Regularized Regression for High-Dimensional Data

Main Reference: Sparsity, the Lasso, and Friends, Model selection and estimation in regression with grouped variables

Key Concepts:

Lasso Ridge Group Lasso Sparsity Regression Model High-Dimensional Data

Table of Contents:

The Failure of Least Squares in High Dimensions

Consider n *i.i.d.* samples $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$ from the linear model $y_i = x_i^\top \beta + \varepsilon_i,\ i = 1, \ldots, n$, with $\mathbb{E}[\varepsilon_i] = 0$. In vector form,

$$y = X\beta + \varepsilon,$$

where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, and $\varepsilon \in \mathbb{R}^n$. The least squares estimator solves

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2.$$

The solution is

$$\hat{\beta} = (X^\top X)^+ X^\top y,$$

where $(X^\top X)^+$ denotes the Moore–Penrose pseudoinverse (which reduces to $(X^\top X)^{-1}$ when the matrix is invertible).

The fitted response vector is

$$\hat{y} = X\hat{\beta} = X(X^\top X)^+ X^\top y,$$

where $P_X := X(X^\top X)^+ X^\top$ is called the projection matrix onto the column space of X.

📌 **High-dimensional Regime** ($p \gg n$)

In this case, $\mathrm{rank}(X) < p$:

- Nonuniqueness If $\hat{\beta}$ is a solution, then $\hat{\beta} + \eta,\ \eta \in \mathrm{null}(X)$, is also a solution. Hence coefficients cannot be interpreted meaningfully.
- High variance The in-sample prediction risk of least squares satisfies

$$\mathrm{Risk} = \mathbb{E}\left[\frac{1}{n}\|X(\hat{\beta} - \beta)\|_2^2\right] \approx \sigma^2 \frac{p}{n},$$

  where σ^2 is the noise variance. Hence the risk increases linearly with p, becoming large when p is not small relative to n.

> Least squares is unstable in high dimensions—regularization is required.

Regularization: Ridge, Lasso, Group Lasso

To handle the issue before, we consider the general penalized least-squares formulation

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + P(\beta),$$

where different choices of $P(\beta)$ lead to different regularization behaviors. Below we describe several important penalties and the structure they induce.

◆ Best-Subset Selection (L0 penalty) The most direct way to enforce sparsity is through the L0 penalty:

$$P(\beta) = \lambda\|\beta\|_0, \quad \|\beta\|_0 = \sum_{j=1}^{p} \mathbf{1}\{\beta_j \neq 0\}.$$

This yields the optimization problem

$$\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_0.$$

It selects a subset of variables explicitly, but the problem is combinatorial and NP-hard, motivating convex relaxations.

◆ Lasso (L1 penalty) Replacing the nonconvex L0 penalty with the convex L1 norm gives the Lasso:

$$P(\beta) = \lambda\|\beta\|_1, \quad \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|.$$

The corresponding estimator solves

$$\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

The L1 penalty promotes coordinate-wise sparsity, setting many coefficients exactly to zero, and can be viewed as a computationally tractable relaxation of best-subset selection.

◆ Ridge (L2 penalty) If the goal is stabilization rather than sparsity, the L2 penalty is used:

$$P(\beta) = \lambda\|\beta\|_2^2, \quad \|\beta\|_2 = \left(\sum_{j=1}^{p} \beta_j^2\right)^{1/2}.$$

Ridge regression solves

$$\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2.$$

This shrinks coefficients continuously and improves conditioning but does not produce exact zeros.

◆ Group Lasso Suppose coefficients are grouped $\beta = (\beta_1, \ldots, \beta_G),\ \beta_g \in \mathbb{R}^{p_g}$. Group Lasso solves:

$$\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda \sum_{g=1}^{G} \sqrt{p_g}\, \|\beta_g\|_2.$$

Because the penalty combines L1 across groups and L2 within groups, it induces group-wise sparsity : entire blocks β_g may be set to zero.

✨ Special cases highlight its relationship to earlier methods:

- $G = 1$: reduces to Ridge (no sparsity).
- $G = p$: each variable is its own group → Lasso.

