Chapter 6: Ridge Regression[a]

In this chapter we consider observations $\{(y_i^0, x_i^0)\}_{i=1}^n$ and assume the following linear model regression model

$$Y_i^0 = \alpha + \beta^\top x_i^0 + \epsilon_i, \quad i = 1, \dots, n \tag{6.1}$$

where $\beta \in \mathbb{R}^p$, $\alpha \in \mathbb{R}$ and where, for all $i, l \in \{1, \dots, n\}$, $\mathbb{E}[\epsilon_i] =$ and $\mathbb{E}[\epsilon_i \epsilon_l] = \sigma^2 \delta_{il}$ for some $\sigma^2 > 0$[b].

Remark: In this chapter the covariates $\{x_i^0\}_{i=1}^n$ are assumed to be deterministic (fixed design setting).

Assume first that $n \geq p$ and that $\text{rank}(\boldsymbol{X}^0) = p$. In this case, we can estimate (α, β) by ordinary least squares (OLS), that is we can estimate α and β using

$$\hat{\alpha} := \bar{y}^0 - \hat{\beta}^\top \bar{x}^0, \quad \hat{\beta} := \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \, \|y - \boldsymbol{X}\beta\|_2^2 = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top y.$$

Remark: This expression for $\hat{\alpha}$ and for $\hat{\beta}$ is obtained by applying Proposition 6.1 below with $\lambda = 0$.

Letting $Y = Y^0 - \frac{1}{n} \sum_{i=1}^n Y_i^0$, the corresponding OLS estimate $\hat{\mu}$ of $\mathbb{E}[Y]$ is given by

$$\hat{\mu} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top y = \mathbf{A}y$$

where $\mathbf{A} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$. Notice that under the model (6.1) the estimator[c] $\hat{\mu}$ is unbiased, that is $\mathbb{E}[\hat{\mu}] = \mathbb{E}[Y]$.

[a] The main reference for this chapter is [11].

[b] Recall that the intercept α in (6.1) allows to have estimators of β which are not affected by a shift of the response variables, that is, which are independent of $c \in \mathbb{R}$ if each y_i^0 is replaced by $y_i^0 + c$.

[c] In this chapter we make the distinction between an estimator, which is a random variable, and an estimate which is a realization of an estimator.

Remark: We focus on the estimation of $\mathbb{E}[Y]$ and not on $\mathbb{E}[Y^0]$ because $\mathbb{E}[Y]$ depends only on the main parameter of interest β.

Variance of the OLS in-sample predictions as p increases

Under the model (6.1) we have $\text{Var}(Y) = \sigma^2\big(\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\big)$. Then, noting that $\boldsymbol{X}^\top \mathbf{1}_n = \mathbf{0}_n$, it follows that under the model (6.1) the variance of the estimator $\hat{\mu}$ is given by

$$\begin{aligned}\text{Var}(\hat{\mu}) = \text{Var}(\boldsymbol{A}Y) &= \boldsymbol{A}\text{Var}(Y)\boldsymbol{A}^\top \\ &= \boldsymbol{A}\sigma^2\boldsymbol{I}_n\boldsymbol{A} - \frac{\sigma^2}{n}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{X}^\top\mathbf{1}_n)\boldsymbol{A} \\ &= \sigma^2\boldsymbol{A}^2 \\ &= \sigma^2\boldsymbol{A}\end{aligned}$$

Using the fact that $\text{tr}(\mathbf{BC}) = \text{tr}(\mathbf{CB})$, we remark that

$$\text{tr}(\mathbf{A}) = \text{tr}\{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\} = p$$

so that, under (6.1), $\hat{\mu}$ is such that $\frac{1}{n}\sum_{i=1}^n \text{Var}(\hat{\mu}_i) = \sigma^2\frac{p}{n}$.

Therefore, under (6.1) and as p grows, the average variance of the OLS estimators $\{\hat{\mu}_i\}_{i=1}^n$ of $\{\mathbb{E}[Y_i]\}_{i=1}^n$ increases, until reaching the value σ^2 when $p = n$.

Remark: If $p > n$ then $\boldsymbol{X}^\top\boldsymbol{X}$ is no longer invertible and therefore $\hat{\mu}$ does not exist.

On the other hand, if we simply estimate $\mathbb{E}[Y]$ by y then the resulting average variance of the estimators $\{Y_i\}_{i=1}^n$ of $\{\mathbb{E}[Y_i]\}_{i=1}^n$ is

$$\frac{1}{n}\sum_{i=1}^n \text{Var}(Y_i) = \sigma^2.$$

In words, as $p \to n$ the average variance of the OLS estimators $\{\hat{\mu}_i\}_{i=1}^n$ converges to the average variance of the naive estimators $\{Y_i\}_{i=1}^n$.

$\implies$ For $p \approx n$ the OLS estimate $\hat{\mu}$ of $\mathbb{E}[Y]$ is not better than the naive estimate y.

