Generalization of LASSO Penalty

We will study three variations of the LASSO penalty:

(i) Elastic Net

(ii) Group LASSO

(iii) Total variation

Elastic Net

Motivation: Consider a feature vector, where there are correlation between components.

$$\underline{X} = [X_1, X_2, X_3 \ldots, X_p]$$

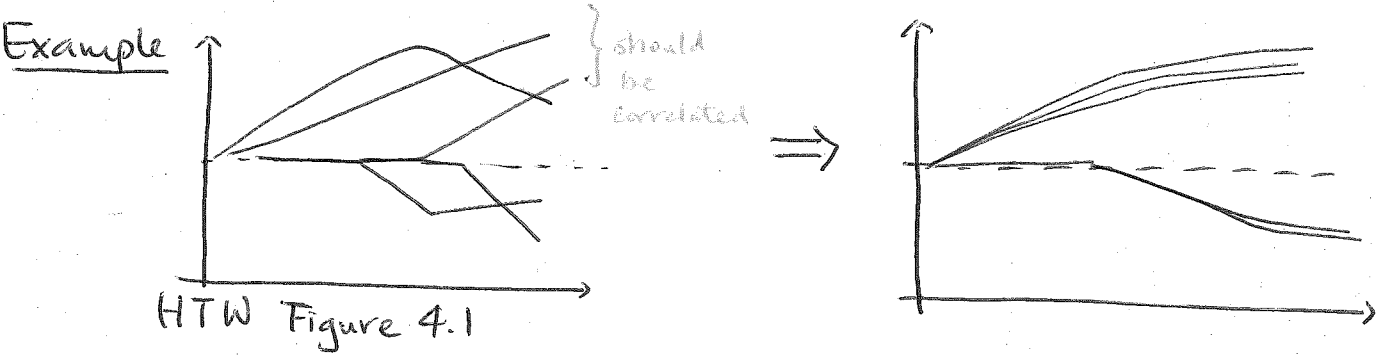
(X_1, X_2: correlated)

Then if we only do LASSO, i.e.

$$\min_{\beta_0, \beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \beta^T \underline{x}_i)^2 + \underbrace{\lambda \|\beta\|_1}_{R(\beta)} \right\}, \quad (1)$$

there will be no guarantee that the correlated features will be simultaneously activated.

Example



The Elastic Net modifies the regularization as

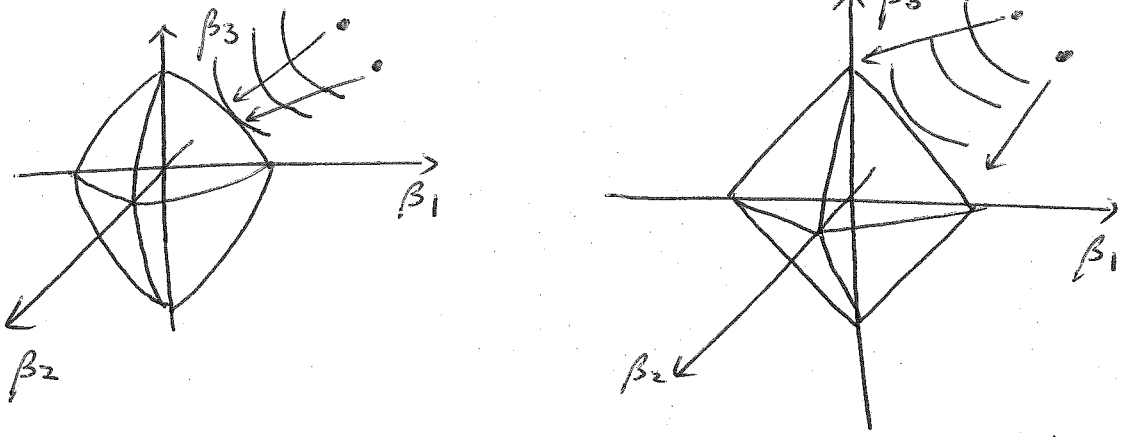
$$R(\beta) \overset{\text{def}}{=} \lambda \left[\frac{1}{2}(1-\alpha)\|\beta\|^2 + \alpha \|\beta\|_1 \right], \quad (2)$$

where α is a constant with $0 \le \alpha \le 1$.

When $\alpha = 1$, then $R(\beta) = \lambda \|\beta\|_1$, standard LASSO

When $\alpha = 0$, then $R(\beta) = \frac{\lambda}{2}\|\beta\|^2$, ridge regression

Why does Elastic Net work?



When the objective function touches the constraint, $\|\beta\|_1$ forces the solution to be sparse, even if the variable are correlated. But the curved surface of elastic net allows two correlated variables individually to land on similar solution.

