Regularization and Sparsity: From Lasso to Structured Penalties

1 Motivation

In high-dimensional settings, classical estimation methods such as ordinary least squares (OLS) often suffer from overfitting, instability, and poor interpretability. Regularization addresses these challenges by introducing additional constraints or penalties that control model complexity. In this lecture, we focus on sparsity-inducing regularization methods, beginning with lasso and extending to structured sparsity penalties.

2 Linear Regression and Regularization

Consider the linear regression model

$$y_i = x_i^\top \beta + \varepsilon_i, \quad i = 1, \dots, n,$$

where $x_i \in \mathbb{R}^p$ and $\beta \in \mathbb{R}^p$.

The ordinary least squares estimator solves

$$\hat{\beta}_{\text{OLS}} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2.$$

When p is large or predictors are highly correlated, OLS becomes unstable.

2.1 Ridge Regression

Ridge regression introduces an ℓ_2 penalty:

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

Ridge regression shrinks coefficients toward zero but does not produce exact zeros.

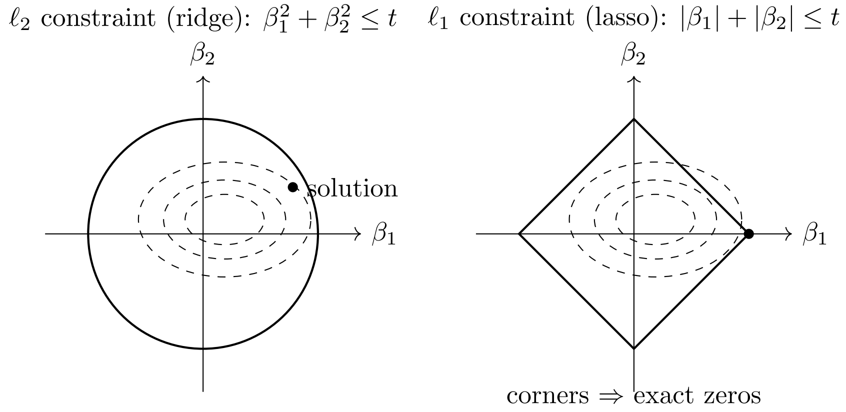
2.2 Lasso

The lasso replaces the ℓ_2 penalty with an ℓ_1 penalty:

$$\hat{\beta}_{\text{lasso}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

The lasso performs variable selection by producing exact zero coefficients.

3 Geometric Interpretation of Sparsity


ℓ_2 constraint (ridge): $\beta_1^2 + \beta_2^2 \le t$

ℓ_1 constraint (lasso): $|\beta_1| + |\beta_2| \le t$

Figure 1: Geometry explains sparsity: the ℓ_1 ball has corners aligned with coordinate axes, making exact zeros more likely than under ℓ_2 regularization.

Regularized estimators can be written in the general form

$$\min_{\beta} \; \mathcal{L}(\beta) + \lambda\,\Omega(\beta),$$

or equivalently,

$$\min_{\beta} \; \mathcal{L}(\beta) \quad \text{subject to} \quad \Omega(\beta) \le t.$$

3.1 ℓ_2 Geometry

The ridge constraint $\sum_j \beta_j^2 \le t$ defines a spherical constraint region with a smooth boundary. Loss contours typically intersect this region away from the coordinate axes, leading to shrinkage without sparsity.

3.2 ℓ_1 Geometry

The lasso constraint $\sum_j |\beta_j| \le t$ defines a diamond-shaped region with sharp corners aligned with the coordinate axes. Loss contours are likely to intersect the constraint at these corners, resulting in exact zeros.

4 Elastic Net

Elastic Net combines ℓ_1 and ℓ_2 penalties:

$$\min_{\beta} \ \mathcal{L}(\beta) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2.$$

Elastic Net was developed to address two limitations of lasso:

- Instability under highly correlated predictors
- The tendency of lasso to select at most n variables when $p \gg n$

The ℓ_2 component encourages correlated predictors to enter the model together (the *grouping effect*), while the ℓ_1 component preserves sparsity.

5 From Individual to Structured Sparsity

5.1 Group Lasso

In many applications, predictors naturally form groups. Let

$$\beta = (\beta_{(1)}, \dots, \beta_{(G)}),$$

where $\beta_{(g)} \in \mathbb{R}^{p_g}$ represents the coefficients in group g.

