dual value d^*, that is, $p^* - d^*$, which will always be equal or larger than 0. When the gap is strictly positive, we would say the **weak duality** holds. While when the gap is equal to 0, that is, $p^* = d^*$, we would say the **strong duality** holds, which is very useful condition we then will use in convex optimization.

With the duality mentioned above, now we come back to the question **how to find the optimal x^* for the original primal problem**? Since the objective function will vary from different primal problems, there is not a formula to explicitly figure out x^*. However, we can find some conditions that x^* would satisfy. It would be even better if these conditions are necessary and sufficient so that we can obtain x^* by solving these conditions. What we need is actually the **Karrush-Kuhn-Tucker (KKT) conditions**. These conditions are strongly related to the (locally) primal optimal x^* and dual optimal (u^*, v^*). However, although commonly stated as necessary conditions (Wikipedia, 2016[21]) for optimal and dual optimal (x^*, u^*, v^*), in fact both of **necessity** and **sufficiency** need extra requirement beyond the KKT conditions, which we will discuss later. **For necessity, what we need is actually strong duality**. Therefore, let us firstly stating KKT conditions as **necessary condition** under strong duality:

- *Karush-Kuhn-Tucker (KKT) conditions* Suppose x^* and (u^*, v^*) are any (locally) primal and dual optimal points with strong duality holds (duality gap equals to 0), then we would have x^* and (u^*, v^*) satisfies the following four conditions:
  - *primal feasibility*: $g_i(x^*) \leq 0,\ h_j(x^*) = 0,\ \forall i \in \{1, ..., m\}\ \forall j \in \{1, ..., r\}$.
  - *dual feasibility*: $u \succeq 0$ (that is, $u_i \geq 0,\ \forall i \in \{1, ..., m\}$).
  - *complementary slackness*: $u_i g_i(x^*) = 0,\ \forall i \in \{1, ..., m\}$.
  - *stationarity*: $\nabla f(x^*) + \sum_{i=1}^{m} u_i \nabla g_i(x^*) + \sum_{j=1}^{r} v_j \nabla h_j(x^*) = 0$.

As is mentioned above, the necessity for KKT conditions need strong duality, which is not always guaranteed in most cases. However, fortunately when the primal problem is convex, we have a very loose condition that ensure the strong duality to hold for the optimization problem, named **Slater's condition**, and we will talk about it in the following part. **Moreover, under convex primal problem, it is also guaranteed that KKT conditions are sufficient conditions for (x^*, u^*, v^*) to be primal and dual optimal**. Thus, in this convex case, KKT conditions will be necessary and sufficient conditions for (x^* and (u^*, v^*) to be primal and dual optimal, and that is really what we want!

(IV) Convex Optimization, Slater's Condition and Subgradient

A function $f : \mathbb{R}^p \to \mathbb{R}$ is **convex** if for any two vectors x, x' in the domain of f and $\forall s \in [0, 1]$, we have

$$f(sx + (1 - s)x') \leq sf(x) + (1 - s)f(x'). \tag{22}$$

Moreover, f will be called **strictly convex** if for any $x \neq x'$ in the domain of f and $\forall s \in (0, 1)$,

$$f(sx + (1 - s)x') < sf(x) + (1 - s)f(x'). \tag{23}$$

For example, the absolute value function $f(x) = |x|$ is convex and but strictly convex. One important property of convex function is that **any local minima must also be global minimal**. This makes finding global minimal equivalent to finding a local minima, which is easier and more practical.

