



### 5.2.2 The least absolute shrinkage and selection operator (LASSO, Tibshirani 1996)

This subsection is devoted to discussions on LASSO, a method that combines the least-square loss with an  $\ell_1$ -constraint, or bound on the sum of the absolute values of the coefficients. Relative to the least-squares solution, this constraint has the effect of shrinking the coefficients and even setting some to zero. In this way, it provides an automatic way for doing model selection in linear regression. Moreover, unlike some other criteria for model selection, the resulting optimization problem is convex, and can be solved efficiently for large problems.

Recall that the linear regression model has the form

$$Y = \beta_0 + \sum_{j=1}^p Z_j \beta_j + \varepsilon.$$

Shrinkage method for variable selection is a continuous process. It is more stable, not suffering from high variability. It has one unified selection framework and easy to make inferences. The lasso is a shrinkage approach like ridge, with subtle but important difference. The LASSO (least absolute shrinkage and selection operator) introduced by Tibshirani (1996, JRSSB) is defined as the  $\ell_1$  optimization problem:


$$\underset{\beta}{\text{minimize}} \quad (\mathbf{Y} - \mathbf{Z}\beta)^\top (\mathbf{Y} - \mathbf{Z}\beta), \quad \text{s.t.} \quad \sum_{j=1}^p |\beta_j| \leq t, \quad \text{pre-sel.} \quad (1)$$

by applying the absolute penalty or magnitude constraint on the parameters. It is often convenient to rewrite the LASSO problem in the so-called Lagrangian form:

$$\underset{\beta}{\text{minimize}} \quad \sum_{i=1}^n (Y_i - \mathbf{Z}_i \beta)^2 + \underbrace{\lambda \sum_{j=1}^p |\beta_j|}_{\text{penalization.}} \quad \text{or} \quad \underset{\beta}{\text{minimize}} \quad (\mathbf{Y} - \mathbf{Z}\beta)^\top (\mathbf{Y} - \mathbf{Z}\beta) + \lambda \|\beta\|_1, \quad =: L(\beta) \quad (2)$$

where  $\lambda \geq 0$  is the so-called tuning parameter. The resulting lasso estimate is denoted by  $\hat{\beta}_\lambda^{\text{lasso}}$ . By Lagrangian duality, there is a one-to-one correspondence between the constrained problem (1) and the Lagrangian form (2): for each value of  $t$  in the range where the constraint  $\|\beta\|_1 \leq t$  is active, there is a corresponding value of  $\lambda$  that yields the same solution from

the Lagrangian form (2). Conversely, the solution  $\hat{\beta}_\lambda^{\text{lasso}}$  solves the bound problem with  $t = \|\hat{\beta}_\lambda^{\text{lasso}}\|_1$ .

 *Remark 4.* Note that in many description of the LASSO, there is a factor  $1/(2n)$  or  $1/n$  appearing in front of  $(\mathbf{Y} - \mathbf{Z}\beta)^\top(\mathbf{Y} - \mathbf{Z}\beta)$  in (1) and (2). Though this makes no difference in (1), and corresponds to a simple reparametrization of  $\lambda$  in (2), this kind of standardization makes  $\lambda$  values comparable for different sample sizes (useful for cross-validation).

In the signal processing literature, the lasso is also known as *basis pursuit*. Note that similar to the ridge regression, we typically center the response and standardize the predictors so that each column is centered ( $\frac{1}{n} \sum_{i=1}^n Z_{ij} = 0$ ) and has unit variance ( $\frac{1}{n} \sum_{i=1}^n Z_{ij}^2 = 1$ ). Thereafter, we fit a model without an intercept. Notice the similarity to the ridge regression problem: the  $L_2$  ridge penalty  $\sum_{i=1}^p \beta_j^2$  is replaced by the  $L_1$  lasso penalty  $\sum_{i=1}^p |\beta_j|$ . The tuning parameter  $\lambda$  controls the amount of shrinkage; the larger  $\lambda$ , the greater amount of shrinkage. In particular,  $\lambda \rightarrow 0$ , the lasso tends to the OLS;  $\lambda \rightarrow \infty$ , the lasso estimate tends to 0 and there is intercept only.

removing the reliance to the measurement unit.