Solving the Elastic Net Problem:

$$\min_{\beta_0, \beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \left[\frac{1}{2}(1-\alpha)\|\beta\|^2 + \alpha \|\beta\|_1 \right] \right\} \quad (3)$$

By coordinate descent, we have

$$\frac{\partial}{\partial \beta_j}(\cdot) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)(-x_{ij}) + \lambda \left[(1-\alpha)\beta_j + \alpha \operatorname{sgn}(\beta_j) \right]$$

$$= \frac{1}{N} \sum_{i=1}^{N} \underbrace{(y_i - \beta_0 - \sum_{k \neq j} x_{ik}\beta_k}_{= r_{ij}} - x_{ij}\beta_j)(-x_{ij}) + \lambda \left[(1-\alpha)\beta_j + \alpha \operatorname{sgn}(\beta_j) \right]$$

$$= -\frac{1}{N} \sum_{i=1}^{N} r_{ij} x_{ij} + \frac{1}{N} \sum_{i=1}^{N} x_{ij}^2 \beta_j + \lambda \left[(1-\alpha)\beta_j + \alpha \operatorname{sgn}(\beta_j) \right]$$

Setting to zero yields

$$\left[\frac{1}{N} \sum_{i=1}^{N} x_{ij}^2 + \lambda(1-\alpha) \right] \beta_j + \alpha\lambda \operatorname{sgn}(\beta_j) = \frac{1}{N} \sum_{i=1}^{N} r_{ij} x_{ij}$$

4.2 The Elastic Net

The lasso does not handle highly correlated variables very well; the coefficient paths tend to be erratic and can sometimes show wild behavior. Consider a simple but extreme example, where the coefficient for a variable X_j with a particular value for λ is $\widehat{\beta}_j > 0$. If we augment our data with an *identical* copy $X_{j'} = X_j$, then they can share this coefficient in infinitely many ways—any $\tilde{\beta}_j + \tilde{\beta}_{j'} = \hat{\beta}_j$ with both pieces positive—and the loss and ℓ_1 penalty are indifferent. So the coefficients for this pair are not defined. A quadratic penalty, on the other hand, will divide $\hat{\beta}_j$ exactly equally between these two twins (see Exercise 4.1). In practice, we are unlikely to have an identical

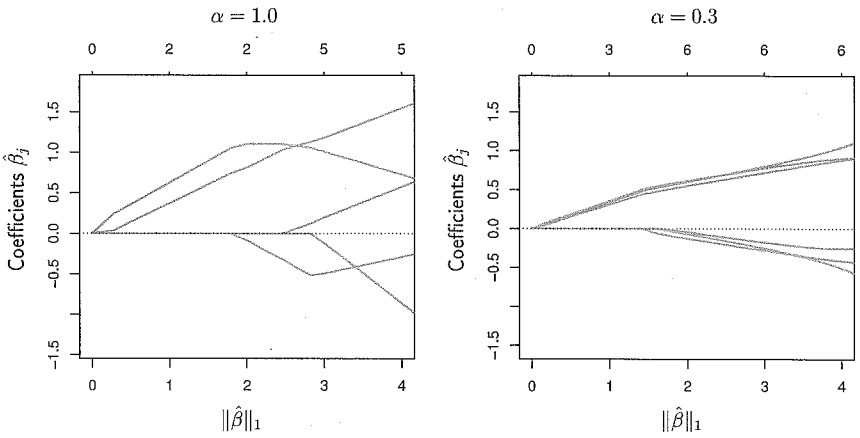


Figure 4.1 *Six variables, highly correlated in groups of three. The lasso estimates ($\alpha = 1$), as shown in the left panel, exhibit somewhat erratic behavior as the regularization parameter λ is varied. In the right panel, the elastic net with ($\alpha = 0.3$) includes all the variables, and the correlated groups are pulled together.*

pair of variables, but often we do have groups of very correlated variables. In microarray studies, groups of genes in the same biological pathway tend to be expressed (or not) together, and hence measures of their expression tend to be strongly correlated. The left panel of Figure 4.1 shows the lasso coefficient path for such a situation. There are two sets of three variables, with pairwise correlations around 0.97 in each group. With a sample size of $N = 100$, the data were simulated as follows:

$$\begin{aligned} &Z_1,\ Z_2 \sim N(0,1) \text{ independent,} \\ &\quad Y = 3 \cdot Z_1 - 1.5 Z_2 + 2\varepsilon, \text{ with } \varepsilon \sim N(0,1), \\ &\quad X_j = Z_1 + \xi_j/5, \text{ with } \xi_j \sim N(0,1) \text{ for } j = 1,2,3, \text{ and} \\ &\quad X_j = Z_2 + \xi_j/5, \text{ with } \xi_j \sim N(0,1) \text{ for } j = 4,5,6. \end{aligned} \tag{4.1}$$

As shown in the left panel of Figure 4.1, the lasso coefficients do not reflect the relative importance of the individual variables.

Why does Elastic Net promote grouping?

