Penalized Regression

University of California, San Diego
Instructor: Armin Schwartzman

Introduction / Motivation

□ We have data $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$ and want to fit a linear model

$$\mathbb{E}(y|\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x}$$

As usual, define the response vector $\mathbf{y}$ and the design matrix $\mathbf{X}$.

□ Penalized regression amounts to performing least squares but with some additional constraints on the coefficients.

□ In numerical analysis (inverse problems), this is known as regularization and it is used when the optimization problem to be solved is under-determined.

□ Most penalties are on the magnitude of the coefficients. It is therefore important to standardize the predictor variables so they are unit-less. Here we assume that the response has been centered so that there is no need for an intercept. (In any case, the intercept is usually not penalized.)

Bias-Variance Decomposition

□ Assume an additive error regression model

$$y = f(\mathbf{x}) + \varepsilon,$$

where ε is independent of $\mathbf{x}$ with $\mathbb{E}\left(\varepsilon\right) = 0$ and $\mathrm{Var}\left(\varepsilon\right) = \sigma^2$.

□ Suppose $(\mathbf{x}_0, y_0)$ is a new observation, where $\mathbf{x}_0$ is fixed and $y_0 = f(\mathbf{x}_0) + \varepsilon_0$.

□ Recall that the EPE at $\mathbf{x}_0$ can be decomposed into:

$$\mathbb{E}\left((y_0 - \widehat{f}(\mathbf{x}_0))^2\right) = \sigma^2 + \mathrm{Bias}^2(\widehat{f}(\mathbf{x}_0)) + \mathrm{Var}(\widehat{f}(\mathbf{x}_0))$$

Ridge regression

- □ This method penalizes the Euclidean norm of the coefficients:

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} = \arg\min_{\mathbf{b}} \; \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 + \lambda\|\mathbf{b}\|_2^2$$

where

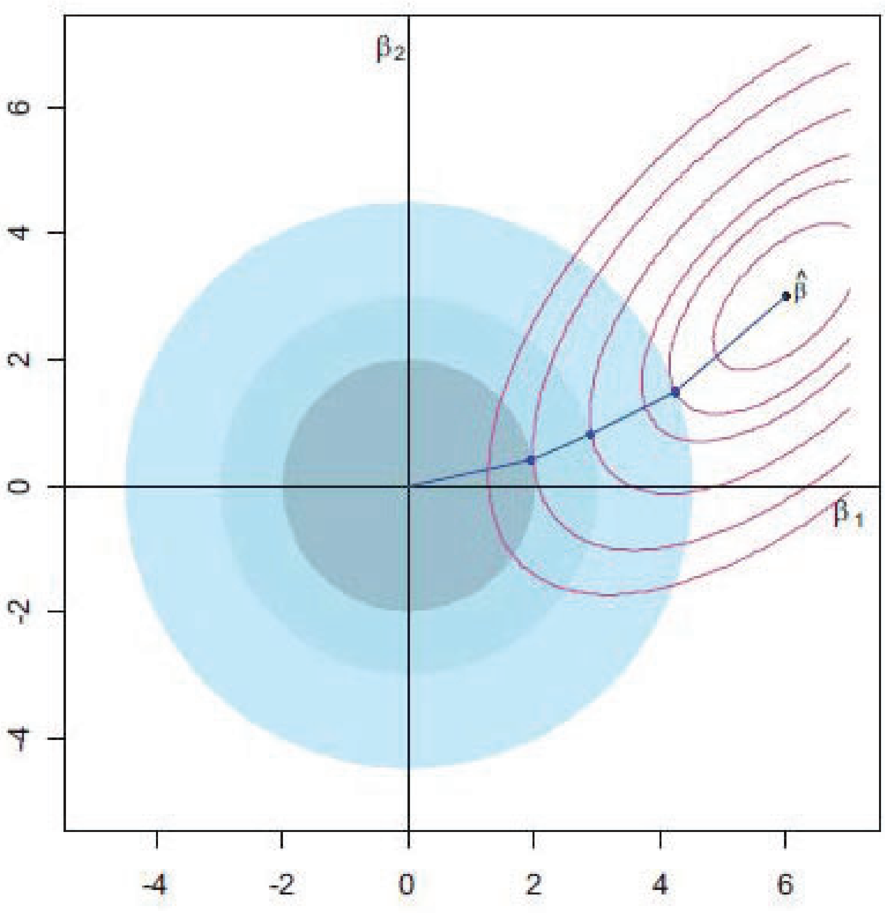
$$\|\mathbf{b}\|_2 = \Big(\sum_j b_j^2\Big)^{1/2} \quad \text{(Euclidean norm)}$$

- □ This is equivalent to solving

$$\min_{\mathbf{b}} \; \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2, \quad \text{subject to } \; \|\mathbf{b}\|_2 \le s$$

- □ λ (or s) is a tuning parameter.
- □ In inverse problems, this is known as Tikhonov regularization.