Linear regression in high dimension and ridge regression

As we just saw, for $p > n$ the OLS estimator of β cannot be computed and for $p \approx n$ the OLS estimator $\hat{\mu}$ performs poorly.

In this context, as discussed in Chapter 1 (see pages 31-32), a first approach that can be used to estimate β is principal component regression (PCR).

Ridge regression is a second possible approach to linear regression with high-dimensional data, which is based on the following lemma.

Lemma 6.1 *Let $\lambda > 0$ and $\gamma_1 \geq \cdots \geq \gamma_p$ be the p eigenvalues of the matrix $\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p$. Then, $\gamma_p \geq \lambda$.*

Proof: Let $l_1 \geq \cdots \geq l_p$ be the eigenvalues of $\boldsymbol{X}^\top \boldsymbol{X}$. Then, since for all $\beta \in \mathbb{R}^p$ we have $\beta^\top \boldsymbol{X}^\top \boldsymbol{X} \beta = \|\boldsymbol{X}\beta\|^2 \geq 0$, it follows that $l_j \geq 0$ for all $j \in \{1, \ldots, p\}$. Then, letting v_j be an eigenvector associated to the eigenvalue l_j, we have

$$(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p) v_j = \boldsymbol{X}^\top \boldsymbol{X} v_j + \lambda v_j = l_j v_j + \lambda v_j = (l_j + \lambda) v_j$$

showing that $l_j + \lambda \geq \lambda$ is an eigenvalue of the matrix $\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p$, with associated eigenvector v_j. The result follows. □

Building on the result of Lemma 6.1, for every $\lambda > 0$ the ridge estimate $(\hat{\alpha}_\lambda, \hat{\beta}_\lambda)$ of (α, β) is defined by

$$\hat{\alpha}_\lambda = \bar{y}^0 - \hat{\beta}_\lambda^\top \bar{x}^0, \quad \hat{\beta}_\lambda = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top y. \tag{6.2}$$

Corresponding optimization problem

As shown in the following proposition, $(\hat{\alpha}_\lambda, \hat{\beta}_\lambda)$ can be interpreted as a penalized least squares estimate of (α, β).

Proposition 6.1 *Let $(\hat{\alpha}_\lambda, \hat{\beta}_\lambda)$ the as defined in (6.2). Then,*

$$(\hat{\alpha}_\lambda, \hat{\beta}_\lambda) = \underset{\alpha\in\mathbb{R},\, \beta\in\mathbb{R}^p}{\operatorname{argmin}} \|y^0 - \alpha - \boldsymbol{X}^0\beta\|_2^2 + \lambda\|\beta\|_2^2. \tag{6.3}$$

It also holds true that

$$\hat{\beta}_\lambda = \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} \|y - \boldsymbol{X}\beta\|_2^2 + \lambda\|\beta\|_2^2.$$

Two important remarks:

1. In (6.3) the intercept is excluded from the penalty term to make $\hat{\beta}_\lambda$ independent of $\hat{\alpha}_\lambda$[a].

2. The input variables $\{x_{(j)}\}_{j=1}^p$ should all be on the same scale to ensure that that the size of the components $\{\beta_j\}_{j=1}^p$ of β is comparable, and thus that the penalty $\lambda\|\beta\|$ appearing in (6.3) makes sense.

   If the variables are not on the same scale we can proceed as follows: Letting $\boldsymbol{D} = \operatorname{diag}(s_1^2, \ldots, s_p^2)$, $\tilde{\boldsymbol{X}}^0 = \boldsymbol{X}^0\boldsymbol{D}^{-1/2}$ and $\gamma = \boldsymbol{D}^{1/2}\beta$, we can rewrite (6.1) as

   $$Y^0 = \alpha + \boldsymbol{X}^0\boldsymbol{D}^{-1/2}\big(\boldsymbol{D}^{1/2}\beta\big) + \epsilon = \alpha + \tilde{\boldsymbol{X}}^0\gamma + \epsilon$$

   and compute the ridge regression estimate $(\hat{\alpha}_\lambda, \hat{\gamma}_\lambda)$ of (α, γ) using the normalized variables $\{\tilde{x}_{(j)}^0\}_{j=1}^p$. We then estimate β using $\tilde{\beta}_\lambda = \boldsymbol{D}^{-1/2}\,\hat{\gamma}_\lambda$.

[a] In particular, if α was in the penalty term then adding an arbitrary constant $c \neq 0$ to each observation y_i^0 would modify the value of all the components of $\hat{\beta}_\lambda$. In this case, the estimated slope parameters would have the undesirable property be affected by an arbitrary shift of the response variables $\{y_i^0\}_{i=1}^n$.

Proof of Proposition 6.1

Let $F(\alpha, \beta) = \sum_{i=1}^{n} \left(y_i^0 - \alpha - \beta^\top x_i^0\right)^2 + \lambda \|\beta\|_2^2$. Simple computations show that F is strictly convex for all $\lambda > 0$, implying that the global minimizer of this function is unique.