Let us take a step further. An optimization problem in (16) will be call **convex optimization problem** if all of f and $g_i,\ i \in \{1, .., m\}$ are convex functions. Suppose now we have an optimization problem, we plan to apply to KKT conditions to find the local optimal so as to find out the global optimal. As we

mentioned above, the necessity of KKT conditions need strong duality. Under convex optimization problem, **Slater's condition** will be helpful since it is **a sufficient condition for strong duality for a convex optimization problem** (Slater, 1950[1]):

- *Slater's condition* For the convex optimization problem

$$\begin{aligned} &minimize\ f(x), \\ &subject\ to\ g_i(x) \leq 0, \quad i \in \{1, 2, ..., m\}, \\ &\qquad\qquad h_j(x) = 0, \quad j \in \{1, 2, ..., r\}, \end{aligned}$$

  the strong duality hold if there $\exists x_0 \in dom(f) \cap \left(\cap_{i=1}^{m} dom(g_i) \right)$, the intersection of domains for f and $g_i,\ i \in \{1, ..., m\}$, such that

$$\begin{aligned} &g_i(x_0) < 0,\ i \in \{1, ..., m\}, \\ &h_j(x_0) = 0,\ j \in \{1, ..., r\}. \end{aligned}$$

Therefore, for a convex optimization problem, if it also satisfies the slater's condition so that strong duality holds, we would have KKT conditions as a necessary condition for x^* and (u^*, v^*) to be global primal and dual optimal. On the other hand, when the primal problem is convex, the KKT conditions are sufficient for the points to be primal and dual optimal (Boyd and Vandenberghe, 2004[8]) (the sufficiency actually need further discussion but this time I haven't study not much in it). Hence, in a nutshell, we have the following important conclusion:

- Conclusion For a convex optimization problem that satisfies Slater's condition, we would have the following two statements equivalent:

$$\begin{aligned} &x^*\ and\ (u^*, v^*)\ are\ global\ primal\ and\ dual\ optimal. \\ &\qquad\qquad \Leftrightarrow \\ &x^*\ and\ (u^*, v^*)\ satisfy\ KKT\ conditions. \end{aligned}$$

Now we get the necessary and sufficient condition for global optimal. That is, **when we want to find out the optimal solution x^* for the original convex primal problem that minimizes $f(x)$ under the constrains, we only need to find some points x^* and (u^*, v^*) such that if they satisfy the KKT conditions and Slater's condition, we can then state that x^* is the global optimal that we want**. That is really cool!

We are almost there! However, before we move back to the solution and computation for group Lasso, **there is still one problem if we plan to use KKT conditions to solve Lasso type optimization: differentiability**. Let us look back to the stationarity condition in KKT conditions:

$$\nabla f(x^*) + \sum_{i=1}^{m} u_i \nabla g_i(x^*) + \sum_{j=1}^{r} v_j \nabla h_j(x^*) = 0. \tag{24}$$

One important observation is that $g_i(x),\ i \in \{1, ..., m\}$ are not always differentiable, especially for Lasso problem, of which the constrain $g_i(x)$ may always contains terms like absolute value function $|x|$ that is not differentiable at point $x = 0$. To deal with this case we raise a new concept in the study of convex function: **subgradient** (or **subderivative** in one dimension). This idea is actually similar to the original meaning of first-order derivative or gradient, which provide a lower bound of the slope for increase at each specific point.

To have a better understand of subgradient, we firstly come to one dimensional subderivative. Suppose $f : \mathbb{R} \to \mathbb{R}$ is a convex function, then a **subderivative** of f at point x_0 is any real number z that satisfies

$$f(x) - f(x_0) \geq z(x - x_0), \quad \forall x \in \mathbb{R}. \tag{25}$$

That is, if we draw a straight line through the point x_0: $l(x) = f(x_0) + z(x - x_0)$, then any slope that enable its line to be beneath $f(x)$ for $\forall x \in \mathbb{R}$: $f(x) \geq l(x) = f(x_0) + z(x - x_0)$ is a subderivative at point x_0. Further, we would call the set contains all subderivative at point x_0 the **subdifferential** at point x_0, denoted as $\partial f(x_0)$:

$$\partial f(x_0) = \{z \in \mathbb{R} : f(x) - f(x_0) \geq z(x - x_0),\ \forall x \in \mathbb{R}\}. \tag{26}$$

When f is differential at point x_0, we would have the subdifferential at point x_0 will only contain one value, that is, the derivative $f'(x_0)$. Thus we come to realize that subdifferential is a more "generalized" form of derivative. An example for subdifferential will be when $f(x) = |x|$, we would have

$$\partial f(x) = \begin{cases} 1, & x > 0, \\ -1, & x < 0, \\ [-1, 1], & x = 0. \end{cases}$$