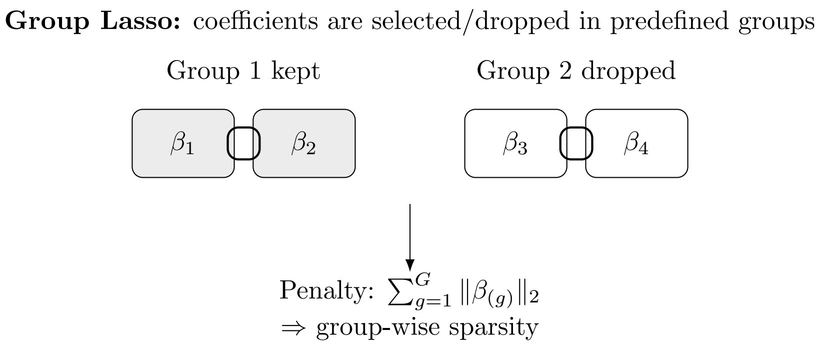

Group Lasso: coefficients are selected/dropped in predefined groups
Group 1 kept
Group 2 dropped
β_1
β_2
β_3
β_4
Penalty: $\sum_{g=1}^{G} \|\beta_{(g)}\|_2$
⇒ group-wise sparsity

Figure 2: Group lasso promotes sparsity at the group level: entire coefficient blocks are either set to zero or kept.

The group lasso solves

$$\min_{\beta} \ \mathcal{L}(\beta) + \lambda \sum_{g=1}^{G} \|\beta_{(g)}\|_2.$$

This penalty enforces sparsity at the group level: entire groups are either selected or excluded.

Examples

- Dummy variables for categorical predictors
- Basis expansions (splines, wavelets)
- Brain regions or functional networks in neuroimaging

5.2 Fused Lasso

Group lasso assumes known groups. When predictors are ordered, it may be more appropriate to penalize differences between neighboring coefficients.

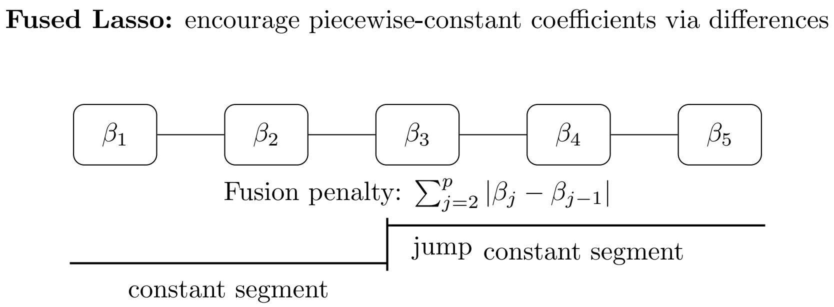

Fused Lasso: encourage piecewise-constant coefficients via differences
β_1
β_2
β_3
β_4
β_5
Fusion penalty: $\sum_{j=2}^{p} |\beta_j - \beta_{j-1}|$
jump
constant segment
constant segment

Figure 3: Fused lasso penalizes successive differences, encouraging piecewise-constant patterns with a small number of change points.

The fused lasso solves

$$\min_{\beta} \ \mathcal{L}(\beta) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|.$$

This penalty encourages both sparsity and piecewise-constant structure.

Applications

- Time series regression
- Genomic data
- Spatial and imaging data

5.3 Sparse Group Lasso

Sparse group lasso combines group-level and within-group sparsity:

$$\min_{\beta} \ \mathcal{L}(\beta) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{g=1}^{G} \|\beta_{(g)}\|_2.$$

This allows selection of important groups while retaining flexibility within groups.

5.4 Graph-Guided Lasso

Fused lasso can be generalized to arbitrary graphs. Let $G = (V, E)$ be a graph over predictors. The graph-guided lasso uses

$$\lambda \sum_{(j,k)\in E} w_{jk}|\beta_j - \beta_k|.$$

This encourages smoothness over graph-connected variables.

5.5 Multitask Lasso

In multitask learning, we observe responses

$$Y^{(t)} = X\beta^{(t)}, \quad t = 1, \dots, T.$$

The multitask lasso solves

$$\min_{\{\beta^{(t)}\}} \sum_{t=1}^{T} \mathcal{L}^{(t)}(\beta^{(t)}) + \lambda \sum_{j=1}^{p} \left\|(\beta_j^{(1)}, \dots, \beta_j^{(T)})\right\|_2.$$

This enforces shared sparsity across tasks.

5.6 Collaborative Learning for Multi-View Data

In many modern applications, each subject is measured using multiple *views* (or modalities).

Multi-view setup. Suppose each subject has V views of predictors:

$$X = \big(X^{(1)}, X^{(2)}, \dots, X^{(V)}\big),$$

where $X^{(v)} \in \mathbb{R}^{n\times p_v}$, and an additive multi-view regression model:

$$y = \sum_{v=1}^{V} X^{(v)}\beta^{(v)} + \varepsilon,$$

with view-specific coefficients $\beta^{(v)} \in \mathbb{R}^{p_v}$.