Theorem Assume $\hat{\beta}_i \hat{\beta}_j > 0$. Define $\rho = x_i^T x_j$.
Then

$$|\hat{\beta}_i - \hat{\beta}_j| \leq \frac{\|y\|_2}{\lambda_2} \sqrt{2(1-\rho)}.$$

Consider the overall minimization

$$\min_{\beta} \underbrace{\|y - X\beta\|^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|^2}_{L(\beta)}$$

Then the optimality at β_i and β_j is

$$\begin{cases} \frac{\partial L}{\partial \beta_i} = -2x_i^T(y - X\beta) + 2\lambda_2 \beta_i + \lambda_1 sgn(\beta_i) = 0 \\ \frac{\partial L}{\partial \beta_j} = -2x_j^T(y - X\beta) + 2\lambda_2 \beta_j + \lambda_1 sgn(\beta_j) = 0 \end{cases}$$

$$\Rightarrow 2(x_j - x_i)^T(y - X\beta) + 2\lambda_2(\beta_i - \beta_j) + \lambda_1(\cancel{sgn(\beta_i)} - sgn(\beta_j)) = 0$$

$= 0$ by assumption

$$\Rightarrow (\beta_i - \beta_j) = \frac{1}{\lambda_2}(x_i - x_j)^T(y - X\beta)$$

$$\Rightarrow (\beta_i - \beta_j)^2 \leq \frac{1}{\lambda_2^2} \underbrace{\|x_i - x_j\|^2}\, \|y - X\beta\|^2 \text{, by Cauchy}$$

$$= \|x_i\|^2 - 2x_i^T x_j + \|x_j\|^2$$
$$= 1 - 2\rho + 1$$
$$= 2(1-\rho)$$

Assume $\sum_i x_{ij}^2 = 1$

So $\|x_i\|_2 = 1$.

Now we need to bound $\|y - X\beta\|^2$.

Note that

$$\underbrace{L(\hat{\beta})}_{\|} \leq L(0) = \|y\|^2$$

$$\|y - X\hat{\beta}\|^2 + \lambda_1\|\hat{\beta}\|_1 + \lambda_2\|\hat{\beta}\|^2$$

so $\|y - X\hat{\beta}\|^2 + \lambda_1\|\hat{\beta}\|_1 + \lambda_2\|\hat{\beta}\|^2 \leq \|y\|^2$

$$\Rightarrow \|y - X\hat{\beta}\|^2 \leq \|y\|^2 - \lambda_1 \underbrace{\|\hat{\beta}\|_1}_{>0} - \lambda_2 \underbrace{\|\hat{\beta}\|^2}_{>0}$$

$$\leq \|y\|^2.$$

Therefore

$$\boxed{|\beta_i - \beta_j| \leq \frac{1}{\lambda_2}\sqrt{2(1-\rho)}\,\|y\|_2.}$$

Why not worry about $\hat{\beta}_i\hat{\beta}_j < 0$?

- if $\hat{\beta}_i\hat{\beta}_j < 0$, then they are on two different paths, so they cannot be grouped.

What if $\rho = -1$?

- Then consider x_i and $-x_j$.

Strictly Convex VS convex for grouping

Lemma Assume $\underline{x}_i = \underline{x}_j$. Let

$J(\cdot)$ = regularization function

Then

(i) if $J(\cdot)$ is strictly convex, then $\hat{\beta}_i = \hat{\beta}_j$, $\forall \lambda$

(ii) if $J(\cdot) = \|\beta\|_1$, then $\hat{\beta}_i \hat{\beta}_j \geq 0$.

Proof Assume $\hat{\beta}_i \neq \hat{\beta}_j$.

(i) Let $\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$, and assume $\underline{x}_i = \underline{x}_j$.

Define $\tilde{\beta} = \begin{bmatrix} \tilde{\beta}_1 \\ \vdots \\ \tilde{\beta}_p \end{bmatrix}$, where $\begin{cases} \tilde{\beta}_i = \frac{1}{2}(\hat{\beta}_i + \hat{\beta}_j) \\ \tilde{\beta}_j = \frac{1}{2}(\hat{\beta}_i + \hat{\beta}_j) \\ \tilde{\beta}_k = \hat{\beta}_k \quad k \neq i \text{ and } k \neq j. \end{cases}$

Then

$$X\tilde{\beta} = \hat{\beta}_1 \underline{x}_1 + \cdots + \boxed{\left(\tfrac{1}{2}(\hat{\beta}_i + \hat{\beta}_j)\right)\underline{x}_i + \left(\tfrac{1}{2}(\hat{\beta}_i + \hat{\beta}_j)\right)\underline{x}_j} + \cdots + \hat{\beta}_p \underline{x}_p.$$

Since $\underline{x}_i = \underline{x}_j$, the sum is same as $\hat{\beta}_i \underline{x}_i + \hat{\beta}_j \underline{x}_j$.

So $X\tilde{\beta} = X\hat{\beta}$.

$\Rightarrow \|y - X\tilde{\beta}\| = \|y - X\hat{\beta}\|$.

Since $J(\cdot)$ is strictly convex, thus

$J(\tilde{\beta}) < J(\hat{\beta})$. So $\tilde{\beta}$ is a better minimizer. Contradiction.

