Sparse linear models

Yuelin Wu

School of Mathematics
Sun Yat-sen University

October 22, 2020

Table of Contents

Table of Contents

Group lasso

In standard l_1 regularization, we assume that there is a 1:1 correspondence between parameters and variables, so that if $\hat{w}_j = 0$, we interpret this to mean that variable j is excluded. But in more complex models, there may be many parameters associated with a given variable. In particular, we may have a vector of weights for each input, w_j.

Group lasso

Here are some examples of group lasso:

- In microarray studies, we often find groups of correlated features, such as genes that operate in the same biological pathway.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.
- When we have qualitative factors among our predictors, we typically code their levels using a set of dummy variables or contrasts.

Example

The lasso does not handle highly correlated variables very well. The coefficient paths tend to be erratic and can sometimes show wild behavior.

- the coefficient for a variable X_j with a particular value for λ is $\widehat{\beta}_j > 0$.
- we augment our data with an identical copy $X_{j'} = X_j$,
- 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 L1 penalty are indifferent.
- A quadratic penalty, on the other hand, will divide β_j exactly equally between these two twins.

Group lasso

Yuan extended the lasso method to the group in 2006, and the group lasso was born. We can group all variables, and then punish the $L2$ norm of each group in the objective function. The effect achieved is to eliminate a whole group of coefficients to zero at the same time, that is, to erase a whole group of variables. This technique It is called the Group Lasso grouping minimum angle regression algorithm.

Group lasso

- If we use an $L1$ regularizer of the form $\|\mathbf{w}\| = \sum_j \sum_c |w_{jc}|$, we may end up with with some elements of $\mathbf{w}_{j,:}$ being zero and some not.
- To prevent this kind of situation, we partition the parameter vector into G groups.

Group lasso

Definition:

$$\underset{\mathbf{w}_0 \in \mathbb{R}, \mathbf{w}_j \in \mathbb{R}^{p_j}}{\text{minimize}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \mathbf{w}_0 - \sum_{j=1}^{J} z_{ij}^T \mathbf{w}_j \right)^2 + \lambda_j \sum_{j=1}^{J} \|\mathbf{w}_j\|_2 \right\}$$

where $\|\mathbf{w}_j\|_2$ is the Euclidean norm of the vector $\mathbf{w}_j$

Group lasso

We often use a larger penalty for larger groups, by setting $\lambda_j = \lambda\sqrt{d_j}$ where d_j is the number of elements in group j.
For example, if we have groups 1, 2 and 3, 4, 5, the objective becomes:

$$J(\mathbf{w}) = \lambda\left[\sqrt{2}\sqrt{\left(w_1^2 + w_2^2\ |\right)} + \sqrt{3}\sqrt{\left(w_3^2 + w_4^2 + w_5^2\right)}\right]$$