The theory of convex analysis tells us that necessary and sufficient conditions for a solution to problem (2) take the form

$$\nabla_{\beta} L(\beta) = -2\mathbf{Z}^\top(\mathbf{Y} - \mathbf{Z}\beta) + \lambda\mathbf{s} = 0 \quad \text{or} \quad -\langle \mathbf{z}_j, \mathbf{Y} - \mathbf{Z}\beta \rangle + \lambda s_j = 0, \quad j = 1, \dots, p. \quad (3)$$

Here each  $s_j$  is an unknown quantity equal to  $\text{sign}(\beta_j)$  if  $\beta_j \neq 0$  and some value lying in  $[-1, 1]$  otherwise, that is, it is a subgradient for the absolute value function (see Chapter 5 of the book “*Statistical learning with sparsity*” for more details). In other words, the solution  $\hat{\beta}_\lambda^{\text{lasso}}$  to problem (2) are the same as solutions  $(\hat{\beta}, \hat{s})$  to problem (3). This system is a form of the so-called **Krush-Kuhn-Tucker (KKT) conditions** for problem (2). Expressing a problem in subgradient form can be useful for designing algorithms for finding its solutions.

Perhaps more importantly, a close look at (1) reveals that, because of the nature of the  $L_1$  constraint, making  $t$  sufficiently small will cause some of the coefficients to be exactly zero. Thus, the lasso does a kind of continuous subset selection or variable selection. A geometric illustration of why LASSO results in sparsity, but ridge does not, is given by the constraint interpretation of their penalties:

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

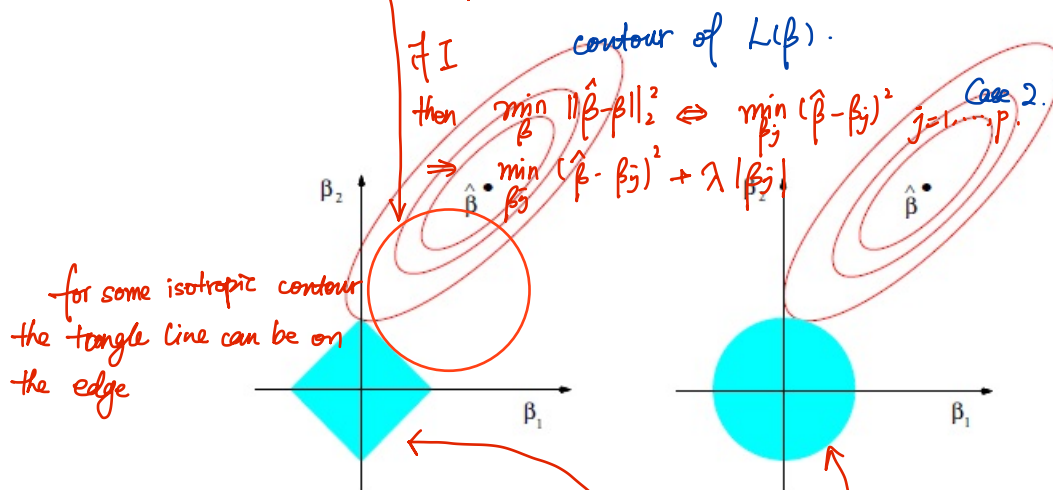


FIGURE 3.11. Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions  $|\beta_1| + |\beta_2| \leq t$  and  $\beta_1^2 + \beta_2^2 \leq t^2$ , respectively, while the red ellipses are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

Case 1 if  $\beta_j \geq 0$ , then

$$\beta_j^* = \hat{\beta}_j - \frac{\lambda}{2}$$

$$\hat{\beta}_j^{\text{LASSO}} = \hat{\beta}_j - \frac{\lambda}{2} \text{ if } \beta_j^* > 0$$

$$\hat{\beta}_j^{\text{LASSO}} = 0 \text{ if } \beta_j^* \leq 0$$

if  $\beta_j < 0$ , then

$$\beta_j^* = \hat{\beta}_j + \frac{\lambda}{2}$$

$$\hat{\beta}_j^{\text{LASSO}} = \hat{\beta}_j + \frac{\lambda}{2} \text{ if } \beta_j^* < 0$$

$$\hat{\beta}_j^{\text{LASSO}} = 0 \text{ if } \beta_j^* \geq 0$$

Case 2.

Often, we believe that many of the  $\beta_j$ 's should be 0. We use the term *sparse* for a model with few nonzero coefficients. Hence, we seek a set of **sparse solutions**. Large enough  $\lambda$  (or small enough  $t$ ) will force some coefficients exactly to be 0. The LASSO can perform model selection for us. Hence, a key property of the  $L_1$ -constraint is its ability to yield sparse solutions. This idea can be applied in many different statistical models.