(ii) if $\hat{\beta}_i \hat{\beta}_j < 0$, then

$$\left(\hat{\beta}_i + \hat{\beta}_j\right)^2 = \hat{\beta}_i^2 + \hat{\beta}_j^2 + 2\hat{\beta}_i\hat{\beta}_j$$

$$< \hat{\beta}_i^2 + \hat{\beta}_j^2$$

$$\Rightarrow \left|\hat{\beta}_i + \hat{\beta}_j\right| < \sqrt{\hat{\beta}_i^2 + \hat{\beta}_j^2} \leq |\hat{\beta}_i| + |\hat{\beta}_j|.$$

Therefore $J(\tilde{\beta}) < J(\hat{\beta})$, where $\tilde{\beta}$ is defined in the same way as in (i).

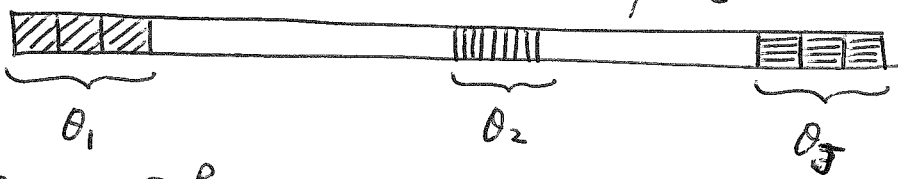
So $\tilde{\beta}$ is a better minimizer. Contradiction.

So we have

$$\hat{\beta}_j = \frac{S_{\lambda\alpha}\left(\frac{1}{N}\sum_{i=1}^{N} r_{ij} x_{ij}\right)}{\frac{1}{N}\sum_{i=1}^{N} x_{ij}^2 + \lambda(1-\alpha)} \qquad (4)$$

Group LASSO

$$\underline{\beta} = [\beta_1 \ \beta_2 \ \ldots\ldots\ldots\ldots \ \beta_p]$$



(i) $\theta_j \in \mathbb{R}^{p_j}$. p_j does not need to be identical.

(ii) This is equivalent to partitioning the vector into J groups.

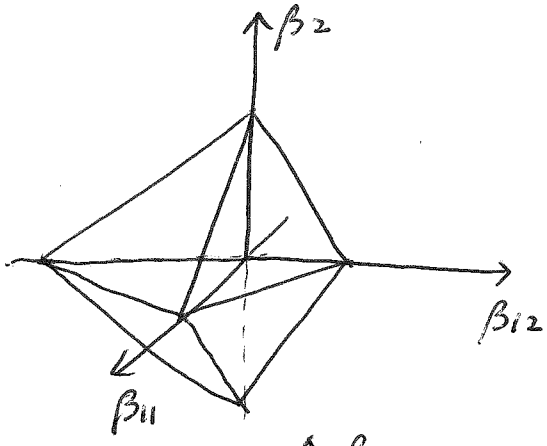
Definition of Group LASSO:

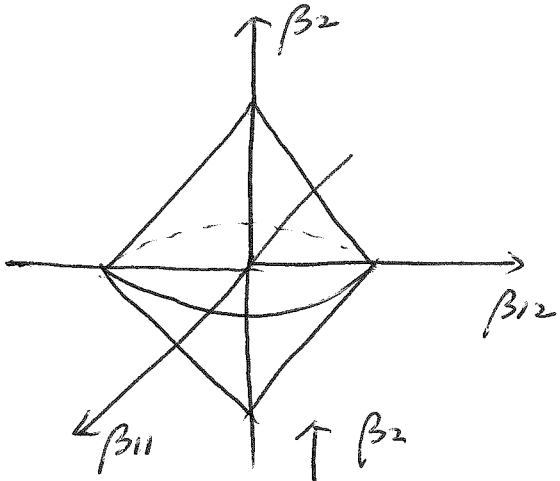
$$\min_{(\theta_0, \underline{\theta}_j)} \left\{ \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \theta_0 - \sum_{j=1}^{J} z_{ij}^T \theta_j\right)^2 + \lambda \underbrace{\sum_{j=1}^{J} \|\underline{\theta}_j\|_2}_{R(\theta)} \right\}.$$

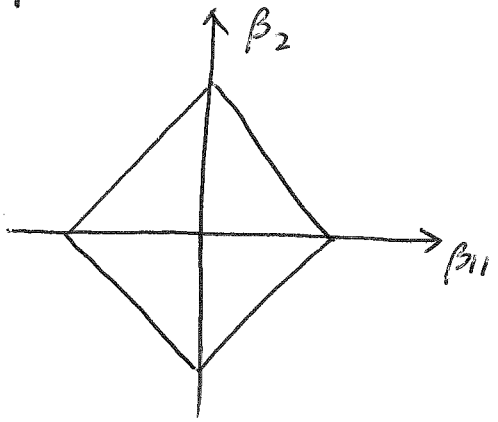
Why does $\sum_{j=1}^{J} \|\underline{\theta}_j\|_2$ promote group?