Group lasso

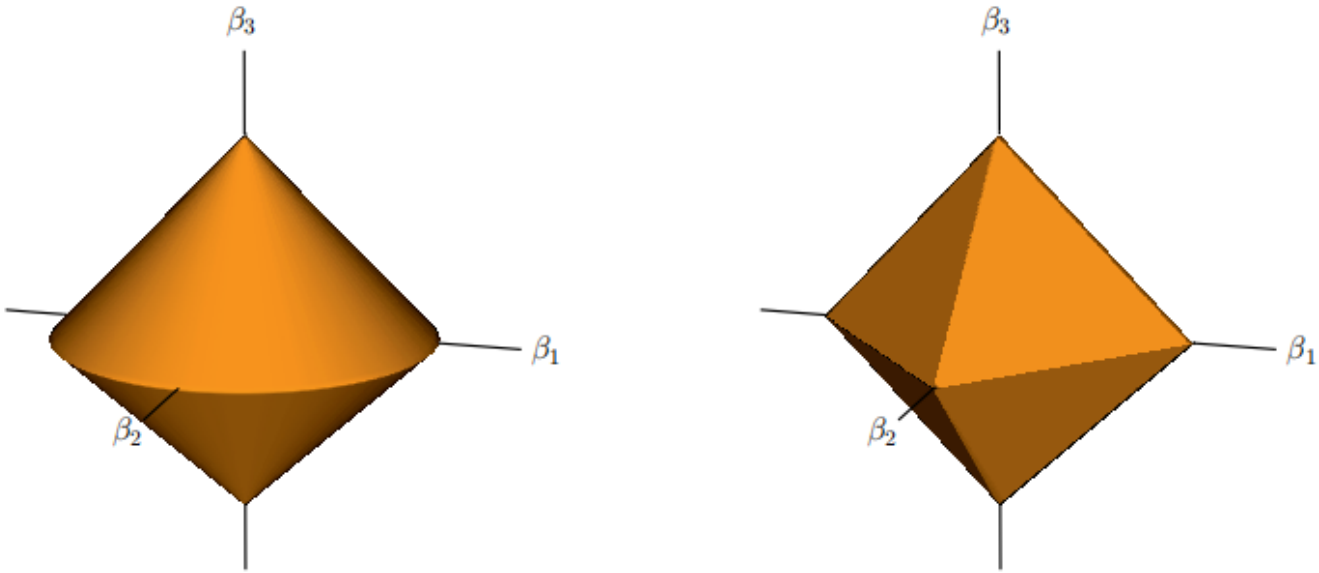


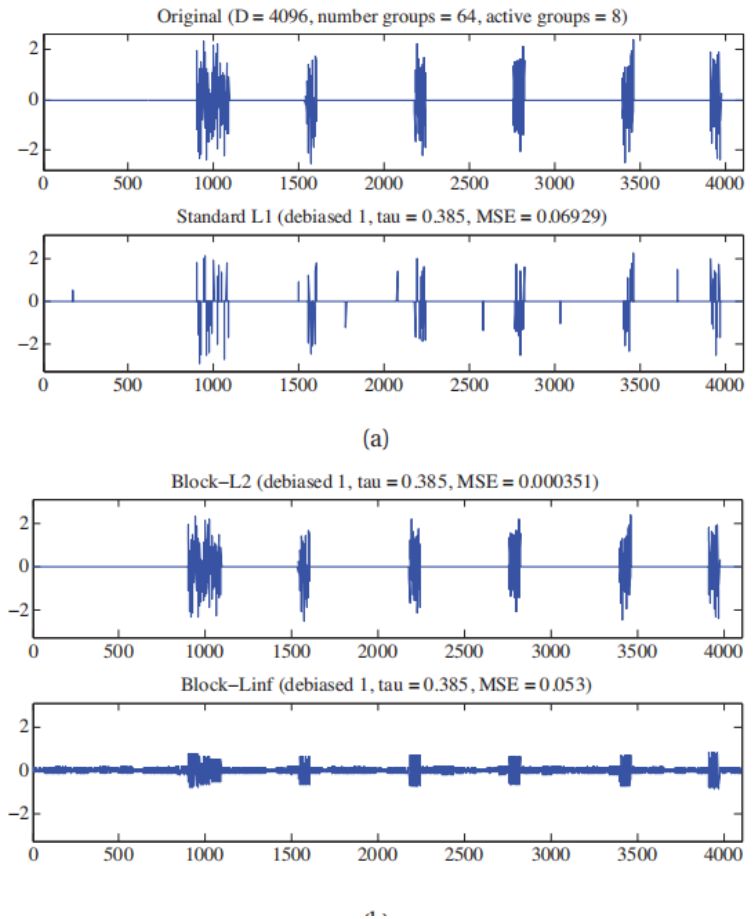
Figure 4.3 *The group lasso ball (left panel) in $\mathbb{R}^3$, compared to the ℓ_1 ball (right panel). In this case, there are two groups with coefficients $\theta_1 = (\beta_1, \beta_2) \in \mathbb{R}^2$ and $\theta_2 = \beta_3 \in \mathbb{R}^1$.*

Group lasso Example

- We have a true signal $\mathbf{w}$ of size $D = 2^{12} = 4096$, divided into 64 groups each of size 64.We randomly choose 8 groups of w and assign them non-zero values.
- In the first example, the values are drawn from a $\mathcal{N}(0, 1)$. In the second example, the values are all set to 1.
- We then pick a random design matrix $\mathbf{X}$ of size $N \times D$, where $N = 2^{10} = 1024$.
- Finally, we generate $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$,where $\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, 10^{-4}\mathbf{I}_N\right)$.