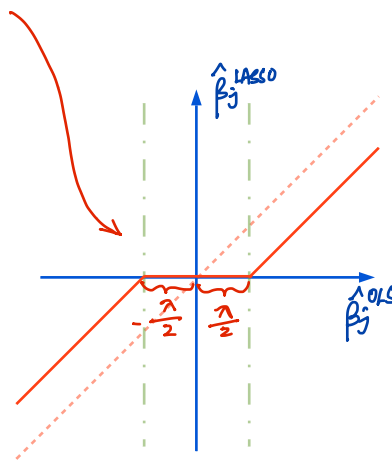
For illustration, consider an orthogonal design case with  $Z^T Z = I$ . The LASSO method is equivalent to: solve  $\beta_j$ 's componentwisely by solving

$$\lambda \uparrow \quad \#\{\hat{\beta}_j^{\text{LASSO}} = 0\} \uparrow.$$

$$\min_{\beta_j} (\beta_j - \hat{\beta}_j^{\text{ls}})^2 + \lambda |\beta_j|$$

The solution to the above problem is *no scaling*.

$$\begin{aligned}
 \hat{\beta}_j^{\text{lasso}} &= \text{sgn}(\hat{\beta}_j^{\text{ls}}) (|\hat{\beta}_j^{\text{ls}}| - \frac{\lambda}{2})_+ \\
 &= \begin{cases} \hat{\beta}_j^{\text{ls}} - \frac{\lambda}{2} & \text{if } \hat{\beta}_j^{\text{ls}} > \frac{\lambda}{2} \\ 0 & \text{if } |\hat{\beta}_j^{\text{ls}}| \leq \frac{\lambda}{2} \\ \hat{\beta}_j^{\text{ls}} + \frac{\lambda}{2} & \text{if } \hat{\beta}_j^{\text{ls}} < -\frac{\lambda}{2} \end{cases}
 \end{aligned}$$



Note that it shrinks big coefficients by a constant  $\frac{\lambda}{2}$  towards zero, and truncates small coefficients to zero exactly, whereas ridge regression does a proportional shrinkage.

*Remark 5.* In the case of an orthonormal design/input matrix  $\mathbf{Z}$ , the ridge estimate and the lasso have explicit solutions. The following is some detailed comparison of different methods under orthogonal design (i.e.  $\mathbf{Z}^\top \mathbf{Z} = \mathbf{I}$ ):

- Computational intractable (NP hard)*
- Best subset (of size  $M$ ):  $\hat{\beta}_j^{\text{ls}}$ , keeps the largest coefficients.
  - Ridge regression:  $\hat{\beta}_j^{\text{ls}}/(1 + \lambda)$ , does a **proportional shrinkage**.
  - LASSO:  $\text{sgn}(\hat{\beta}_j^{\text{ls}})(|\hat{\beta}_j^{\text{ls}}| - \frac{\lambda}{2})_+$ , transform each coefficient by a constant factor first, then truncate it at zero with a certain threshold. This is called “**Soft thresholding**”, used often in wavelet-based smoothing.

**Computation of the LASSO solution** The lasso problem is a convex program, specially a quadratic program (QP) with a convex constraint. As such, there are many sophisticated QP methods for solving the lasso. However, there is a particularly simple and effective computational algorithm, that gives insights into how the lasso works. For convenience, we rewrite the criterion in Lagrangian form:

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^n (y_i - \sum_{j=1}^p Z_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j|. \quad (4)$$

By convention, we will assume *scaling* that both  $Y$  and the predictors  $z_j$  have been standardized. In this case, the intercept term  $\beta_0$  can be omitted. The Lagrangian form is especially convenient for numerical computation of the solution by a simple procedure known as *coordinate descent*.