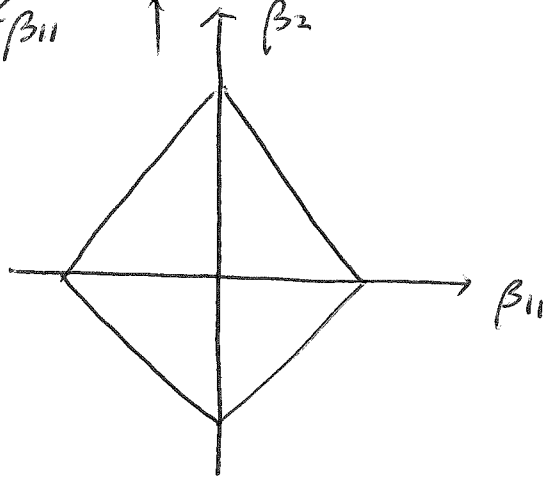
(i) if $\underline{\beta}$ is individually sparse, then very likely all $\|\underline{\theta}_j\|_2$ will have value. But if $\underline{\beta}$ is group sparse, then only a few groups $\|\underline{\theta}_j\|_2$ is activated.

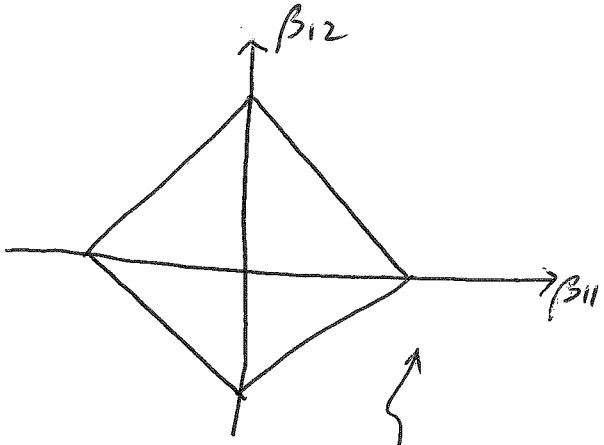
(ii) if $p_j = 1$, then $\|\underline{\theta}_j\|_1 = |\theta_j|$



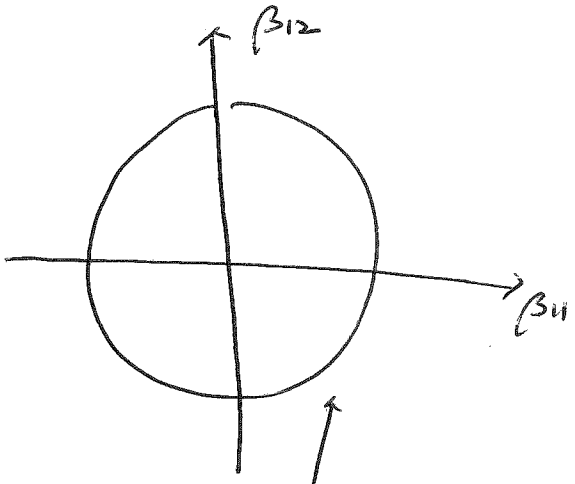








Same group,
but sparsity is
still individual



Same group.
So activated together

B1

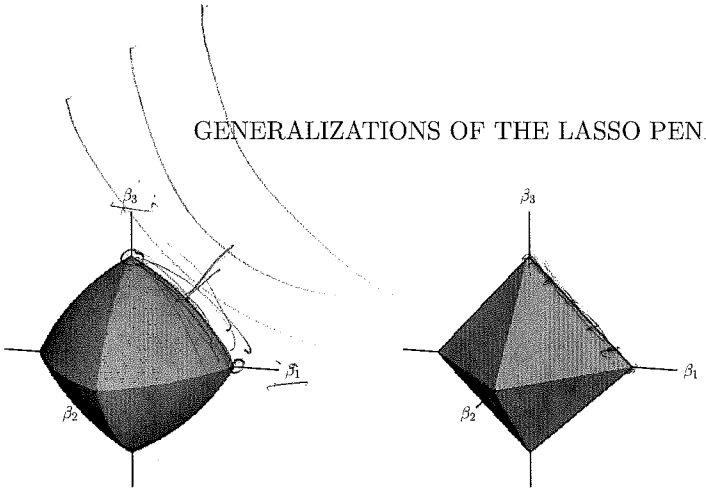


Figure 4.2 *The elastic-net ball with $\alpha = 0.7$ (left panel) in $\mathbb{R}^3$, compared to the ℓ_1 ball (right panel). The curved contours encourage strongly correlated variables to share coefficients (see Exercise 4.2 for details).*

coefficient takes the form

$$\widehat{\beta}_j = \frac{\mathcal{S}_{\lambda\alpha}\big(\sum_{i=1}^N r_{ij}x_{ij}\big)}{\sum_{i=1}^N x_{ij}^2 + \lambda(1-\alpha)}, \tag{4.4}$$

where $\mathcal{S}_\mu\big(z\big) := \text{sign}(z)\,(z-\mu)_+$ is the soft-thresholding operator, and $r_{ij} := y_i - \widehat{\beta}_0 - \sum_{k\neq j} x_{ik}\widehat{\beta}_k$ is the partial residual. We cycle over the updates (4.4) until convergence. Friedman et al. (2015) give more details, and provide an efficient implementation of the elastic net penalty for a variety of loss functions.

