Lasso, Ridge regression, and Elastic Net

Andreas Elsener

Motivation

Linear regression model: let $X \in \mathbb{R}^{n \times p}, \beta \in \mathbb{R}^p$, where n is the number of observations, p the number of predictors, and $\varepsilon \sim \mathcal{N}(0, 1_{n \times n})$

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

- The goal of a regression analysis is a good fit *and* an interpretable model.
- Forward/Backward selection, C_p-Mallows, AIC, BIC, etc.
- New idea: restrict the number of non-zero variables.

The optimization problem then looks like for some $s \leq p$

$$\text{minimize } \|y - X\beta\|_2^2, \text{ subject to } \|\beta\|_0 = \sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}} \leq s$$

BUT: This optimization problem is not convex!!!

Lasso

Solution: Take the $\|\cdot\|_1$ - norm as a convex surrogate of the $\|\cdot\|_0$-"norm". The optimization problem can then be formulated for some $t > 0$ as

$$\text{minimize } \|y - X\beta\|_2^2, \text{ subject to } \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j| \leq t.$$

Formulate this minimization problem as an unconstrained problem by introducing a Lagrange multiplier $\lambda > 0$. The Lasso estimator is then given by

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

Lasso

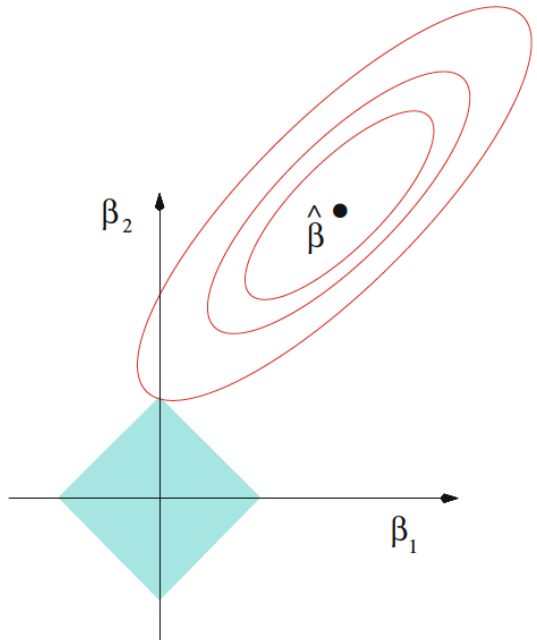
Why does the Lasso set some variables to 0? To see this consider the case $p = 2$, $n > p$, and remember that

$$\hat{\beta}_{\mathrm{LS}} = (X^T X)^{-1} X^T y.$$

We then have (blackboard) that

$$\|y - X\beta\|_2^2 = y^T y + (\beta - \hat{\beta}_{\mathrm{LS}})^T X^T X (\beta - \hat{\beta}_{\mathrm{LS}}) - \hat{\beta}_{\mathrm{LS}} X^T X \hat{\beta}_{\mathrm{LS}}.$$

The solid blue region corresponds to $|\beta_1| + |\beta_2| \leq s$. From Hastie et al., "The Elements of Statistical Learning", Springer:

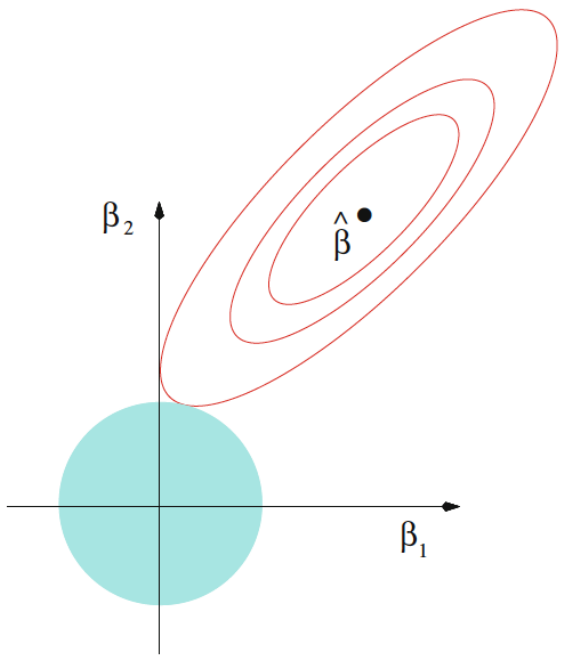


Ridge regression

Another possibility that doesn't have the same effect is to penalize by the $\|\cdot\|_2$ of the predictors.