Now we come to the high-dimensional case, and a similar concept as subderivative will be defined which is now called subgradient. Suppose $f : \mathbb{R}^p \to \mathbb{R}$ is a convex function, then a **subgradient** of f at point x_0 is any vector $z \in \mathbb{R}^p$ such that

$$f(x) - f(x_0) \geq \langle z, x - x_0 \rangle, \quad \forall x \in \mathbb{R}^p. \tag{27}$$

Also, **subdifferential** at point x_0 is now defined as

$$\partial f(x_0) = \{z \in \mathbb{R}^p : f(x) - f(x_0) \geq \langle z, x - x_0 \rangle,\ \forall x \in \mathbb{R}^p\}. \tag{28}$$

And we would also have when f is differential at point x_0, the subdifferential at point x_0 will only contain one value, which is the gradient $\nabla f(x_0)$. Now for a convex function $f : \mathbb{R}^p \to \mathbb{R}$, we have the following important remark (Wikipedia, 2016[22]):

- Remark A point x_0 is a global optimal for a convex function f if and only if zero is contained in the subdifferential $\partial f(x_0)$ at this point.

With this remark, we finally can extend the KKT conditions to deal with the non-differentiable case, where the extended stationarity condition now is stated as

- *(extended) stationarity*:

$$0 \in \partial f(x^*) + \sum_{i=1}^{m} u_i \partial g_i(x^*) + \sum_{j=1}^{r} v_j \partial h_j(x^*). \tag{29}$$

Finally, we are ready to come with sufficient concepts to solve the problem of solution and computation for group Lasso!

(V) Solution and Computation for Group Lasso

As we define in equation (15), the objective function for group Lasso is

$$\hat{\beta}_{group} = \arg\min_{\beta} \ \Big\{||y - \sum_{j=1}^{J} X_j\beta_j||_2^2 + \lambda \sum_{j=1}^{J} \sqrt{p_j}||\beta_j||_2\Big\},$$

where $\beta = (\beta_1, \beta_2, ..., \beta_J)^T$ and each β_j is a p_j-dimensional vector. Transform it into the convex optimization problem standard form (16), we would have

$$f(\beta) = f(\beta_1, \beta_2, ..., \beta_J) = ||y - \sum_{j=1}^{J} X_j\beta_j||_2^2 \tag{30}$$

$$g(\beta) = g(\beta_1, \beta_2, ..., \beta_J) = \sum_{j=1}^{J} \sqrt{p_j}||\beta_j||_2 - t. \tag{31}$$

Notice that the t in equation (31) is actually the original constrain on β such that $\sum_{j=1}^{J} \sqrt{p_j}||\beta_j||_2 \leq t, t \geq 0$. Actually the reason why group Lasso can be written in the form (15) is that we would always **let $t\lambda$ as a constant (it is my own understanding)**, so it can be reduce when we just plan to minimize the objective function. That is also intuitive, since smaller t will always means we need larger penalty on β, which means larger λ.

Now since we want to use KKT conditions to solve this convex optimization problem, and also $g_j(\beta)\lambda\sqrt{p_j}||\beta_j||_2, \ j \in \{1, ..., J\}$, is non-differentiable at the all-zero vector, therefore **the first thing we should do is to calculate the subdifferential for both f and g_j so as to state the stationarity condition for optimal**. Let us begin with taking subdifferential of $f(\beta)$ (30). Since $\{\beta_1, \beta_2, ..., \beta_J\}$ are separable in the objective function (15), we can thus only take the subdifferential on β_j, which we will also denote as $\frac{\partial}{\partial\beta_j}$:

$$\begin{aligned}
\frac{\partial}{\partial\beta_j} f(\beta) =& \frac{\partial}{\partial\beta_j}\Big\{(y - \sum_{i=1}^{J} X_i\beta_i)^T(y - \sum_{i=1}^{J} X_i\beta_i)\Big\} \\
=& \frac{\partial}{\partial\beta_j}\Big\{y^Ty - 2\sum_{i=1}^{J}\beta_i^T x_i^T y + (\sum_{i=1}^{J}\beta_i^T X_i^T)(\sum_{i=1}^{J} X_i\beta_i)\Big\} \\
=& -2X_j^T y + 2\sum_{i=1, i\neq j}^{J} X_j^T X_i\beta_i + 2X_j^T X_j\beta_j \\
=& -2X_j^T(y - \sum_{i=1}^{J} X_i\beta_i), \quad j \in \{1, ..., J\}
\end{aligned} \tag{32}$$