For all $\beta \in \mathbb{R}^p$ let $\alpha_\beta = \operatorname{argmin}_{\alpha \in \mathbb{R}} F(\alpha, \beta)$ so that to prove the proposition we need to show that

$$F(\hat{\alpha}_\lambda, \hat{\beta}_\lambda) = \min_{\alpha \in \mathbb{R},\, \beta \in \mathbb{R}^p} F(\alpha, \beta) = \min_{\beta \in \mathbb{R}^p} F(\alpha_\beta, \beta).$$

We have

$$0 = \frac{\partial}{\partial \alpha} F(\alpha, \beta)\Big|_{(\alpha,\beta)=(\alpha_\beta,\beta)} \Leftrightarrow \alpha_\beta = \bar{y}^0 - \beta^\top \bar{x}^0, \quad \forall \beta \in \mathbb{R}^p \tag{6.4}$$

and thus

$$\begin{aligned}
\operatorname*{argmin}_{\beta \in \mathbb{R}^p} F(\alpha_\beta, \beta) &= \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \|y^0 - \alpha_\beta - \boldsymbol{X}^0 \beta\|_2^2 + \lambda \|\beta\|_2^2 \\
&= \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \|y - \boldsymbol{X}\beta\|_2^2 + \lambda \|\beta\|_2^2 \\
&= (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top y \\
&= \hat{\beta}_\lambda.
\end{aligned}$$

Using (6.4) it follows that $\alpha_{\hat{\beta}_\lambda} = \bar{y}^0 - \hat{\beta}_\lambda^\top \bar{x}^0 = \hat{\alpha}_\lambda$ and the proof is complete. □

$\hat{\beta}_\lambda$ as a shrinkage estimator of β

Proposition 6.1 shows that ridge regression imposes a penalty on the size of β. The strength of the penalty depends on the parameter λ, with the larger λ the smaller $\|\hat{\beta}_\lambda\|$.

This claim is formalized in the following proposition.

Proposition 6.2 *Assume that $\boldsymbol{X}^\top\boldsymbol{X}$ is invertible. Then, under the model (6.1), we have $\|\mathbb{E}[\hat{\beta}_\lambda]\| \leq \|\mathbb{E}[\hat{\beta}_{\lambda_0}]\|$ for all $\lambda > \lambda_0 \geq 0$.*

Proof: For every $\lambda \geq 0$ remark that, under the model (6.1), we have

$$\mathbb{E}[\hat{\beta}_\lambda] = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top\mathbb{E}[Y] = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top\boldsymbol{X}\beta \tag{6.5}$$

and let $\boldsymbol{B}_\lambda = \boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top\boldsymbol{X}$.

Let $\lambda > \lambda_0 \geq 0$ and remark that

$$\|\mathbb{E}[\hat{\beta}_{\lambda_0}]\|^2 - \|\mathbb{E}[\hat{\beta}_\lambda]\|^2 = \beta^\top(\boldsymbol{B}_{\lambda_0} - \boldsymbol{B}_\lambda)\beta$$

so that to prove the proposition it is enough to show that $\boldsymbol{B}_{\lambda_0} - \boldsymbol{B}_\lambda \succ 0$.

Since the matrix $\boldsymbol{X}^\top\boldsymbol{X}$ is assumed to be invertible, the matrices $\boldsymbol{B}_{\lambda_0}$ and $\boldsymbol{B}_\lambda$ are invertible and thus

$$\boldsymbol{B}_{\lambda_0} - \boldsymbol{B}_\lambda \succ 0 \Leftrightarrow \boldsymbol{B}_\lambda^{-1} - \boldsymbol{B}_{\lambda_0}^{-1} \succ 0.$$

Since $\lambda > \lambda_0$, and noting that since $\boldsymbol{X}^\top\boldsymbol{X}$ is invertible the matrices $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ and $(\boldsymbol{X}^\top\boldsymbol{X})^{-2}$ are positive definite[a], we have

$$\begin{aligned}
&\boldsymbol{B}_\lambda^{-1} - \boldsymbol{B}_{\lambda_0}^{-1} \\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\Big((\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_p)(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_p) - (\boldsymbol{X}^\top\boldsymbol{X} + \lambda_0\boldsymbol{I}_p)(\boldsymbol{X}^\top\boldsymbol{X} + \lambda_0\boldsymbol{I}_p)\Big)(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\Big(2\boldsymbol{X}^\top\boldsymbol{X}(\lambda - \lambda_0) + (\lambda^2 - \lambda_0^2)\boldsymbol{I}_p\Big)(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \\
&= 2(\lambda - \lambda_0)(\boldsymbol{X}^\top\boldsymbol{X})^{-1} + (\lambda^2 - \lambda_0^2)(\boldsymbol{X}^\top\boldsymbol{X})^{-2} \\
&\succ 0
\end{aligned}$$

and the result follows. □

[a] Indeed, $\boldsymbol{X}^\top\boldsymbol{X}$ is semi-definite positive and invertible and thus positive definite. Then, the fact that for $k \in \mathbb{N}$ the matrix $(\boldsymbol{X}^\top\boldsymbol{X})^{-k}$ is positive definite can be directly checked from the eigen-decomposition of the matrix $\boldsymbol{X}^\top\boldsymbol{X}$.