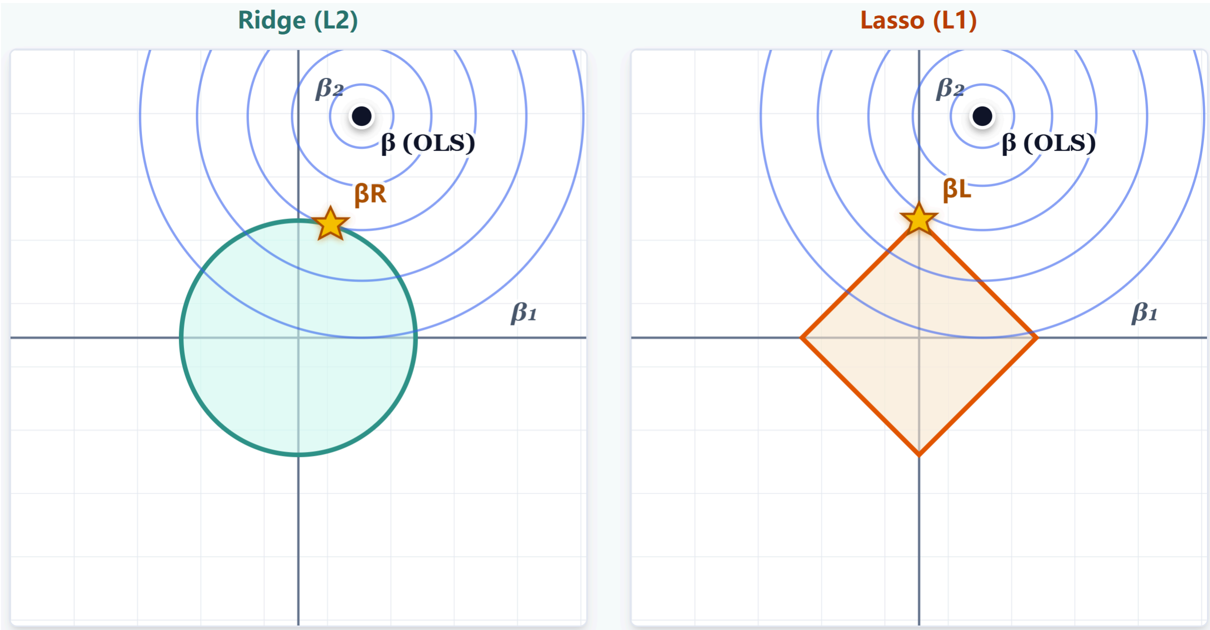
Geometric Intuition for Sparsity

Regularizers differ in geometry, which determines whether sparsity occurs.

🔶 Lasso (L1 penalty) The L1 ball has sharp corners aligned with the coordinate axes. Because these corners lie exactly on the axes, the loss contours frequently touch them, which makes $\beta_j = 0$ a common optimal solution. Hence Lasso naturally induces coordinate-wise sparsity .

🔶 Ridge (L2 penalty) The L2 ball is smooth and round, with no corners. Quadratic loss contours almost never touch the constraint boundary on an axis; instead, they intersect in smooth interior points. Thus Ridge shrinks coefficients but does not set them exactly to zero.

See the figure below for intuition.



🔶 Group Lasso Group Lasso In Group Lasso, the feasible region is a Cartesian product of L2 balls, one for each group. The boundary of each ball is smooth internally, but nondifferentiable points appear where the entire group norm $||\beta_g||_2$ reaches zero. These are "group corners." As a result:

- sparsity occurs at the group level,
- but not within a group.

Thus Group Lasso induces group sparsity , weaker than Lasso's elementwise sparsity but still stronger than Ridge's pure shrinkage.

See the figure below for intuition. In this example, the parameter vector is $(\beta_{11}, \beta_{12}, \beta_2)$, where β_{11} and β_{12} belong to group 1, and β_2 forms its own group.