4.3 The Group Lasso

There are many regression problems in which the covariates have a natural group structure, and it is desirable to have all coefficients within a group become nonzero (or zero) simultaneously. The various forms of group lasso penalty are designed for such situations. A leading example is when we have qualitative factors among our predictors. We typically code their levels using a set of dummy variables or contrasts, and would want to include or exclude this group of variables together. We first define the group lasso and then develop this and other motivating examples.

Consider a linear regression model involving J groups of covariates, where for $j = 1, \ldots, J$, the vector $Z_j \in \mathbb{R}^{p_j}$ represents the covariates in group j. Our goal is to predict a real-valued response $Y \in \mathbb{R}$ based on the collection of covariates $(Z_1, \ldots, Z_J)$. A linear model for the regression function $\mathbb{E}(Y \mid Z)$

4.3.2 Sparse Group Lasso

When a group is included in a group-lasso fit, all the coefficients in that group are nonzero. This is a consequence of the ℓ_2 norm. Sometimes we would like sparsity both with respect to which groups are selected, and which coefficients are nonzero within a group. For example, although a biological pathway may be implicated in the progression of a particular type of cancer, not all genes in the pathway need be active. The *sparse group lasso* is designed to achieve such within-group sparsity.

In order to achieve within-group sparsity, we augment the basic group lasso (4.11) with an additional ℓ_1-penalty, leading to the convex program

$$\underset{\{\theta_j \in \mathbb{R}^{p_j}\}_{j=1}^J}{\text{minimize}} \left\{ \frac{1}{2}\|\mathbf{y} - \sum_{j=1}^J \mathbf{Z}_j\theta_j\|_2^2 + \lambda \sum_{j=1}^J \left[(1-\alpha)\|\theta_j\|_2 + \alpha\|\theta_j\|_1\right] \right\}, \quad (4.17)$$

with $\alpha \in [0,1]$. Much like the elastic net of Section 4.2, the parameter α creates a bridge between the group lasso ($\alpha = 0$) and the lasso ($\alpha = 1$). Figure 4.5 contrasts the group lasso constraint region with that of the sparse group lasso for the case of three variables. Note that in the two horizontal axes, the constraint region resembles that of the elastic net.

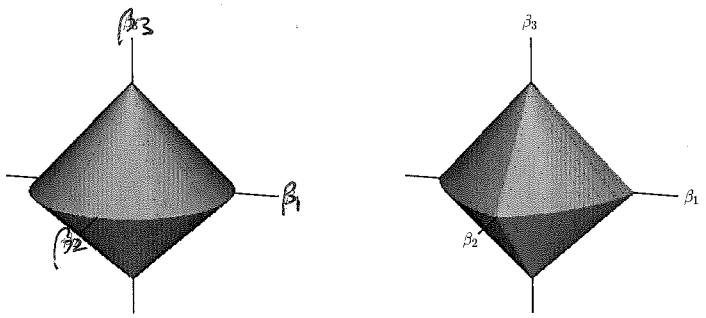


Figure 4.5 *The group lasso ball (left panel) in* $\mathbb{R}^3$*, compared to the sparse group-lasso ball with* $\alpha = 0.5$ *(right panel). Depicted are two groups with coefficients* $\theta_1 = (\beta_1, \beta_2) \in \mathbb{R}^2$ *and* $\theta_2 = \beta_3 \in \mathbb{R}^1$*.*

Since the optimization problem (4.17) is convex, its optima are specified by zero subgradient equations, similar to (4.13) for the group lasso. More precisely, any optimal solution must satisfy the condition

$$-\mathbf{Z}_j^T(\mathbf{y} - \sum_{\ell=1}^J \mathbf{Z}_\ell\widehat{\theta}_\ell) + \lambda(1-\alpha)\cdot\widehat{s}_j + \lambda\alpha\widehat{t}_j = 0, \text{ for } j = 1, \cdots, J, \quad (4.18)$$

where $\widehat{s}_j \in \mathbb{R}^{p_j}$ belongs to the subdifferential of the Euclidean norm at $\widehat{\theta}_j$,

Note: if $p_j = 1$, then $\|\theta_j\|_2 = |\theta_j|$.