Variance of $\hat{\beta}_\lambda$ under the model (6.1)

Proposition 6.2 implies that, when $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible, unlike the OLS estimator $\hat{\beta}$ the ridge estimator $\hat{\beta}_\lambda$ is biased under the model (6.1).

As shown in the following proposition, $\hat{\beta}_\lambda$ has however the advantage to have a smaller variance.

Proposition 6.3 *Assume that $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible. Then, under the model (6.1), we have $\mathrm{Var}(\hat{\beta}_{\lambda_0}) - \mathrm{Var}(\hat{\beta}_\lambda) \succ 0$ for all $\lambda > \lambda_0 \geq 0$.*

Proof: Recall that under the model (6.1) we have $\mathrm{Var}(Y) = \sigma^2\big(\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\big)$ and note that $\boldsymbol{X}^\top \mathbf{1}_n = \mathbf{0}_n$. Therefore, under the model (6.1), for all $\lambda > 0$ we have

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_\lambda) &= \sigma^2(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \mathrm{Var}(Y) \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \\
&= \sigma^2(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \big(\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\big) \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \\
&= \sigma^2(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}.
\end{aligned}$$

Let $\lambda > \lambda_0 \geq 0$ and note that, since by assumption the matrix $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible, we have

$$\mathrm{Var}(\hat{\beta}_{\lambda_0}) - \mathrm{Var}(\hat{\beta}_\lambda) \succ 0 \Leftrightarrow \mathrm{Var}(\hat{\beta}_\lambda)^{-1} - \mathrm{Var}(\hat{\beta}_{\lambda_0})^{-1} \succ 0.$$

Simple computations show that

$$\frac{\mathrm{Var}(\hat{\beta}_\lambda)^{-1} - \mathrm{Var}(\hat{\beta}_{\lambda_0})^{-1}}{\sigma^2} = 2(\lambda - \lambda_0)\boldsymbol{I}_p + (\lambda^2 - \lambda_0^2)(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$$

and, since $(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \succ 0$, the proposition is proved. □

Remark: Compared to $\hat{\beta}$, for all $\lambda > 0$ and under (6.1) the estimator $\hat{\beta}_\lambda$ has therefore a larger bias and a smaller variance, and a natural question is which of these two estimators has the lowest mean squared error (MSE). It can be shown (see [11], Theorem 1.2) that there exists a $\lambda > 0$ such that $\hat{\beta}_\lambda$ has a smaller MSE than $\hat{\beta}$ under (6.1), that is that under (6.1) there exists a $\lambda > 0$ such that

$$\mathbb{E}[\|\hat{\beta}_\lambda - \beta\|^2] < \mathbb{E}[\|\hat{\beta} - \beta\|^2].$$

A useful technical lemma

Lemma 6.2 *Let* $\lambda > 0$ *and* $\mathbf{A}^{(\lambda)} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top$. *Then,* $a_{ii}^{(\lambda)} \in [0, 1)$ *for all* $i \in \{1, \dots n\}$.

Proof: We have

$$\boldsymbol{I}_p - \mathbf{A}^{(\lambda)} = \boldsymbol{X}\Big((\boldsymbol{X}^\top \boldsymbol{X})^{-1} - (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\Big)\boldsymbol{X}^\top$$

and therefore, recalling that for two invertible matrices $\boldsymbol{C}$ and $\boldsymbol{B}$ we have $\boldsymbol{C} \succ \boldsymbol{B} \Leftrightarrow \boldsymbol{B}^{-1} \succ \boldsymbol{C}^{-1}$, it follows that $\boldsymbol{I}_p - \mathbf{A}^{(\lambda)}$ is a positive definite matrix (since $\lambda > 0$).

Therefore, all the diagonal elements of the matrix $\boldsymbol{I}_p - \mathbf{A}^{(\lambda)}$ are strictly positive[a], showing that $a_{ii}^{(\lambda)} < 1$ for all i.

On the other hand, since $\mathbf{A}^{(\lambda)}$ is semi-definite positive then $a_{ii}^{(\lambda)} \geq 0$ for all i. The proof is complete. □

[a] Indeed, if $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ is positive definite and e.g. $m_{11} \leq 0$ then for $v = (1, 0, \dots, 0) \in \mathbb{R}^n$ we have $v^\top \boldsymbol{M} v = m_{11} \leq 0$.