Taking the subdifferential of $g(\beta), \ j \in \{1, ..., J\}$ (31) is a litter more tricky. First since $g(\beta)$ is separable with respect to β_j, we can rewrite $g(\beta)$ as $g(\beta) = \sum_{j=1}^{J} g_j(\beta_j)$, where $g_j(\beta_j) = \sqrt{p_j}||\beta_j||_2$ is only relate to β_j. Thus we have $\frac{\partial}{\partial\beta_j} g(\beta) = \frac{\partial}{\partial\beta_j} g_j(\beta_j)$, and then we could have the following discussions:

- When $\beta_j = (\beta_{j1}, \beta_{j2}, ..., \beta_{jp_j})^T \neq 0$, that is, not all element of β_j is 0, we would have

$$\frac{\partial}{\partial \beta_j} g_j(\beta) = \frac{\partial}{\partial \beta_j} \sqrt{p_j \sum_{k=1}^{p_j} \beta_{jk}^2} = \begin{pmatrix} \frac{\partial}{\partial \beta_{j1}} \lambda \sqrt{p_j \sum_{k=1}^{p_j} \beta_{jk}^2} \\ \frac{\partial}{\partial \beta_{j2}} \lambda \sqrt{p_j \sum_{k=1}^{p_j} \beta_{jk}^2} \\ \vdots \\ \frac{\partial}{\partial \beta_{jp_j}} \lambda \sqrt{p_j \sum_{k=1}^{p_j} \beta_{jk}^2} \end{pmatrix} = \frac{\sqrt{p_j}}{\sqrt{\sum_{k=1}^{p_j} \beta_{jk}^2}} \begin{pmatrix} \beta_{j1} \\ \beta_{j2} \\ \vdots \\ \beta_{jk} \end{pmatrix} = \frac{\sqrt{p_j}\beta_j}{||\beta_j||_2}. \tag{33}$$

- When $\beta_j = (\beta_{j1}, \beta_{j2}, ..., \beta_{jp_j})^T = 0$, that is, β_j is a zero vector, denote $z \in \mathbb{R}^{p_j}$ as any subgradient of g_j at point $\beta_j = 0$, and then according to the definition of subgradient (27), we would have

$$g_j(\beta') - g_j(0) \geq \langle z, \beta' \rangle, \quad \forall \beta' \in \mathbb{R}^{p_j}. \tag{34}$$

  Since $g_j(\beta_j) = \sqrt{p_j}||\beta_j||_2$ and $g_j(0) = 0$, the solving equation (34) is equivalent with finding $z = (z_1, z_2, ..., z_{p_j})^T$ such that

$$\begin{aligned} & \sqrt{p_j}||\beta'||_2 \geq \langle z, \beta' \rangle \\ \Leftrightarrow \quad & p_j||\beta'||_2^2 \geq |\langle z, \beta' \rangle|^2 \\ \Leftrightarrow \quad & ||\beta'||_2^2 \geq |\langle \frac{z}{\sqrt{p_j}}, \beta' \rangle|^2, \quad \forall \beta' \in \mathbb{R}^{p_j} \end{aligned} \tag{35}$$

  Now we state that $||\frac{z}{\sqrt{p_j}}||_2^2 \leq 1$ is sufficient and necessary solution for the above equation (35), and provide the proof as follow:

  - *Sufficiency*: If condition $||\frac{z}{\sqrt{p_j}}||_2^2 \leq 1$ holds, by using Cauchy-Schwarz inequality, we would have

$$|\langle \frac{z}{\sqrt{p_j}}, \beta' \rangle|^2 \leq ||\frac{z}{\sqrt{p_j}}||_2^2 ||\beta'||_2^2 \leq ||\beta'||_2^2.$$

    Hence sufficiency is proved.