### Single predictor: Soft thresholding.

Let us first consider a single predictor setting, based on samples  $\{Z_i, y_i\}_{i=1}^n$ . The problem then is to solve

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^n (y_i - Z_i \beta)^2 + \lambda |\beta|. \quad (5)$$

The standard approach to this univariate minimization problem would be to take the gradient (first derivative) with respect to  $\beta$ , and set it to zero. There is a complication, however,

because the absolute value function  $|\beta|$  does not have a derivative at  $\beta = 0$ . However, we can proceed by direct inspection of the function (5), and find that

$$\hat{\beta} = \begin{cases} \frac{1}{n}\langle \mathbf{Z}, \mathbf{Y} \rangle - \lambda & \text{if } \frac{1}{n}\langle \mathbf{Z}, \mathbf{Y} \rangle > \lambda, \\ 0 & \text{if } \frac{1}{n}\langle \mathbf{Z}, \mathbf{Y} \rangle \leq \lambda, \\ \frac{1}{n}\langle \mathbf{Z}, \mathbf{Y} \rangle + \lambda & \text{if } \frac{1}{n}\langle \mathbf{Z}, \mathbf{Y} \rangle < -\lambda, \end{cases}$$

which can write succinctly as

$$\hat{\beta} = S_{\lambda}\left(\frac{1}{n}\langle \mathbf{Z}, \mathbf{Y} \rangle\right).$$

Here the soft-thresholding operator  $S_{\lambda}(x) = \text{sgn}(x)(|x| - \lambda)_+$  translates its argument  $x$  toward zero by the amount  $\lambda$ , and sets it to zero if  $|x| \leq \lambda$ .

### Multiple predictors: Cyclic coordinate descent

Using this intuition from the univariate case, we can now develop a simple coordinatewise scheme for solving the full lasso problem (2). More precisely, we repeatedly cycle through the predictors in some fixed (but arbitrary) order (say  $j = 1, \dots, p$ ), where at the  $j$ th step, we update the coefficient  $\beta_j$  by minimizing the objective function in this coordinate while holding fixed all other coefficients  $\{\hat{\beta}_k, k \neq j\}$  at their current values.

Writing the objective in (2) as

$$\frac{1}{2n} \sum_{i=1}^n (y_i - \sum_{k \neq j} Z_{ik} \beta_k - Z_{ij} \beta_j)^2 + \lambda \sum_{k \neq j} |\beta_k| + \lambda |\beta_j|, \quad (6)$$

we see that solution for each  $\beta_j$  can be expressed succinctly in terms of the partial residual  $r_i^{(j)} = y_i - \sum_{k \neq j} x_{ik} \hat{\beta}_k$ , which removes from the outcome the current fit from all but the  $j$ th predictor. In terms of the partial residual, the  $j$ th coefficient is updated as

$$\hat{\beta}_j = S_{\lambda}\left(\frac{1}{n}\langle \mathbf{Z}_j, \mathbf{r}^{(j)} \rangle\right).$$

Equivalently, the update can be written as

$$\hat{\beta}_j \leftarrow S_{\lambda}\left(\hat{\beta}_j + \frac{1}{n}\langle \mathbf{Z}_j, \mathbf{r} \rangle\right),$$

where  $r_i = y_i - \sum_{i=1}^p Z_{ij} \hat{\beta}_j$  are the full residuals. The overall algorithm operates by applying the above soft-thresholding update repeatedly in a cyclical manner, updating the coordinates of  $\hat{\beta}$  (and hence the residual vectors) along the way.

Why does this algorithm work? The lasso criterion in (2) is a convex function of  $\beta$  and so has no local minima. The algorithm just described corresponds to the method of cyclical coordinate descent, which minimizes this convex objective function along each coordinate at a time. Under relatively mild conditions (which apply here), such coordinate-wise minimization schemes applied to a convex function converge to a global optimum. It is important to note that some conditions are required, because there are instances, involving nonseparable penalty functions, in which coordinate descent schemes can become “jammed” (see Chapter 5, *“Statistical learning with sparsity”* for further details).

Other popular algorithms include:

- Shooting algorithm (Fu 1998; Zhang and Lu, 2007)
- LARS solution path (Efron *et al.* 2001)
  1. the most efficient algorithm for LASSO with R package “LARS”;
  2. designed for standard linear regression

Lastly, the tuning parameter  $t$  or  $\lambda$  should be adaptively chosen to minimize the MSE or PE. Common tuning methods include the tuning error, cross validation (CV), generalized cross validation (GCV), AIC, BIC and others; see chapter 7 of Hastie, Tibshirani and Friedman (2011).