Ridge regression



Ridge regression: fit

□ Ridge regression can be solved explicitly:

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$$

□ The fitted values are

$$\hat{\mathbf{y}} = \mathbf{H}_\lambda\mathbf{y} = \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$$

□ The parameter λ controls the effective degrees of freedom:

$$\mathrm{df}(\lambda) = \mathrm{trace}[\mathbf{H}_\lambda] = \mathrm{trace}\left[\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\right]$$

(This is not necessarily an integer.)

□ When $\lambda = 0$, we get $\mathrm{df}(\lambda) = p$.

Ridge regression: Bias vs. variance

□ Assume that the linear model holds:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \mathbb{E}(\boldsymbol{\varepsilon}) = 0, \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$$

Then:

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}}] = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}$$
$$\mathrm{Var}[\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}}] = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\sigma^2$$

□ We can see that the ridge regression estimator is biased, but it also has smaller variance than the OLS estimator.

□ Choosing λ properly may result in a reduced EPE with respect to OLS.

□ The tuning parameter λ can be chosen by minimizing AIC, GCV or CV prediction error.

Generalized cross-validation (GCV)

□ GCV is an approximation to the leave-one-out EPE:

$$\mathrm{GCV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\hat{y}_i - y_i}{1 - \mathrm{df}(\lambda)/n}\right]^2$$

□ This formula depends on the effective degrees of freedom $\mathrm{df}(\lambda)$.

LASSO (Least Absolute Shrinkage and Selection Operator)

□ This method penalizes the ℓ^1-norm of the coefficients:

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{lasso}} = \arg\min_{\mathbf{b}} \ \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 + \lambda\|\mathbf{b}\|_1$$