  - *Necessity*: Suppose z is a solution for equation (35) while $||\frac{z}{\sqrt{p_j}}||_2^2 > 1$. Since we know for Cauchy-Schwarz inequality, the equality holds if and only if $\frac{z}{\sqrt{p_j}}$ and β' is linear dependent, therefore by the fact that β' is arbitrary in $\mathbb{R}_{p_j}$, we know there must $\exists \beta_0' \in \mathbb{R}_{p_j}$ such that the equality for Cauchy-Schwarz inequality holds. That is, for β_0', we would have

$$|\langle \frac{z}{\sqrt{p_j}}, \beta_0' \rangle|^2 = ||\frac{z}{\sqrt{p_j}}||_2^2 ||\beta_0'||_2^2 \geq ||\frac{z}{\sqrt{p_j}}||_2^2 ||\beta_0'||_2^2,$$

    which is in contradiction to equation (35). Therefore, if z is a solution, we must have $||\frac{z}{\sqrt{p_j}}||_2^2 \leq 1$. Hence necessity is proved.

  Therefore, the subgradient for $g_j(\beta_j)$ at point $\beta_j = 0$ is any vector z such that $||\frac{z}{\sqrt{p_j}}||_2^2 \leq 1$. We thus have the subdifferential for $g_j(\beta_j)$ at point 0 as

$$\frac{\partial}{\partial \beta_j} g_j(0) = \{z : z \in \mathbb{R}^{p_j}, ||\frac{z}{\sqrt{p_j}}||_2^2 \leq 1\}. \tag{36}$$

Combining the subdifferential of $f(\beta)$ (32), and $g(\beta)$ (33,36) on β_j, **we now have the extended stationarity condition (29) in KKT conditions written as**:

$$0 \in \{\frac{\partial}{\partial \beta_j} f(\beta) + \lambda \frac{\partial}{\partial \beta_j} g(\beta)\}\Big|_{\beta=\hat{\beta}} = \begin{cases} -2X_j^T(y - \sum_{i=1}^J X_i \hat{\beta}_i) + \frac{\lambda \sqrt{p_j}\hat{\beta}_j}{||\hat{\beta}_j||_2}, & \hat{\beta}_j \neq 0, \\ -2X_j^T(y - \sum_{i=1}^J X_i \hat{\beta}_i) + \{\lambda z : z \in \mathbb{R}^{p_j}, ||\frac{z}{\sqrt{p_j}}||_2^2 \leq 1\}, & \hat{\beta}_j = 0, \end{cases} \tag{37}$$

where $j = 1, ..., J$, and $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_J)^T$ is the optimal solution for the group Lasso problem (15). Motice that in the definition (31) of $g(\beta) = \sum_{j=1}^{J} \sqrt{p_j}||\beta_j||_2 - t$, we could easily choose a small β' such that $g(\beta') < 0$, therefore the **Slater's condition holds**. Also with the fact that it is a convex problem, we thus now know that **in this case the KKT conditions, especially stationarity condition (37), are both sufficient and necessary for solution $\hat{\beta}$ to be global optimal**.

A step further, to simplify the stationarity condition, we could rewrite (37) into the form

$$-2X_j^T(y - \sum_{i=1}^{J} X_i\hat{\beta}_i) + \lambda\sqrt{p_j}s_j = 0, \quad j \in \{1, ..., J\}, \tag{38}$$

where $s_j = \frac{\hat{\beta}_j}{||\hat{\beta}_j||_2}$ if $\hat{\beta}_j = 0$ and s_j is any vector with $||s_j||_2 \leq 1$ otherwise. Now we want to find the optimal $\hat{\beta}_j$ and assume the other $\hat{\beta}_k$, $k \neq j$ is already known, that is, assume we have already known $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_J)$ except $\hat{\beta}_j$, then we would first define a new notation r_j:

$$r_j = y - \sum_{i=1, i\neq j}^{J} X_i\hat{\beta}_i. \tag{39}$$

Thus the stationarity condition (38) now can be written as

$$-2X_j^T(r_j - X_j\hat{\beta}_j) + \lambda\sqrt{p_j}s_j = 0, \quad j \in \{1, ..., J\}. \tag{40}$$