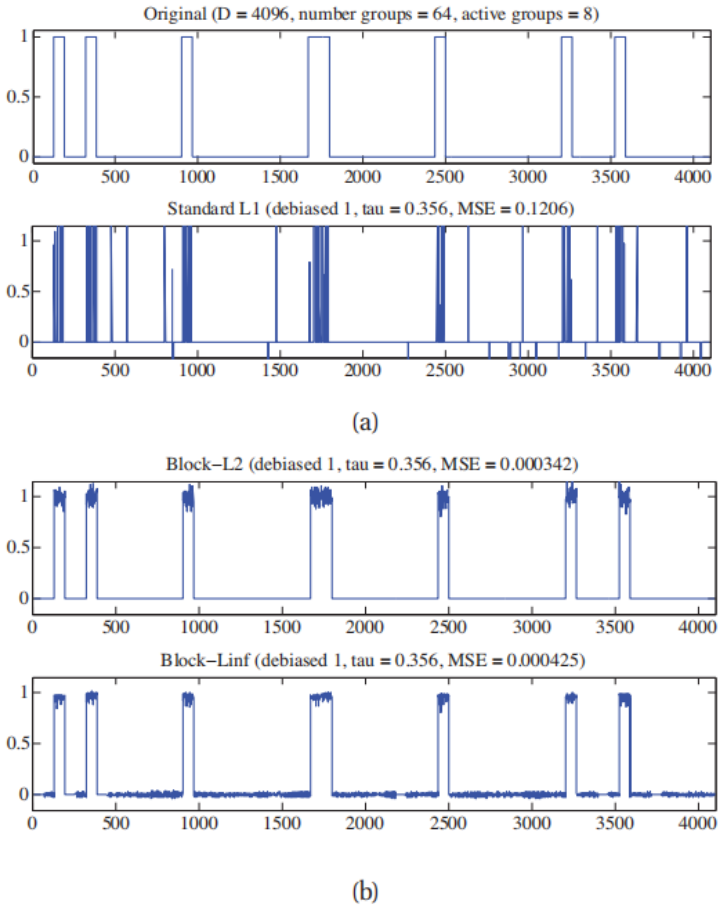
Group lasso Example

The values are drawn from a $\mathcal{N}(0, 1)$:



Group lasso Example

The values are all set to 1:



Group lasso

Group lasso is equivalent to MAP estimation using the following prior $p\left(\mathbf{w} \mid \gamma, \sigma^2\right) \propto \exp\left(-\frac{\gamma}{\sigma} \sum_{g=1}^{G} \|\mathbf{w}_g\|_2\right)$ Now one can show (Exercise 13.10) that this prior can be written as a GSM, as follows

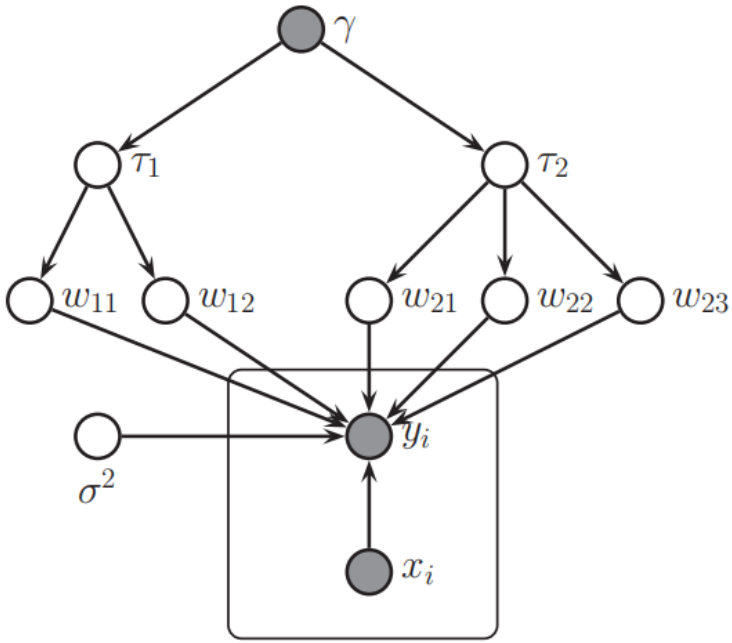
$$\mathbf{w}_g \mid \sigma^2, \tau_g^2 \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \tau_g^2 \mathbf{I}_{d_g}\right)$$

$$\tau_g^2 \mid \gamma \sim \mathrm{Ga}\left(\frac{d_g + 1}{2}, \frac{\gamma}{2}\right)$$

where d_g is the size of group g

Group lasso

So we see that there is one variance term per group, each of which comes from a Gamma prior, whose shape parameter depends on the group size, and whose rate parameter is controlled by γ