Choosing the penality parameter λ

When p is large compared to n and λ is too small then $\hat{\beta}_\lambda$ will be such that $\|y - \boldsymbol{X}\hat{\beta}_\lambda\|_2^2 \approx 0$, in which case we will over-fit the data. On the other hand, it is clear from (6.2) that $\hat{\beta}_\lambda \to 0$ as $\lambda \to \infty$, and thus that if λ is too large then we will under-fit the data.

In practice we choose λ so that the model has good out-of-sample predictive performance. One way to achieve this is to use cross validation.

Letting $\hat{\beta}_{-i,\lambda}$ be the ridge estimate of β computed from all the observations but (y_i, x_i), in leave-one-out ordinary cross validation (OCV) we let $\lambda = \hat{\lambda}$ where $\hat{\lambda}$ is, for some set $\Lambda \subseteq [0, \infty)$, defined by

$$\hat{\lambda} = \underset{\lambda \in \Lambda}{\operatorname{argmin}} \, \mathrm{OCV}_{\mathrm{ridge}}(\lambda), \quad \mathrm{OCV}_{\mathrm{ridge}}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - x_i^\top \hat{\beta}_{-i,\lambda}\right)^2.$$

Remark: In practice Λ is often a finite set and $\hat{\lambda}$ is obtained by computing $\mathrm{OCV}_{\mathrm{ridge}}(\lambda)$ for all $\lambda \in \Lambda$.

This definition of $\mathrm{OCV}_{\mathrm{ridge}}(\lambda)$ suggests that we need to perform n regressions to compute this quantity. However, by Theorem 6.1 below, for all $\lambda > 0$ we have

$$\mathrm{OCV}_{\mathrm{ridge}}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - x_i^\top \hat{\beta}_\lambda)^2}{(1 - a_{ii}^{(\lambda)})^2} = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_{\lambda,i})^2}{(1 - a_{ii}^{(\lambda)})^2} \tag{6.6}$$

with $\mathbf{A}^{(\lambda)}$ as defined in Lemma 6.2 and with $\hat{\mu}_\lambda = \boldsymbol{X}\hat{\beta}_\lambda$ the ridge estimate of $\mathbb{E}[Y]$. Therefore, only one regression is needed to compute $\mathrm{OCV}_{\mathrm{ridge}}(\lambda)$.

Remark: By lemma 6.2 we have $a_{ii}^{(\lambda)} \in [0, 1)$ for all $i \in \{1, \ldots, n\}$ and all $\lambda > 0$, and thus $\mathrm{OCV}_{\mathrm{ridge}}(\lambda)$ is well-defined for all $\lambda > 0$.

A key result for cross-validation

The equality in (6.6) is obtained by applying the following theorem with $\boldsymbol{M} = \boldsymbol{I}_p \lambda$.

Theorem 6.1 *Let $\boldsymbol{M} \in \mathbb{R}^{p \times p}$ be a semi-definite positive matrix such that the matrix $(\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{M})$ is invertible. Let*

$$\beta_{\boldsymbol{M}} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|y - \boldsymbol{X}\beta\|_2^2 + \beta^\top \boldsymbol{M} \beta$$

and assume that, for all $i \in \{1, \ldots, n\}$, the function

$$\mathbb{R}^p \ni \beta \mapsto \sum_{l \neq i} \left(y_l - \beta^\top x_l\right)^2 + \beta^\top \boldsymbol{M} \beta$$

has a unique global minimizer $\beta_{-i,\boldsymbol{M}} \in \mathbb{R}^p$. Let

$$\boldsymbol{A}^{(\boldsymbol{M})} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{M})^{-1} \boldsymbol{X}^\top$$

and assume that $|a_{ii}^{(\boldsymbol{M})}| \neq 1$ for all $i \in \{1, \ldots, n\}$. Then,

$$\frac{1}{n} \sum_{i=1}^n \left(y_i - x_i^\top \beta_{-i,\boldsymbol{M}}\right)^2 = \frac{1}{n} \sum_{i=1}^n \frac{(y_i - x_i^\top \beta_{\boldsymbol{M}})^2}{(1 - a_{ii}^{(\boldsymbol{M})})^2}.$$

Proof of Theorem 6.1

Let $i \in \{1, \ldots, n\}$, $\tilde{y}^{(\boldsymbol{M},-i)}$ denote the vector y where the ith element has been replaced by $x_i^\top \beta_{-i,\boldsymbol{M}}$ and

$$L_{-i}(\beta) = \sum_{l \neq i}^{n} (y_l - x_l^\top \beta)^2 + \beta^\top \boldsymbol{M} \beta.$$