To solve the above condition we have the following discussion:

- When $\hat{\beta}_j$=0, then in this case we would have (40) is equivalent to $-2X_j^T r_j + \lambda\sqrt{p_j}s_j = 0$, where s_j is a vector with $||s_j||_2 \leq 1$. That is, we would have the condition equivalent to

$$||X_j^T r_j||_2 \leq \frac{\lambda\sqrt{p_j}}{2}. \tag{41}$$

  Since now the KKT conditions are sufficient and necessary for optimal solution $\hat{\beta}_j$, therefore we thus know that **when (41) holds, $\hat{\beta}_j = 0$ will be our optimal**. The term $||X_j^T r_j||_2$ is often named **entry term**, because **when this term is smaller than a specific number as (41) does, we would have all coefficients in that group $\hat{\beta}_j$ equal to 0**, which is actually an important property of group Lasso.

- When $\hat{\beta}_j \neq 0$, then in this case we have $s_j = \frac{\hat{\beta}_j}{||\hat{\beta}_j||_2}$. Thus now we want to solve (40) with the form

$$-2X_j^T(r_j - X_j\hat{\beta}_j) + \lambda\sqrt{p_j}\frac{\hat{\beta}_j}{||\hat{\beta}_j||_2} = 0,$$

  which directly leads to the solution:

$$\hat{\beta}_j = \left(X_j^T X_j + \frac{\lambda\sqrt{p_j}}{2||\hat{\beta}_j||_2}\right)^{-1} X_j^T r_j. \tag{42}$$

With these discussion and combining the conclusion in (41) and (42), we now can have our numerical solution for group Lasso in general case as follow:

- Solutions for Group Lasso in General Cases For a group Lasso problem, suppose we have already know the optimal solution for $(\hat{\beta}_1, ..., \hat{\beta}_{j-1}, 0, \hat{\beta}_{j+1}, ..., \hat{\beta}_J)$ except $\hat{\beta}_j$, then we could have the optimal solution for $\hat{\beta}_j$ as

$$\hat{\beta}_j = \begin{cases} 0 & , \ ||X_j^T r_j||_2 \leq \frac{\lambda\sqrt{p_j}}{2}, \\ \left(X_j^T X_j + \frac{\lambda\sqrt{p_j}}{2||\hat{\beta}_j||_2} I\right)^{-1} X_j^T r_j & , \ otherwise. \end{cases} \tag{43}$$

Now we got the solution for group Lasso in general case. However, as we can seen in (43), the solution for $\hat{\beta}_j$ still have term $\frac{1}{||\hat{\beta}_j||_2}$ in it, which is depend on $\hat{\beta}_j$. To further simply this solution, Yuan and Lin (2006[10]) made an assumption in their original paper for group Lasso. They **assume samples are orthonormal within each group, that is, $X_j^T X_j = I$ for $j \in \{1, ..., J\}$.** In this case, let us take a deeper look into the solution when $||X_j^T r_j||_2 > \frac{\lambda\sqrt{p_j}}{2}$:

$$\hat{\beta}_j = \left(X_j^T X_j + \frac{\lambda\sqrt{p_j}}{2||\hat{\beta}_j||_2} I\right)^{-1} X_j^T r_j.$$

Since $X_j^T X_j = I$, we thus have

$$\left(X_j^T X_j + \frac{\lambda\sqrt{p_j}}{2||\hat{\beta}_j||_2} I\right)^{-1} = \left(I + \frac{\lambda\sqrt{p_j}}{2||\hat{\beta}_j||_2} I\right)^{-1} = \left[(1 + \frac{\lambda\sqrt{p_j}}{2||\hat{\beta}_j||_2})I\right]^{-1} = \frac{1}{1 + \frac{\lambda\sqrt{p_j}}{2||\hat{\beta}_j||_2}} I.$$