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$

Advantage: analytic solution.
Disadvantage: no variable selection, just shrinkage, see exercise 1 of series 11. From Hastie et al., "The Elements of Statistical Learning", Springer:



Elastic net

If $p > n$ the lasso selects at most n variables. Also if there is a group of highly correlated predictors, then the lasso tends to select only one variable from a group and ignore the others.
To overcome these limitations the idea is to combine ridge regression and lasso. For $\lambda_1, \lambda_2 > 0$ the elastic net estimator is defined as

$$\hat{\beta}_{\mathsf{EN}} = \arg\min_{\beta} \left\{ \|y - X\beta\|_2^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1 \right\}.$$

Bayesian interpretation of the lasso

This example is taken from the lecture notes "Mathematical Statistics" written by Sara van de Geer. Consider the model

$$X_i = \theta_i + \varepsilon_i, i = 1, \ldots, n, \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1).$$

We then have $X_i \sim \mathcal{N}(\theta_i, 1)$ and the X_i are independent. The n parameters θ_i are all unknown. Define the estimator

$$\hat{\theta} = \arg\min_{\vartheta} \sum_{i=1}^{n} (X_i - \vartheta_i)^2 + 2\lambda \sum_{i=1}^{n} |\vartheta_i|,$$

where $\lambda > 0$ is a regularization parameter. The estimator is then given by

$$\hat{\theta}_i = \begin{cases} X_i - \lambda, & \text{if } X_i > \lambda, \\ 0, & \text{if } |X_i| \leq \lambda, \\ X_i + \lambda, & \text{if } X_i < -\lambda. \end{cases}$$

Suppose that $\theta_1, \ldots, \theta_n \overset{\text{i.i.d.}}{\sim} w(z) = \frac{1}{\tau\sqrt{2}} \exp\left[-\frac{\sqrt{2}|z|}{\tau}\right]$, where $w(z)$ is the density of the double-exponential distribution for $z \in \mathbb{R}$.

Bayesian interpretation of the lasso

To compute the posterior distribution $w(\vartheta|X_1,\ldots,X_n)$.

$$\begin{aligned} w(\vartheta|X_1,\ldots,X_n) &= \frac{w(\vartheta,X_1,\ldots,X_n)}{w(X_1,\ldots,X_n)} = \frac{w(\vartheta,X_1,\ldots,X_n)}{w(X_1,\ldots,X_n)}\frac{w(\vartheta)}{w(\vartheta)} \\ &\propto w(X_1,\ldots,X_n|\vartheta)w(\vartheta) \\ &= (2\pi)^{-n/2}\exp\left[-\frac{\sum_{i=1}^n(X_i-\vartheta_i)^2}{2}\right] \\ &\times (2\pi\tau)^{-n/2}\exp\left[-\frac{\sqrt{2}\sum_{i=1}^n|\vartheta_i|}{\tau}\right] \end{aligned}$$

Thus, we see that $\hat{\theta}$ with regularization parameter $\lambda = 2\sqrt{2}/\tau$ is the maximum a posteriori estimator.

Matrix completion via nuclear norm penalization

e.g. Netflix, Spotify, etc.

Users \ Films	Titanic	Goldfinger	The Da Vinci Code	Ocean's 13
Alice	1	NA	3.5	5
Bob	NA	1.4	NA	NA
Anna	1	NA	3.4	4.7
John	NA	NA	NA	2.8

- The matrix contains (possibly noisy) ratings and many missing entries (NA = not available). How can we fill in the gaps/predict the missing entries?
- We assume that the (noisy) observations are drawn randomly from the set of entries.
- Since the ratings/tastes are similar the matrix is assumed to have a low rank.

Matrix completion via nuclear norm penalization

We could penalize by the rank of the matrix but as in the regression case this quantity is not convex. Therefore, we use the nuclear norm as its convex surrogate:

$$\|B\|_{\text{nuclear}} = \sum_{k=1}^{q} \sigma_k,$$

where σ_k are the singular values of the matrix B. A matrix with rank r has exactly r non-zero singular values.
A possible estimator of the unknown entries is given by

$$\hat{B} = \arg \min_{B \in \mathbb{R}^{p \times q}} \|B|_{\Omega} - B^0|_{\Omega}\|_F^2 + \lambda \|B\|_{\text{nuclear}},$$

where Ω is the set of observed entries.