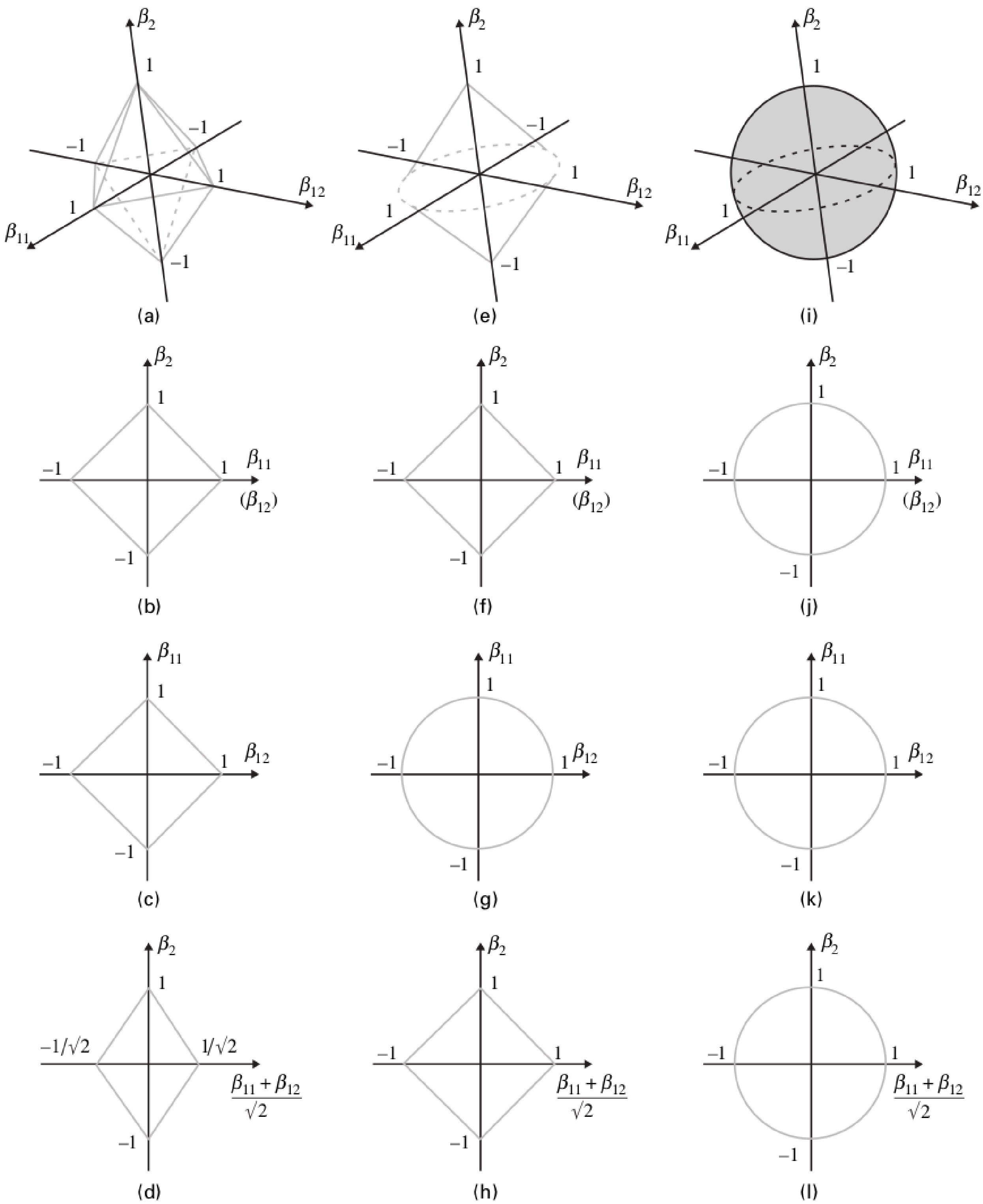


Fig. 1. (a)–(d) l_1-penalty, (e)–(h) group lasso penalty and (i)–(l) l_2-penalty

Coordinate Descent for Regularized Regression

Regularized regression problems often take the form

$$\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + P(\beta),$$

where the penalty $P(\beta)$ is separable across coordinates or groups. This separability is precisely what makes Coordinate Descent (CD) extremely effective.