By combining the above two equations, we thus now have

$$\hat{\beta}_j = \frac{1}{1 + \frac{\lambda\sqrt{p_j}}{2||\hat{\beta}_j||_2}} X_j^T r_j. \tag{44}$$

The following step in the most tricky one in this section! Let take ℓ_2-norm $||\cdot||_2$ on both sides of the above equation (44), then we would have

$$\begin{aligned} ||\hat{\beta}_j||_2 = ||\frac{1}{1 + \frac{\lambda\sqrt{p_j}}{2||\hat{\beta}_j||_2}} I X_j^T r_j||_2 \quad &\Rightarrow \quad ||\hat{\beta}_j||_2 = \frac{1}{1 + \frac{\lambda\sqrt{p_j}}{2||\hat{\beta}_j||_2}} ||X_j^T r_j||_2 \\ &\Rightarrow \quad ||\hat{\beta}_j||_2 + \frac{\lambda\sqrt{p_j}}{2} = ||X_j^T r_j||_2 \\ &\Rightarrow \quad ||\hat{\beta}_j||_2 = ||X_j^T r_j||_2 - \frac{\lambda\sqrt{p_j}}{2}. \end{aligned} \tag{45}$$

With the above equation (45), we further simplify $\hat{\beta}_j$ as

$$\hat{\beta}_j = \frac{1}{1 + \frac{\lambda\sqrt{p_j}}{2||X_j^T r_j||_2 - \lambda\sqrt{p_j}}} X_j^T r_j = \left(1 - \frac{\lambda\sqrt{p_j}}{2||X_j^T r_j||_2}\right) X_j^T r_j. \tag{46}$$

We are almost there! Magically, notice that when $||X_j^T r_j||_2 \leq \frac{\lambda\sqrt{p_j}}{2}$, solutions of $\hat{\beta}_j$ can also be written as

$$\hat{\beta}_j = 0 = \left(1 - \frac{\lambda\sqrt{p_j}}{2||X_j^T r_j||_2}\right)_+ X_j^T r_j.$$

Therefore we finally come to the solution for $\hat{\beta}_j$ under within group orthonormality $X_j^T X_j = I$:

- Solutions for Group Lasso in Within-Group Orthonormal Cases When the sample are orthonormal each within group, that is, $X_j^T X_j = I$ for $j \in \{1, ..., J\}$, we could have the optimal solutions for $\hat{\beta}_j$ as

$$\hat{\beta}_j = \Big(1 - \frac{\lambda\sqrt{p_j}}{2||X_j^T r_j||_2}\Big)_+ X_j^T r_j. \tag{47}$$

What a beautiful solution! Now we can see the "group" property for group Lasso we mentioned before from (47): when $||X_j^T r_j||_2 \leq \frac{\lambda\sqrt{p_j}}{2}$, we would have $\hat{\beta}_j = 0$, that is, **all coefficients for variables in the j-th group will become 0 simultaneously**; otherwise, the coefficients for variables in j-th group will be $\hat{\beta}_j = \Big(1 - \frac{\lambda\sqrt{p_j}}{2||X_j^T r_j||_2}\Big) X_j^T r_j$, of which **almost every coefficient will be intuitively non-zero**.

(VI) Connect the Solutions of Lasso and Group Lasso

We know Lasso regression is actually a special form of group Lasso regression where each group contain only one element, that is, $p_j = 1$ for $j \in 1, ..., J$. We plan to end the discussion of this chapter as well as providing some insight by connecting the solutions of Lasso regression and group Lasso regression.