Group lasso algorithm

We rewrite the relevant optimization problem in a more compact matrix-vector notation:

$$\delta \underset{(\theta_1,\ldots,\theta_J)}{\text{minimize}} \left\{ \frac{1}{2} \left\| \mathbf{y} - \sum_{j=1}^{J} \mathbf{Z}_j \theta_j \right\|_2^2 + \lambda \sum_{j=1}^{J} \|\theta_j\|_2 \right\}$$

For this problem, the zero subgradient equations

$$-\mathbf{Z}_j^T \left(\mathbf{y} - \sum_{\ell=1}^{J} \mathbf{Z}_\ell \widehat{\theta}_\ell \right) + \lambda \widehat{s}_j = 0, \qquad \text{for } j = 1, \cdots J$$

- whenever $\widehat{\theta}_j \neq 0$, then $\widehat{s}_j = \widehat{\theta}_j / \left\| \widehat{\theta}_j \right\|_2$.
- when $\widehat{\theta}_j = 0$, then $\widehat{s}_j$ is any vector with $\|\widehat{s}_j\|_2 \leq 1$.

Block coordinate descent

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

One method for solving the zero subgradient equations is by holding fixed all block vectors $\left\{ \widehat{\theta}_k, k \neq j \right\}$, and then solving for $\widehat{\theta}_j$. Doing so amounts to performing block coordinate descent on the group lasso objective function.

We rewrite

$$-\mathbf{Z}_j^T \left(\mathbf{r}_j - \mathbf{Z}_j \hat{\theta}_j \right) + \lambda \widehat{s}_j = 0$$

where $\mathbf{r}_j = \mathbf{y} - \sum_{k \neq j} \mathbf{Z}_k \widehat{\theta}_k$ is the j^{th} partial residual.

- by the subgradient $\widehat{s}_j$, we must have $\hat{\theta}_j = 0$ if $\left\| \mathbf{Z}_j^T \mathbf{r}_j \right\|_2 < \lambda$.
- otherwise the minimizer $\widehat{\theta}_j$ must satisfy

$$\widehat{\theta}_j = \left(\mathbf{Z}_j^T \mathbf{Z}_j + \frac{\lambda}{\|\widehat{\theta}_j\|_2} \mathbf{I} \right)^{-1} \mathbf{Z}_j^T \mathbf{r}_j$$

Group lasso algorithm

$$\widehat{\theta}_j = \left(\mathbf{Z}_j^T \mathbf{Z}_j + \frac{\lambda}{\|\widehat{\theta}_j\|_2} \mathbf{I} \right)^{-1} \mathbf{Z}_j^T \mathbf{r}_j$$

This update is similar to the solution of a ridge regression problem, except that the underlying penalty parameter depends on $\left\|\widehat{\theta}_j\right\|_2$.

Unfortunately, Equation does not have a closed-form solution for $\widehat{\theta}_j$ unless $\mathbf{Z}_j$ is orthonormal.

In this special case, we have the simple update

$$\widehat{\theta}_j = \left(1 - \frac{\lambda}{\|\mathbf{Z}_j^T \mathbf{r}_j\|_2} \right)_+ \mathbf{Z}_j^T \mathbf{r}_j$$

where $(t)_+ := \max\{0, t\}$ is the positive part function.

Table of Contents

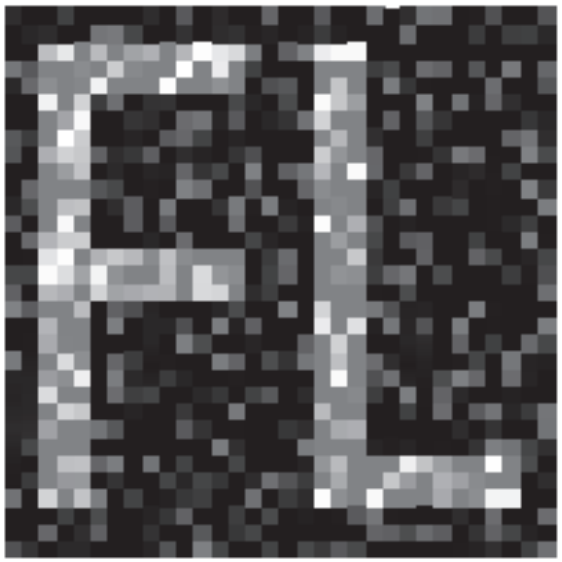
Fused lasso

In some problem settings (e.g., functional data analysis),

- We want neighboring coefficients to be similar to each other
- We want coefficients to being sparse.