Then, $\nabla L_{-i}(\beta) = -2 \sum_{l \neq i} x_l \big(y_l - x_l^\top \beta\big) + 2\boldsymbol{M}\beta$ for all $\beta \in \mathbb{R}^p$, and thus

$$\begin{aligned}
\nabla L_{-i}(\beta_{-i,\boldsymbol{M}}) = 0 &\Leftrightarrow -2 \sum_{l \neq i}^{n} x_l \big(y_l - x_l^\top \beta_{-i,\boldsymbol{M}}\big) + 2\boldsymbol{M}\beta_{-i,\boldsymbol{M}} = 0 \\
&\Leftrightarrow -2 \sum_{l=1}^{n} x_l \big(\tilde{y}_l^{(\boldsymbol{M},-i)} - x_l^\top \beta_{-i,\boldsymbol{M}}\big) + 2\boldsymbol{M}\beta_{-i,\boldsymbol{M}} = 0 \\
&\Leftrightarrow -2\boldsymbol{X}^\top \tilde{y}^{(\boldsymbol{M},-i)} + 2\big(\boldsymbol{X}\boldsymbol{X}^\top + \boldsymbol{M}\big)\beta_{-i,\boldsymbol{M}} = 0 \\
&\Leftrightarrow \beta_{-i,\boldsymbol{M}} = (\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{M})^{-1} \boldsymbol{X}^\top \tilde{y}^{(\boldsymbol{M},-i)}.
\end{aligned}$$

Using this expression for $\beta_{-i,\boldsymbol{M}}$, we obtain

$$\begin{aligned}
x_i^\top \beta_{-i,\boldsymbol{M}} &= \big(a_i^{(\boldsymbol{M})}\big)^\top \tilde{y}^{(\boldsymbol{M},-i)} \\
&= \big(a_i^{(\boldsymbol{M})}\big)^\top y + \big(a_i^{(\boldsymbol{M})}\big)^\top \big(\tilde{y}^{(\boldsymbol{M},-i)} - y\big) \\
&= \big(a_i^{(\boldsymbol{M})}\big)^\top y + a_{ii}^{(\boldsymbol{M})} \big(x_i^\top \beta_{-i,\boldsymbol{M}} - y_i\big) \\
&= x_i^\top \beta_{\boldsymbol{M}} - a_{ii}^{(\boldsymbol{M})} \big(y_i - x_i^\top \beta_{-i,\boldsymbol{M}}\big)
\end{aligned}$$

showing that

$$y_i - x_i^\top \beta_{\boldsymbol{M}} = (1 - a_{ii}^{(\boldsymbol{M})}) \big(y_i - x_i^\top \beta_{-i,\boldsymbol{M}}\big).$$

The result follows. □

Generalized cross validation: preliminaries

Let $\boldsymbol{G} \in O(n)$ and consider the transformation $y \mapsto y_{\boldsymbol{G}} := \boldsymbol{G}y$ and $\boldsymbol{X} \mapsto \boldsymbol{X}_{\boldsymbol{G}} := \boldsymbol{G}\boldsymbol{X}$ of the data.

Then, it is easily checked that the resulting ridge regression estimate $\hat{\beta}_{\boldsymbol{G},\lambda}$ of β is given by

$$\hat{\beta}_{\boldsymbol{G},\lambda} \in \underset{\alpha\in\mathbb{R},\, \beta\in\mathbb{R}^p}{\operatorname{argmin}} \|y_{\boldsymbol{G}} - \boldsymbol{X}_{\boldsymbol{G}}\beta\|_2^2 + \lambda\|\beta\|_2^2 = \hat{\beta}_\lambda$$

while, letting $\hat{\mu}_\lambda^{(\boldsymbol{G})} = \boldsymbol{X}_{\boldsymbol{G}}\hat{\beta}_\lambda = \boldsymbol{G}\hat{\mu}_\lambda$, the resulting OCV criterion is

$$\mathrm{OCV}_{\mathrm{ridge}}^{(\boldsymbol{G})}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \frac{\left(y_{\boldsymbol{G},i} - \hat{\mu}_{\lambda,i}^{(\boldsymbol{G})}\right)^2}{\left(1 - a_{ii}^{(\boldsymbol{G},\lambda)}\right)^2}$$

where

$$\boldsymbol{A}^{(\boldsymbol{G},\lambda)} = \boldsymbol{X}_{\boldsymbol{G}}(\boldsymbol{X}_{\boldsymbol{G}}\boldsymbol{X}_{\boldsymbol{G}}^\top + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}_{\boldsymbol{G}}^\top = \boldsymbol{G}\boldsymbol{A}_\lambda\boldsymbol{G}^\top.$$

Therefore, applying the rotation $\boldsymbol{G}$ to the observations $\{(y_i, x_i)\}_{i=1}^n$ leaves the ridge regression estimate $\hat{\beta}_\lambda$ unchanged but, in general, modifies the OCV criterion.

Given this dependence of OCV (and therefore of the resulting choice of λ) to the choice of $\boldsymbol{G}$ one can wonder what is a "bad" rotation $\boldsymbol{G}$ of the data in term of cross validation that we should avoid.

The generalized cross validation criterion

Intuitively, if $\boldsymbol{G}$ is such that we have highly uneven values of $a_{ii}^{(\boldsymbol{G},\lambda)}$ then the value of $\mathrm{OCV}_{\mathrm{ridge}}^{(\boldsymbol{G})}(\lambda)$ will tend to be dominated by a small number of data points.