Let us come back to the setting of Lasso with n samples and J variables. Denote $X = (x_1, ..., x_J)$ as the design matrix where x_j is a n-dimensional column vector represent the j-th variable. We also **assume orthonormality between variables** such that we have $X^T X = I$. In this case, we have proven the closed form solution for Lasso estimator in (10):

$$\begin{aligned}\hat{\beta}_j =& \operatorname{sgn}(\hat{\beta}_j^{OLS})\Big(|\hat{\beta}_j^{OLS}| - \frac{\lambda}{2}\Big)_+ \\ =& \operatorname{sgn}(x_j^T y)\Big(|x_j^T y| - \frac{\lambda}{2}\Big)_+,\end{aligned} \tag{48}$$

where $\hat{\beta}^{OLS} = (X^T X)^{-1} X^T y = X^T y$ and thus $\hat{\beta}_j^{OLS} = x_j^T y$. Then we come to the group Lasso. **Since each group only 1 variable, therefore the within-group orthonormality is naturally satisfied**. So according to (47), we have

$$\begin{aligned}\hat{\beta}_j =& \Big(1 - \frac{\lambda\sqrt{p_j}}{2||X_j^T r_j||_2}\Big)_+ X_j^T r_j \\ =& \Big(1 - \frac{\lambda}{2|x_j^T r_j|}\Big)_+ x_j^T r_j.\end{aligned} \tag{49}$$

By "connect the solutions" we actually mean to show that (49) is a same solution as (48). Since X is orthonormal, we have

$$x_j^T x_k = 0, \quad \forall k \neq j.$$

Therefore we could simplify $x_j^T r_j$ as

$$x_j^T r_j = x_j^T \big(y - \sum_{k \neq j}^{J} \hat{\beta}_k x_k\big) = x_j^T y - \sum_{k \neq j}^{J} \hat{\beta}_k x_j^T x_k = x_j^T y.$$

And hence we have

$$\Big(1 - \frac{\lambda}{2|x_j^T r_j|}\Big)_+ x_j^T r_j = \Big(1 - \frac{\lambda}{2|x_j^T y|}\Big)_+ x_j^T y = \Big(|x_j^T y| - \frac{\lambda}{2}\Big)_+ \frac{x_j^T y}{|x_j^T y|} = \operatorname{sgn}(x_j^T y)\big(|x_j^T y| - \frac{\lambda}{2}\big)_+, \tag{50}$$

which finally states the solutions (48) for Lasso regression and those (49) for group Lasso regression with groups containing only one element each is the same.

Extremely nice work!

References

[1] Slater, M. (1950). Lagrange Multipliers Revisited, Cowles Commis (Vol. 403). *sion Discussion Paper, Mathematics*.

[2] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.

[3] Frank, L. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2), 109-135.

[4] Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37(4), 373-384.

[5] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

[6] Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning (Vol. 1)*. Springer, Berlin: Springer series in statistics.

[7] Bishop, C. M., & Tipping, M. E. (2003). Bayesian regression and classification. *Nato Science Series sub Series III Computer And Systems Sciences*, 190, 267-288.

[8] Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

[9] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301-320.

[10] Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1), 49-67.

[11] Friedman, J., Hastie, T., & Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. *arXiv preprint* arXiv:1001.0736.

[12] Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3), 273-282.

[13] Buhlmann, P., & Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

[14] Simon, N., & Tibshirani, R. (2012). Standardization and the group lasso penalty. *Statistica Sinica*, 22(3), 983.

[15] Gordon, G., & Tibshirani, R. (2012). Karush-kuhn-tucker conditions. *Optimization*, 10(725/36), 725.

[16] Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2), 231-245.

[17] Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7, 1456-1490.

[18] Reid, S., & Tibshirani, R. (2015). Sparse regression and marginal testing using cluster prototypes. *Biostatistics*, kxv049.

[19] Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC Press.

[20] Slater's condition. (2016, March 13). In *Wikipedia, The Free Encyclopedia*. `https://en.wikipedia.org/w/index.php?title=Slater%27s_condition&oldid=709859313`

[21] Duality (optimization). (2016, April 5). In *Wikipedia, The Free Encyclopedia*. `https://en.wikipedia.org/w/index.php?title=Duality_(optimization)&oldid=713801186`

[22] Subderivative. (2016, March 18). In *Wikipedia, The Free Encyclopedia*. `https://en.wikipedia.org/w/index.php?title=Subderivative&oldid=710770961`.