CD solves the optimization problem by **cyclically updating one coordinate (or one group) at a time while holding all others fixed**. When a single coordinate is isolated, the optimization problem often admits a **closed-form update rule**, enabling CD to scale efficiently in high-dimensional settings.

Fix all coordinates except β_j. Define the partial residual

$$r_j = y - \sum_{k \neq j} X_k \beta_k.$$

The full optimization problem reduces to a 1-dimensional subproblem:

$$\min_{\beta_j} \frac{1}{2} \|r_j - X_j \beta_j\|_2^2 + P_j(\beta_j).$$

This is the key idea:

> Regularizers like L1、L2、Group-Lasso are coordinate-separable (or block-separable), and their 1-D subproblems have analytic solutions.

🔶 Ridge Regression – Closed Form Update

For Ridge, the subproblem is

$$\frac{1}{2} \|r_j - X_j \beta_j\|_2^2 + \lambda \beta_j^2.$$

This is a simple quadratic, yielding:

$$\beta_j \leftarrow \frac{X_j^\top r_j}{\|X_j\|_2^2 + 2\lambda}.$$

Insight:

- Ridge shrinks all coefficients smoothly but never to zero.
- No thresholding appears → no sparsity.

🔶 Lasso – Soft-Thresholding Update

For Lasso, the subproblem becomes

$$\frac{1}{2} \|r_j - X_j \beta_j\|_2^2 + \lambda |\beta_j|.$$

The solution has the famous soft-thresholding form:

$$\beta_j \leftarrow S\left(\frac{X_j^\top r_j}{\|X_j\|_2^2}, \frac{\lambda}{\|X_j\|_2^2} \right),$$

where $S(z, \gamma) = \mathrm{sign}(z) \max(|z| - \gamma, 0)$.

Insight:

- Soft-thresholding is exactly where Lasso sparsity comes from.
- If the correlation $X_j^\top r_j$ is below a threshold, the coordinate collapses to 0.
- This echoes the geometric interpretation: L1's corners produce zeros.

🔶 Group Lasso – Block Coordinate Descent

Now updates happen at the **group level**, not per coordinate. Let group g have parameter block β_g and design block X_g. The subproblem is

$$\min_{\beta_g} \frac{1}{2}\|r_g - X_g\beta_g\|_2^2 + \lambda\sqrt{p_g}\|\beta_g\|_2.$$

The update is the block shrinkage:

$$\beta_g \leftarrow \left(1 - \frac{\lambda\sqrt{p_g}}{\|X_g^\top r_g\|_2}\right)_+ (X_g^\top r_g).$$

Insight: The Mechanism Behind Block Sparsity

- If $\|X_g^\top r_g\|_2 < \lambda\sqrt{p_g}$, then the entire group is shrunk to zero.
- If the group is retained, the update scales the **whole vector** β_g proportionally.
- It never sets individual coordinates inside a group to zero independently.

This behavior matches the geometric intuition of Group Lasso: L1 across groups + L2 within groups → each group enters or leaves the model as a whole.

📊 As a summary,

Method	Penalty	Sparsity Type	Coordinate Descent Update	Key Insight
Ridge	$\lambda\|\beta\|_2^2$	❌ None	$\beta_j \leftarrow \frac{X_j^\top r_j}{\|X_j\|_2^2+2\lambda}$	Smooth L2 ball → shrinkage without zeros
Lasso	$\lambda\|\beta\|_1$	✅ Element-wise sparsity	$\beta_j \leftarrow S\left(\frac{X_j^\top r_j}{\|X_j\|_2^2}, \frac{\lambda}{\|X_j\|_2^2}\right)$	L1 corners → soft-thresholding produces zeros
Group Lasso	$\lambda\sum_{g=1}^{G}\sqrt{p_g}\,\|\beta_g\|_2$	✅ Group-wise sparsity	$\beta_g \leftarrow \left(1 - \frac{\lambda\sqrt{p_g}}{\|X_g^\top r_g\|_2}\right)_+ (X_g^\top r_g)$	L1 across groups + L2 within groups → block soft-thresholding