To avoid this problem, a natural idea is to apply OCV using a rotation $\boldsymbol{G}_* \in O(n)$ of the data such that

$$a_{ii}^{(\boldsymbol{G}_*,\lambda)} = a_{ll}^{(\boldsymbol{G}_*,\lambda)}, \quad \forall i, l \in \{1, \dots, n\}.$$

Remark: It can be shown that such a matrix $\boldsymbol{G}_*$ indeed exists (see [13], Section 6.2.3, page 258).

Noting that

$$\mathrm{tr}(\boldsymbol{A}^{(\boldsymbol{G},\lambda)}) = \mathrm{tr}(\boldsymbol{G}\boldsymbol{A}^{(\lambda)}\boldsymbol{G}^\top) = \mathrm{tr}(\boldsymbol{A}^{(\lambda)}), \quad \forall \boldsymbol{G} \in O(n),$$

it follows that $\boldsymbol{G}^*$ is such that $a_{ii}^{(\boldsymbol{G}_*,\lambda)} = \mathrm{tr}(\boldsymbol{A}^{(\lambda)})/n$ for all i.

Therefore,

$$\begin{aligned}
\mathrm{OCV}_{\mathrm{ridge}}^{(\boldsymbol{G}_*)}(\lambda) &= \frac{1}{n} \frac{\sum_{i=1}^n \left(y_{\boldsymbol{G}_*,i} - \hat{\mu}_{\lambda,i}^{(\boldsymbol{G}_*)}\right)^2}{(1 - \mathrm{tr}(\boldsymbol{A}^{(\lambda)})/n)^2} \\
&= \frac{n\|y - \hat{\mu}_\lambda\|^2}{(n - \mathrm{tr}(\boldsymbol{A}^{(\lambda)}))^2} \\
&=: \mathrm{GCV}_{\mathrm{ridge}}(\lambda).
\end{aligned}$$

Choosing λ which minimizes $\lambda \mapsto \mathrm{GCV}_{\mathrm{ridge}}(\lambda)$ is called generalized cross validation.

Remark: By Lemma 6.2 we have $\mathrm{tr}(\boldsymbol{A}^{(\lambda)}) \in [0, n)$ for all $\lambda > 0$ and thus $\mathrm{GCV}_{\mathrm{ridge}}(\lambda)$ is well defined for every $\lambda > 0$.

Bayesian perspective of ridge regression

Consider the following Bayesian linear regression model

$$\beta \sim \mathcal{N}_p(0, \boldsymbol{I}_p\sigma^2/\lambda), \quad Y_i \sim \mathcal{N}_1(x_i^\top \beta, \sigma^2), \quad i = 1, \ldots, n. \tag{6.7}$$

By definition, the posterior distribution of β given the observation y is $\pi(\beta|y) \propto \pi(y|\beta)\pi(\beta)$, and simple computations show that

$$\beta|y \sim \mathcal{N}_p(\hat{\beta}_\lambda, (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\sigma^2). \tag{6.8}$$

We therefore see that the ridge estimator $\hat{\beta}_\lambda$ is both the posterior mean and the posterior mode of β in the Bayesian model (6.7). Hence, in (6.7), the prior distribution for β acts as a penalty on $\|\beta\|$. In other words, the prior distribution leads the posterior distribution to favour values of β such that $\|\beta\|$ is small.

To interpret the posterior variance of β note that under the model (6.7) we have

$$\mathrm{Var}(\hat{\beta}_\lambda) = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\sigma^2$$

while, using the fact that $\boldsymbol{I}_p = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)$ and (6.5), it is easily checked that, under (6.7),

$$b(\beta) := \mathbb{E}[\hat{\beta}_\lambda|\, \beta] - \beta = \mathbb{E}[\hat{\beta}_\lambda|\, \beta] - \beta = \big(\frac{1}{\lambda}\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{I}_p\big)^{-1}\beta.$$

Therefore, $(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\sigma^2 = \mathrm{Var}(\hat{\beta}_\lambda) + \mathbb{E}_{\text{prior}}[b(\beta)b(\beta)^\top]$ which, with (6.8), shows that the Bayesian posterior covariance matrix for β can be viewed as the sum of the covariance matrix of $\hat{\beta}_\lambda$ (under (6.7)) and of the prior expected squared bias of $\hat{\beta}_\lambda$ (under (6.7)).

An illustrative example

We let $n = 40$, $p = 50$ and simulate the covariates $\{x_i^0\}_{i=1}^n$ using $X_{ij}^0 \overset{\text{iid}}{\sim} \mathcal{U}(0,1)$ and the response variable $\{y_i^0\}_{i=1}^n$ using

$$Y_i^0 = \beta_*^\top x_i^0 + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}_1(0,1), \quad i = 1, \ldots, n$$

where $\beta_{*,j}$ is a random draw from the $\mathcal{U}(0,1)$ distribution for $j = 1, \ldots, 10$ while $\beta_{*,j} = 0$ for $j > 10$. For this example we consider the linear model (6.1) without intercept and estimate β using the non-centred data $\{(y_i^0, x_i^0)\}_{i=1}^n$.