Example: some observations that change over time (K time stamps)

$$\underline{Y} \in \mathbb{R}^{N \times K}, \quad \underline{X} \in \mathbb{R}^{N \times p}, \quad \underline{\Theta} \in \mathbb{R}^{p \times K}$$

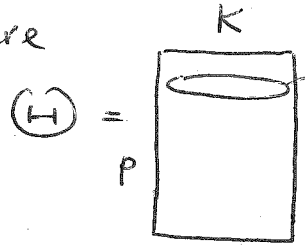
So we have

$$\underline{Y} = \underline{X}\,\underline{\Theta} + \underline{E}, \quad \underline{E} \text{ is some kind of error.}$$

The problem can be formulated as

$$\min_{\underline{\Theta} \in \mathbb{R}^{p \times K}} \left\{ \frac{1}{2} \|\underline{Y} - \underline{X}\,\underline{\Theta}\|_F^2 + \lambda \sum_{j=1}^{p} \|\underline{\theta}_j'\|_2 \right\}$$

where



$\underline{\theta}_j' \in \mathbb{R}^K$ is the j^{th} row of $\underline{\Theta}$.

Compare: if we only use $\sum_{j=1}^{p}\sum_{k=1}^{K} |\Theta_{jk}|$, i.e. $\|\Theta\|_1$, then we have no way of controlling the group sparsity.

$\sum_{j=1}^{p} \|\underline{\theta}_j'\|_2$ can do group sparsity because once the j^{th} row is activated, all elements on the row are activated.

Example:

Compressed sensing of multiple frames:

$$\begin{cases} \underline{y}_1 = S_1 \underline{\Phi}\, \underline{\theta}_1 \\ \underline{y}_2 = S_2 \underline{\Phi}\, \underline{\theta}_2 \\ \vdots \\ \underline{y}_K = S_K \underline{\Phi}\, \underline{\theta}_K \end{cases}$$

$S_1, S_2, \ldots, S_K$ are binary masks (random mixing matrix.)

$\underline{\Phi}$ = basis function, e.g. wavelet

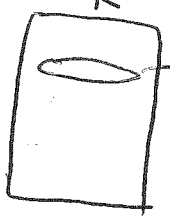
$\underline{\theta}_1, \ldots, \underline{\theta}_K$ are wavelet coefficients

$$\Rightarrow \begin{bmatrix} \underline{y}_1 \\ \vdots \\ \underline{y}_K \end{bmatrix} = \begin{bmatrix} S_1 \\ \vdots \\ S_K \end{bmatrix} \begin{bmatrix} \underline{\Phi} & & \\ & \ddots & \\ & & \underline{\Phi} \end{bmatrix} \begin{bmatrix} \underline{\theta}_1 \\ \vdots \\ \underline{\theta}_K \end{bmatrix}.$$

Since the wavelet coefficients should be similar, we can do

$$\min \left\| \begin{bmatrix} y_1 \\ \vdots \\ y_K \end{bmatrix} - \begin{bmatrix} S_1 \\ \vdots \\ S_K \end{bmatrix} \begin{bmatrix} \Phi & & \\ & \ddots & \\ & & \Phi \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_K \end{bmatrix} \right\|^2 + \lambda \sum_{j=1}^{P} \| \theta_j' \|_2,$$

where $\Theta = [\theta_1 \dots \theta_K] = P \times K$



Variation: We can also do something like this

$$\min_{\theta_j \in \mathbb{R}^{P_j}} \left\{ \frac{1}{2N} \| y - \sum_{j=1}^{J} Z_j \theta_j \|^2 + \lambda \sum_{j=1}^{J} \left[(1-\alpha) \| \theta_j \|_2 + \alpha \| \theta_j \|_1 \right] \right\}.$$

by combining group LASSO with Elastic Net.
This is called the sparse group LASSO.

Computing Group LASSO solution

we ignored θ_0.

Rewrite (5) as follows:

$$\min_{(\theta_1, \dots, \theta_J)} \left\{ \frac{1}{2N} \| y - \sum_{j=1}^{J} Z_j \theta_j \|_2^2 + \lambda \sum_{j=1}^{J} \| \theta_j \|_2 \right\}.$$

$$\frac{\partial}{\partial \theta_j}(\cdot) = 0 \Rightarrow - Z_j^T \left(y - \sum_{l=1}^{J} Z_l \hat{\theta}_l \right) + \lambda \hat{s}_j = 0,$$

where $\hat{s}_j \in \partial \| \cdot \|_2$ is the sub-differential of $\| \cdot \|_2$.

$$= \begin{cases} \dfrac{\hat{\theta}_j}{\| \theta_j \|_2}, & \hat{\theta}_j \neq 0 \\ \{ \| \hat{s}_j \|_2 \leq 1 \}, & \hat{\theta}_j = 0 \end{cases}$$

Therefore, we have

$$= - Z_j^T \left(r_j - Z_j \hat{\theta}_j \right) + \lambda \hat{s}_j = 0 \qquad r_j = y - \sum_{k \neq j} Z_k \hat{\theta}_k.$$