### 5.2.3 Other penalization methods

In past two decades, the revolutionary work is the development of regularization framework:

$$\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{Z}) + \lambda J(\boldsymbol{\beta})$$

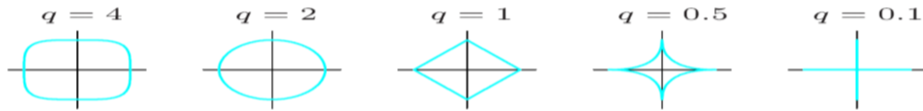
where  $L(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{Z})$  is the loss function,  $J(\boldsymbol{\beta})$  is the penalty function and  $\lambda$  is the tuning parameter. There is a vast literature on the topic, for instance,

1. for OLS,  $L = \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^d \beta_j x_{ij})^2$ ;
2. for MLE methods,  $L$  is the negative log likelihood;
3. for the Cox's proportional hazard models,  $L$  is the negative partial likelihood;
4. in supervised learning,  $L$  is the hinge loss function (SVM, support vector machine), or exponential loss (AdaBoost).

For the penalty function of the form

$$J_q(|\boldsymbol{\beta}|) = \lambda \|\boldsymbol{\beta}\|_q^q = \sum_{j=1}^p |\beta_j|^q, \quad q \geq 0,$$

1.  $J_0(|\boldsymbol{\beta}|) = \sum_{j=1}^p I(\beta_j \neq 0)$  ( $L_0$ -norm; Donoho and Johnstone, 1988);
2.  $J_1(|\boldsymbol{\beta}|) = \sum_{j=1}^p |\beta_j|$  (LASSO; Tibshirani, 1996) and its variants (Adaptive lasso, SCAD, MCP; Zou, 2005; Fan and Li, 2001; Zhang, 2010);
3.  $J_2(|\boldsymbol{\beta}|) = \sum_{j=1}^p |\beta_j|^2$  (Ridge; Hoerl and Kennard, 1970);
4.  $J_\infty(|\boldsymbol{\beta}|) = \max_j |\beta_j|$  (Supnorm penalty; Zhang et al. 2008).



**Figure 3.13:** *Contours of constant value of  $\sum_j |\beta_j|^q$  for given values of  $q$ .*

### 5.2.4 A Glimpse at the Theory (to be updated....)

To study their theoretical properties of different methods, we define the following notations:

- data  $(\mathbf{Z}_i, Y_i)$ ,  $i = 1, \dots, n$ ;
- $n$ : sample size;
- $p$ : the number of predictors,  $\mathbf{Z}_i \in \mathbb{R}^p$ ;
- The full index set  $\mathcal{S} = \{1, 2, \dots, p\}$ ;
- The selected index set given by a procedure is  $\hat{\mathcal{A}}$ , its size is  $|\hat{\mathcal{A}}|$ ;
- The linear coefficient vector  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ ;
- The true linear coefficients  $\boldsymbol{\beta}_0 = (\beta_{10}, \dots, \beta_{p0})^\top$ ;
- The true model  $\mathcal{A}_0 = \{j : j = 1, \dots, p, |\beta_{j0}| \neq 0\}$ .

As mentioned before, shrinkage method for variable selection is a continuous process. It has one unified selection framework and easy to make inferences. In theory, we could derive the oracle properties and valid asymptotic inferences. In general, we standardize the inputs before applying the methods, that is to center  $\mathbf{X}$  to mean 0 and variance 1, and center  $\mathbf{Y}$  to mean 0. We often fit a model without intercept.

*An ideal variable selection shall satisfy or the best results we can expect from a selection procedure are:*

- (1) we are able to obtain the correct model structure, i.e. keeping all the important variables in the model while filtering all the noisy variable out of the model;
- (2) it has some optimal inference properties, such as consistency, optimal convergence rate, asymptotic normality, most efficient (?).

To introduce these properties, we consider a true model

$$y_i = \beta_{00} + \sum_{j=1}^p Z_{ij}\beta_{j0} + \varepsilon.$$



Important index set (or active set):  $\mathcal{A}_0 = \{j : \beta_{j0} \neq 0, j = 1, 2, \dots, p\}$ ;

Unimportant index set (or inactive set):  $\mathcal{A}_0^c = \{j : \beta_{j0} = 0, j = 1, 2, \dots, p\}$ .

An **oracle** performs as if the true model were known;

The **selection consistency** means

$$\hat{\beta}_j \neq 0 \quad \text{for } j \in \mathcal{A}_0; \quad \hat{\beta}_j = 0 \quad \text{for } j \in \mathcal{A}_0^c;$$

The **estimation consistency and asymptotic normality** means

$$\sqrt{n}(\hat{\beta}_{\mathcal{A}_0} - \beta_{\mathcal{A}_0}) \xrightarrow{d} N(\mathbf{0}, \Sigma_I),$$

where  $\beta_{\mathcal{A}_0} = \{\beta_j, j \in \mathcal{A}_0\}$  and  $\Sigma_I$  the covariance matrix if knowing the true model.