Collaborative (multi-view) lasso. Collaborative regression addresses settings where the same outcome is predicted using multiple data views, but the predictors across views are not directly comparable. Rather than enforcing similarity in coefficients, collaboration is imposed at the level of fitted values.

Two-view setup. Let $X \in \mathbb{R}^{n\times p_x}$ and $Z \in \mathbb{R}^{n\times p_z}$ denote two views of predictors measured on the same n subjects, and let $y \in \mathbb{R}^n$ be the response. Each view has its own coefficient vector, β_x

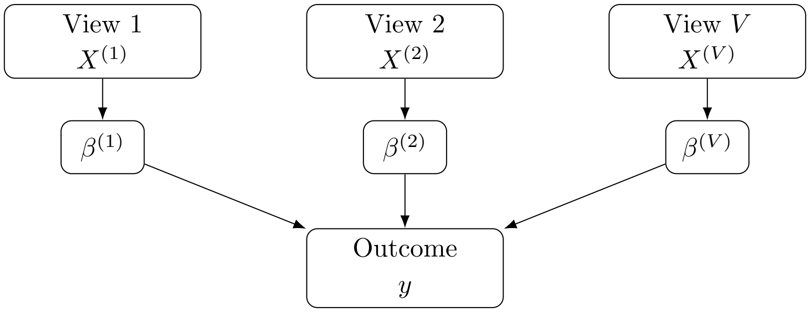


Figure 4: Collaborative (multi-view) learning: multiple feature views contribute to one outcome, with a penalty that encourages shared sparsity across views.

and β_z.

Collaborative regression objective.

$$\min_{\beta_x,\beta_z} \|y - X\beta_x\|_2^2 + \|y - Z\beta_z\|_2^2 + \rho\|X\beta_x - Z\beta_z\|_2^2 + \lambda_x\|\beta_x\|_1 + \lambda_z\|\beta_z\|_1.$$

Interpretation of terms.

- The first two terms measure goodness-of-fit for each view separately.
- The collaboration term $\|X\beta_x - Z\beta_z\|_2^2$ encourages the two views to produce similar predictions.
- The ℓ_1 penalties induce sparsity within each view.

Key idea. Collaboration is enforced at the *prediction level*, not the coefficient level. This allows different views to rely on different features while still agreeing on the predicted outcome.

When is this useful? Collaborative regression is particularly appropriate when:

- Predictors differ across views (e.g., multimodal data),
- Coefficients are not directly comparable,
- Agreement in predictions is scientifically meaningful.

6 Unifying Perspective

All methods discussed can be written as

$$\min_{\beta} \mathcal{L}(\beta) + \lambda\,\Omega(\beta),$$

where the penalty $\Omega(\beta)$ encodes prior structural assumptions.

Key takeaway:

Regularization is a mechanism for encoding structure and prior knowledge into learning algorithms.

References

- Zou, H. and Hastie, T. (2005). *Regularization and variable selection via the elastic net.* Journal of the Royal Statistical Society: Series B, 67(2), 301–320.
  (Elastic Net: bridges ridge and lasso; stabilizes variable selection under collinearity.)

- Yuan, M. and Lin, Y. (2006). *Model selection and estimation in regression with grouped variables.* Journal of the Royal Statistical Society: Series B, 68(1), 49–67.
  (Group Lasso: enforces sparsity at the group level.)

- Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). *Sparsity and smoothness via the fused lasso.* Journal of the Royal Statistical Society: Series B, 67(1), 91–108.
  (Fused Lasso: sparsity + smoothness through penalties on coefficient differences.)

- Jacob, L., Obozinski, G., and Vert, J.-P. (2009). *Group lasso with overlap and graph lasso.* Proceedings of the 26th International Conference on Machine Learning (ICML).
  (Graph-guided and overlapping group sparsity.)

- Obozinski, G., Wainwright, M. J., and Jordan, M. I. (2011). *Support union recovery in high-dimensional multivariate regression.* The Annals of Statistics, 39(1), 1–47.
  (Multitask Lasso: shared sparsity across related tasks.)

- Gross, S. M. and Tibshirani, R. (2015). *Collaborative regression.* Biostatistics, 16(2), 326–338.
  (Collaborative learning: joint sparsity across multiple data views for the same outcome.)

Extended Reading

- Ding, D. Y., Li, S., Narasimhan, B., and Tibshirani, R. (2022). *Cooperative learning for multiview analysis.* Proceedings of the National Academy of Sciences, 119(38), e2202113119.
  (Cooperative learning: modern framework for multiview integration extending collaborative regression.)