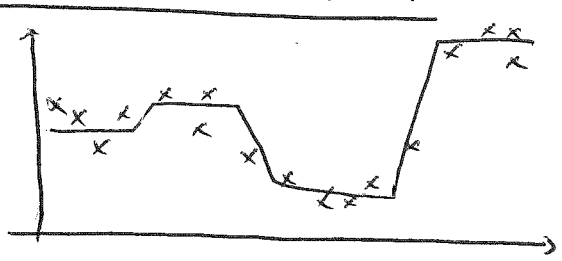
This gives

$$\hat{\theta}_j = \begin{cases} \left(\underline{Z}_j^T \underline{Z}_j + \frac{\lambda}{\|\hat{\theta}_j\|_2} I \right)^{-1} \underline{Z}_j^T \underline{r}_j, & \|\underline{Z}_j^T \underline{r}_j\|_2 \geq \lambda \\ 0, & \|\underline{Z}_j^T \underline{r}_j\|_2 < \lambda \end{cases}$$

Then we can apply coordinate-descent:

$$\hat{\theta}_j = \left(1 - \frac{\lambda}{\|\underline{Z}_j^T \underline{r}_j\|_2}\right)_+ \underline{Z}_j^T \underline{r}_j, \text{ when } \underline{Z}_j^T \underline{Z}_j = I, \text{ or when } \underline{Z}_j \text{ is orthogonal.}$$

Total Variation



Motivation: Find a curve (possibly piecewise smooth) to fit the data.

Idea: $\min_{\theta \in \mathbb{R}^N} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \theta_i)^2 + \lambda \sum_{i=2}^{N} |\theta_i - \theta_{i-1}| \right\}$.

You want the deviation of $\theta_i - \theta_{i-1}$ to be small.

This is a kind of sparsity in the gradient domain.

In general, we can do (7)

$$\min_{(\beta_0, \beta)} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j\right)^2 + \lambda \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \right\}.$$

The difficulty of (7) is that coordinate-descent does not work because $\sum |\beta_j - \beta_{j-1}|$ is not separable. The correlation between β_j and β_{j-1} makes coordinate descent impossible.

Total Variation in Imaging Applications

The typical linear shift invariant model:

$$y = Ax + \eta, \qquad x \in \mathbb{R}^n, \; y \in \mathbb{R}^m, \; A \in \mathbb{R}^{m \times n}, \; \eta \overset{iid}{\sim} N(0, \sigma^2 I)$$

Example:

(i) Denoising: $A = I$

(ii) Deblurring: A = circulant matrix

$$\text{e.g. } A = \begin{bmatrix} 2 & 1 & 0 & \cdots & 1 \\ 1 & 2 & 1 & 0 & \cdots 0 \\ 0 & 1 & 2 & 1 & \cdots 0 \\ & & \ddots & & \\ 1 & 0 & \cdots & 1 & 2 \end{bmatrix}$$

← usually denote through convolution

$$\underbrace{Ax}_{\text{matrix-vector product}} = \underbrace{h * x}_{\text{blur kernel}}$$

(iii) Inpainting: $A = S$

$$\text{e.g. } S = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

← sub-samples of I.

(iv) Super-resolution: $A = SH$

S: sampling; H: convolution

The optimization people care about:

$$\hat{x} = \underset{x}{\operatorname{argmin}} \; \|Ax - y\|^2 + \lambda R(x)$$

($R(x)$: regularization function.)

Total Variation (Rudin-Fetami-Osher 1992)

$$R(x) = \|x\|_{TV}.$$

There are several different versions of $\|\cdot\|_{TV}$.

1D: $$\|x\|_{TV} = \sum_{i=2}^{n} |x_i - x_{i-1}| + |x_1 - x_n|$$

($|x_1 - x_n|$: boundary condition.)

In term of matrix

$$\|x\|_{TV} = \|Dx\|_1 = \left\| \begin{bmatrix} 1 & -1 & & \\ & 1 & -1 & \\ & & \ddots & \ddots \\ -1 & & & 1 \end{bmatrix} x \right\|_1$$

2D:

Anisotropic TV:

$\nabla_1 x$ = horizontal gradient of x

$\nabla_2 x$ = vertical gradient of x

$[1 \ -1] * x$ $\qquad$ $\begin{bmatrix} 1 \\ -1 \end{bmatrix} * x$

$$\|x\|_{TV_1} = \|\nabla_1 x\|_1 + \|\nabla_2 x\|_1 \qquad \text{def } D$$

$$= \sum_{i=1}^{n} \left(|\nabla_1 x|_i + |\nabla_2 x|_i \right) = \left\| \begin{bmatrix} \nabla_1 \\ \nabla_2 \end{bmatrix} x \right\|_1$$

Isotropic TV:

$$\|x\|_{TV_2} = \sum_{i=1}^{n} \sqrt{(\nabla_1 x)_i^2 + (\nabla_2 x)_i^2}$$

Note that $\|x\|_{TV_2} \neq \left\| \begin{bmatrix} \nabla_1 \\ \nabla_2 \end{bmatrix} x \right\|_2$

$$= \sqrt{\sum_{i=1}^{n} (\nabla_1 x)_i^2 + (\nabla_2 x)_i^2}$$

Therefore, the final optimization becomes

$$\hat{x} = \arg\min_x \|Ax - y\|^2 + \lambda \|Dx\|_1$$

Fast algorithms to solve this problem:

ADMM (will be discussed)