### Consistency

More formally, let the true parameter be  $\beta_0$  and their estimates by  $\hat{\beta}_n$ . The subscript ‘ $n$ ’ means the sample size  $n$ .

Definition of **estimation consistency**:

$$\hat{\beta}_n - \beta_0 \xrightarrow{p} \mathbf{0}, \quad \text{as } n \rightarrow \infty.$$

Definition of **model selection consistency**:

$$P\left(\{j : \hat{\beta}_j \neq 0\} = \{j : \beta_{0j} \neq 0\}\right) \rightarrow 1, \quad \text{as } n \rightarrow \infty.$$

Definition of **sign consistency**:

$$P\left(\hat{\beta}_n =_s \beta_0\right) \rightarrow 1, \quad \text{as } n \rightarrow \infty.$$

where

$$\hat{\beta}_n =_s \beta_0 \Leftrightarrow \text{sgn}(\hat{\beta}_n) = \text{sgn}(\beta_0).$$

Note that sign consistency is stronger than model selection consistency.

**The following are some existing theoretical results of LASSO estimate:**

- Knight and Fu (2000 AoS, titled “Asymptotics for LASSO-type estimators” ) have shown that
  - Estimation consistency: The LASSO solution has estimation consistency for fixed  $p$ .

$$\hat{\beta}^{\text{lasso}}(\lambda_n) \xrightarrow{p} \beta, \quad \text{as } \lambda_n = o(n)$$

It is root- $n$  consistent and asymptotically normal.

- Model selection property: For  $\lambda_n \propto n^{\frac{1}{2}}$ , as  $n \rightarrow \infty$ , there is a non-vanishing positive probability for LASSO to select the true model.
  - Zhao and Yu (2006). On Model Selection Consistency of Lasso. JMLR, 7, 2541-2563.
    - If an irrelevant predictor is highly correlated with the predictors in the true model, LASSO may not be able to distinguish it from the true predictors with any amount of data and any amount of regularization.
    - Under **Irrepresentable Condition** (IC), the LASSO is model selection consistent in both fixed  $p$  and large  $p$  settings.
- The IC condition is represented as

$$| [\mathbf{X}_n^\top(1)\mathbf{X}_n(1)]^{-1} \mathbf{X}_n^\top(1)\mathbf{X}_n(2) | < 1 - \eta$$

i.e., the regression coefficients of irrelevant covariates  $\mathbf{X}_n(2)$  on the relevant covariates  $\mathbf{X}_n(1)$  is constrained. It is almost the necessary and sufficient condition for LASSO to select the true model.

- Special cases for LASSO to be selection consistent
 

Assume the true model size  $|\mathcal{A}_0| = q$ . The underlying model must satisfy a nontrivial condition if the LASSO variable selection is consistent (Zou, 2006). The LASSO is always consistent in model selection under the following special cases:

  - (1) when  $p = 2$
  - (2) when the design is orthogonal
  - (3) when the covariates have bounded constant correlation,  $0 < r_n < \frac{1}{1+cq}$
  - (4) when the design has power decay correlation with  $\text{corr}(\mathbf{x}_i, \mathbf{x}_j) = \rho_n^{|i-j|}$ , where  $|\rho_n| \leq c < 1$

- More comments

Suppose  $p \gg n$ . Of course, we prefer a parsimonious model (Occam's Razor). Ridge regression produces coefficient values for each of the  $p$  variables. But because of its  $\ell_1$  penalty, the LASSO will set many of the variables exactly equal to 0. That is, LASSO produces sparse solutions. So LASSO takes care of model selection for us, and we can even see when variables jump into the model by looking at the LASSO path.

Zou and Hastie (2005) proposed the **elastic net**, which is a convex combination of ridge and the LASSO. Paper asserts that the elastic net can improve error over LASSO and still produce sparse solutions.

Frank and Friedman (1993) introduced **bridge regression**, which generalizes  $\ell_q$  norms. Regularization ideas extended to other contexts: Park (Ph.D. Thesis, 2006) computes  $\ell_1$  regularized paths for generalized linear models and among many others.

Other developments not reviewed here are equally important to the advancement of modern linear models. Students are encouraged to read more literature to enhance their understanding on the topics.