Fused lasso Example

Left:Noisy image. Right: Fused lasso estimate using 2d lattice prior.



Fused lasso

We can model this by using a prior of the form

$$p\left(\mathbf{w} \mid \sigma^2\right) \propto \exp\left(-\frac{\lambda_1}{\sigma}\sum_{j=1}^{D}|w_j| - \frac{\lambda_2}{\sigma}\sum_{j=1}^{D-1}|w_{j+1} - w_j|\right)$$

we often use $X = I$

$$J\left(\mathbf{w}, \lambda_1, \lambda_2\right) = \sum_{i=1}^{N}\left(y_i - w_i\right)^2 + \lambda_1\sum_{i=1}^{N}|w_i| + \lambda_2\sum_{i=1}^{N-1}|w_{i+1} - w_i|$$

GSM interpretation of fused lasso

The fused lasso model is equivalent to the following hierarchical model

$$\mathbf{w} \mid \sigma^2, \boldsymbol{\tau}, \boldsymbol{\omega} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \boldsymbol{\Sigma}(\boldsymbol{\tau}, \boldsymbol{\omega})\right)$$

$$\tau_j^2 \mid \gamma_1 \sim \text{Expon}\left(\frac{\gamma_1^2}{2}\right), \quad j = 1 : D$$

$$\omega_j^2 \mid \gamma_2 \sim \text{Expon}\left(\frac{\gamma_2^2}{2}\right), \quad j = 1 : D - 1$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}$, and $\boldsymbol{\Omega}$ is a tridiagonal precision matrix with

$$\text{main diagonal} = \left\{\frac{1}{\tau_j^2} + \frac{1}{\omega_{j-1}^2} + \frac{1}{\omega_j^2}\right\}$$

$$\text{off diagonal} = \left\{-\frac{1}{\omega_j^2}\right\}$$

where we have defined $\omega_0^{-2} = \omega_D^{-2} = 0$

Fused lasso

It is possible to generalize the EM algorithm to fit the fused lasso model, by exploiting the Markov structure of the Gaussian prior for efficiency. Direct solvers can also be derived. However, this model is undeniably more expensive to fit than the other variants we have considered.

Table of Contents

Elastic net

Although lasso has proved to be effective as a variable selection technique, it has several problems

- If there is a group of variables that are highly correlated , then the lasso tends to select only one of them, chosen rather arbitrarily.
- In $D > N$, lasso can select at most N variables before it saturates.
- If $N > D$, but the variables are correlated, it has been empirically observed that the prediction performance of ridge is better than that of lasso.

Example

The lasso does not handle highly correlated variables very well. The coefficient paths tend to be erratic and can sometimes show wild behavior.

- the coefficient for a variable X_j with a particular value for λ is $\widehat{\beta}_j > 0$.
- we augment our data with an identical copy $X_{j'} = X_j$,
- 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 L1 penalty are indifferent.
- A quadratic penalty, on the other hand, will divide β_j exactly equally between these two twins.

Elastic net

The elastic net makes a compromise between the ridge and the lasso penalties; it solves the convex program:

$$\text{minimize}_{(\beta_0,\beta)} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - x_i^T \beta\right)^2 + \lambda \left[\frac{1}{2}(1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1 \right] \right\}$$

where $\alpha \in [0, 1]$ is a parameter that can be varied. By construction, the penalty applied to an individual coefficient (disregarding the regularization weight $\lambda > 0$) is given by

$$\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha\,|\beta_j|$$

Example

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:

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

Elastic net

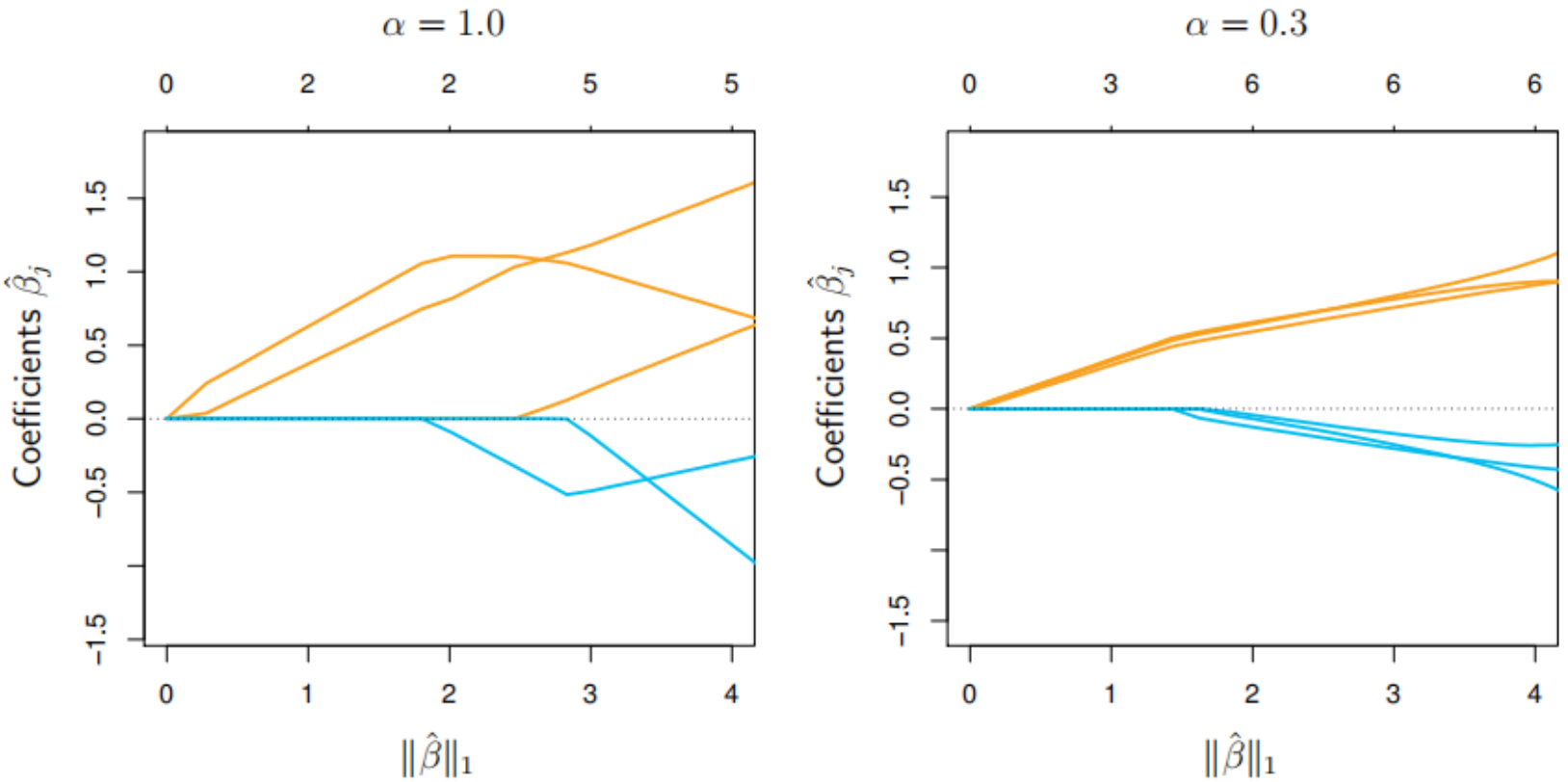


Figure 4.1 *Six variables, highly correlated in groups of three. The lasso estimates ($\alpha = 1$), as shown in the left panel, exhibit somewhat erratic behavior as the regularization parameter λ is varied. In the right panel, the elastic net with ($\alpha = 0.3$) includes all the variables, and the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

Example

- By adding some component of the ridge penalty to the L1-penalty, the elastic net automatically controls for strong within-group correlations.
- for any $\alpha < 1$ and $\lambda > 0$, the elastic-net problem is strictly convex: a unique solution exists irrespective of the correlations or duplications in the X_j

The elastic net has an additional tuning parameter α that has to be determined. In practice, it can be viewed as a higher-level parameter.