where

$$\|\mathbf{b}\|_1 = \sum_j |b_j| \quad \text{(the } \ell^1\text{-norm)}$$

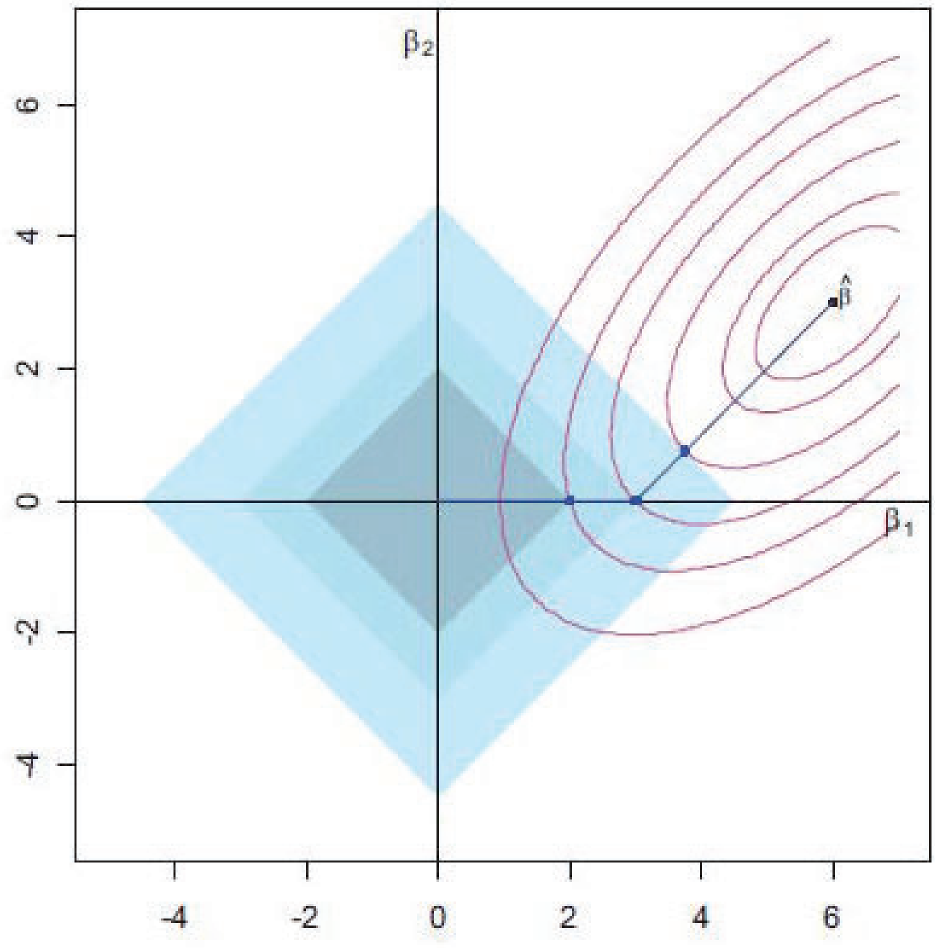
□ Equivalent to

$$\min_{\mathbf{b}} \ \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2, \quad \text{subject to } \ \|\mathbf{b}\|_1 \le s$$

□ λ (or s) is a tuning parameter.

□ No closed-form expression is available in general. Numerically, this is a convex problem, with dedicated software available (due to its popularity).

LASSO



□ This method is more recent but dates back (at least) to Tibshirani (1996) and has been studied extensively since around 2005.

□ Unlike ridge regression, the LASSO often sets some coefficients to 0. This is because the ℓ_1-ball has corners.

□ The latest theoretical developments show that, if the variables are only weakly correlated and the nonzero coefficients are large enough, then the LASSO recovers the true underlying model with high probability. See for example (Zhao and Yu, 2007).

The elastic-net

□ This variant, suggested by Zou and Hastie (2005), combines the ℓ_2 and ℓ_1 penalties:

$$\widehat{\boldsymbol{\beta}}^{\text{elastic}}_{\lambda_1,\lambda_2} = \arg\min_{\mathbf{b}} \ \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 + \lambda_1\|\mathbf{b}\|_1 + \lambda_2\|\mathbf{b}\|_2^2$$

□ Note that there are two tuning parameters now (λ_1 and λ_2).

□ This method was designed to select variables like the LASSO while shrinking together the coefficients of correlated predictors like ridge regression.

Best subset selection

□ This method penalizes the number of nonzero coefficients:

$$\widehat{\boldsymbol{\beta}}^{\text{best}}_{q} = \arg\min_{\mathbf{b}} \ \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2, \quad \text{subject to} \ \ \|\mathbf{b}\|_0 \leq q$$

where

$$\|\mathbf{b}\|_0 = \#\{j : b_j \neq 0\} \quad \text{(number of nonzero entries)}$$

□ Equivalently, the model thus obtained is the best model of size at most q.

□ Ridge regression and the LASSO are convex relaxations of best subset selection.

The case of orthogonal predictors

□ Suppose that $\mathbf{X}$ has orthonormal column vectors, meaning,

$$\mathbf{X} = [\mathbf{X}_1 \cdots \mathbf{X}_p] \quad \text{with} \quad \mathbf{X}_j^\top \mathbf{X}_k = \mathbb{I}\{j = k\}$$

□ Let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)$ denote the least squares coefficients when regressing $\mathbf{y}$ on $\mathbf{X}$, namely

$$\widehat{\beta}_j = \mathbf{X}_j^\top \mathbf{y}$$

□ Order these coefficients according to their absolute value. Let R_j denote the rank (in decreasing order) of $|\widehat{\beta}_j|$ among $|\widehat{\beta}_1|, \ldots, |\widehat{\beta}_p|$.

The case of orthogonal predictors

□ When the predictors are orthogonal, we have the following closed-form expressions:

Method	Formula for jth coefficient
Best subset (size q)	$\widehat{\beta}_j \mathbb{I}\{R_j \leq q\}$
Ridge (with parameter λ)	$\widehat{\beta}_j / (1 + \lambda)$
LASSO (with parameter λ)	$\operatorname{sign}(\widehat{\beta}_j) \left(\lvert\widehat{\beta}_j\rvert - \lambda/2\right)_+$

□ Note that best subset selection corresponds to hard thresholding of the least squares coefficients, ridge regression corresponds to shrinkage, and LASSO corresponds to soft thresholding.