- Set on subjective grounds.
- Set in a cross-validation scheme.

Elastic net

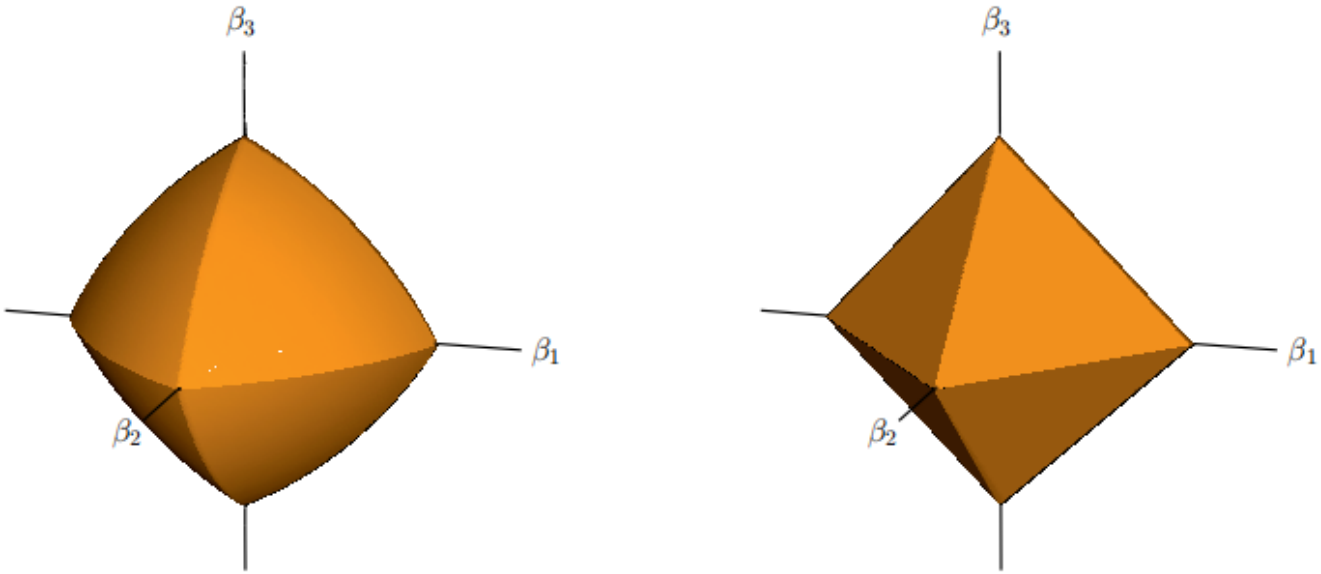


Figure 4.2 *The elastic-net ball with $\alpha = 0.7$ (left panel) in $\mathbb{R}^3$, compared to the ℓ_1 ball (right panel). The curved contours encourage strongly correlated variables to share coefficients (see Exercise 4.2 for details).*

Algorithm

- The elastic-net problem (4.2) is convex in the pair $(\beta_0, \beta) \in \mathbb{R} \times \mathbb{R}^p$, and a variety of different algorithms can be used to solve it.
- Coordinate descent is particularly effective, and the updates are a simple extension of those for the lasso.

Coordinate Descent:

- Simply center the covariates x_{ij}
- $\widehat{\beta}_0 = \bar{y} = \frac{1}{N}\sum_{j=1}^{N} y_j$
- the coordinate descent update for the j^{th} coefficient takes the form:

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

where $\mathcal{S}_\mu(z) := \text{sign}(z)(z-\mu)_+$ 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.

GSM interpretation of elastic net

The implicit prior being used by the elastic net is a product of Gaussian and Laplace distributions

$$p\left(\mathbf{w} \mid \sigma^2\right) \propto \exp\left(-\frac{\gamma_1}{\sigma}\sum_{j=1}^{D}|w_j| - \frac{\gamma_2}{2\sigma^2}\sum_{j=1}^{D} w_j^2\right)$$

This can be written as a hierarchical prior as follows

$$w_j \mid \sigma^2, \tau_j^2 \sim \mathcal{N}\left(0, \sigma^2\left(\tau_j^{-2} + \gamma_2\right)^{-1}\right)$$

$$\tau_j^2 \mid \gamma_1 \sim \text{Expon}\left(\frac{\gamma_1^2}{2}\right)$$

Clearly if $\gamma_2 = 0$, this reduces to the regular lasso.

Table of Contents

Nonconvex Penalties

When p is large and the number of relevant variables is small, lasso may not be enough. In order to reduce the set of chosen variables sufficiently, lasso may end up over-shrinking the retained variables. For this reason there has been interest in nonconvex penalties.

Nonconvex Penalties

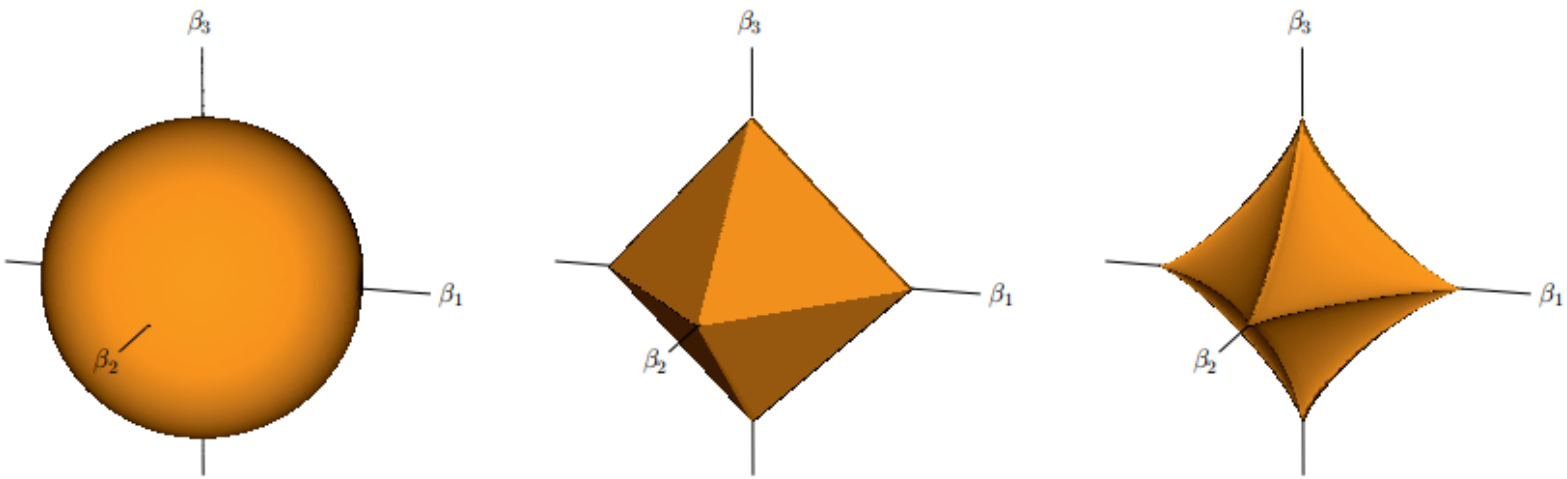


Figure 4.12 *The ℓ_q unit balls in $\mathbb{R}^3$ for $q = 2$ (left), $q = 1$ (middle), and $q = 0.8$ (right). For $q < 1$ the constraint regions are nonconvex. Smaller q will correspond to fewer nonzero coefficients, and less shrinkage. The nonconvexity leads to combinatorially hard optimization problems.*

Nonconvex Penalties

Unfortunately, along with nonconvexity comes combinatorial computational complexity; even the simplest case of L0 can be solved exactly only for $p \approx 40$ or less.

For this and related statistical reasons alternative nonconvex penalties have been proposed.

Adaptive lasso

Zou (2006) proposed the adaptive lasso as a means for fitting models sparser than lasso. Using a pilot estimate $\tilde{\beta}$, the adaptive lasso solves

$$\text{minimize}_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right\}$$

where $w_j = 1/\left|\tilde{\beta}_j\right|^{\nu}$

Adaptive lasso

If $p < N$ one can use the least-squares solutions as the pilot estimates.

If $p > N$, the least-squares estimates are not defined, but the univariate regression coefficients can be used for the pilot estimates and result in good recovery properties under certain conditions.

Adaptive lasso

The adaptive lasso penalty can be seen as an approximation to the ℓ_q penalties with $q = 1 - \nu$.

One advantage of the adaptive lasso is that given the pilot estimates, the criterion is convex in β.

Thanks !