From the results presented in Figure 6.1, we see that for this example OCV allows to choose a λ such that the mean squared error (MSE) of $\hat{\mu}_\lambda$ (for estimating $\mathbb{E}[Y^0]$) is very close to the one we could achieve in the ideal scenario where we could choose λ knowing $\mathbb{E}[Y^0]$.

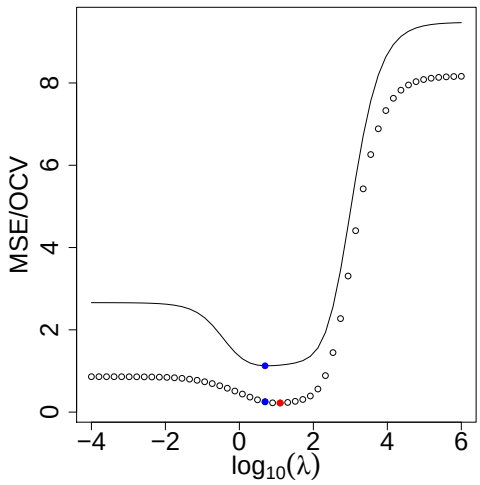

Figure 6.1: The dots show the mapping $\lambda \mapsto \text{MSE}(\lambda) := \frac{1}{n}\|\hat{\mu}_\lambda - \boldsymbol{X}^0\beta^*\|^2$ and the solid line the mapping $\lambda \mapsto \text{OCV}_{\text{ridge}}(\lambda)$. The red dot is for $\lambda^* = \text{argmin}_\lambda \text{MSE}(\lambda)$ and the blue dots are for $\hat{\lambda} = \text{argmin}_\lambda \text{OCV}_{\text{ridge}}(\lambda)$.

Chapter 7: LASSO Regression[a]

As in the previous chapter we consider observations $\{(y_i^0, x_i^0)\}_{i=1}^n$ and assume the following linear regression model

$$Y_i^0 = \alpha + \beta^\top x_i^0 + \epsilon_i, \quad i = 1, \dots, n \tag{7.1}$$

where $\beta \in \mathbb{R}^p$, $\alpha \in \mathbb{R}$ and where, for all $i, l \in \{1, \dots, n\}$, $\mathbb{E}[\epsilon_i] =$ and $\mathbb{E}[\epsilon_i \epsilon_l] = \sigma^2 \delta_{il}$ for some $\sigma^2 > 0$.

For $r \geq 0$ let

$$(\hat{\alpha}_\lambda^{(r)}, \hat{\beta}_\lambda^{(r)}) \in \underset{\alpha \in \mathbb{R},\ \beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\| y^0 - \alpha - \boldsymbol{X}^0 \beta \right\|_2^2 + 2\lambda \sum_{j=1}^p |\beta_j|^r \tag{7.2}$$

with the convention $\sum_{j=1}^p |\beta_j|^q = \sum_{j=1}^p \mathbf{1}_{\mathbb{R}\setminus\{0\}}(\beta_j)$ when $r = 0$.

The ridge estimator $(\hat{\alpha}_{2\lambda}, \hat{\beta}_{2\lambda})$ corresponds to the case $r = 2$ and, as we saw in the previous chapter, shrinks the regression coefficient towards zero. However, none of the components of $\hat{\beta}_{2\lambda}$ is shrink exactly to zero.

In practice, it is sometimes convenient (notably to facilitate the interpretation of the fitted model) to have some regression coefficients exactly equal to zero, so that the corresponding effect is dropped from the model. In particular, for large p we often would like to have a sparse estimate $\tilde{\beta}$ of β, where $\tilde{\beta}_j = 0$ for many $j \in \{1, \dots, p\}$. As we will see in this chapter, this is exactly what LASSO regression achieves.

Remark: In the terminology introduced at the very beginning of Chapter 1, LASSO regression is therefore a feature selection method.

[a] The main reference for this chapter is [4, Chapter 2].

Preliminaries

A natural way to obtain a sparse estimate $\tilde{\beta}$ of β is to penalize for the number of non-zero coefficients, that is to let $r = 0$ in (7.2).

However, computing $\hat{\beta}_\lambda^{(r)}$ for $r = 0$ is computationally hard, one reason being that for $r = 0$ the function

$$(\alpha, \beta) \mapsto \left\| y^0 - \alpha - \boldsymbol{X}^0 \beta \right\|_2^2 + 2\lambda \sum_{j=1}^{p} |\beta_j|^r \tag{7.3}$$

is non-convex and may have several local minima.

The LASSO estimator $(\tilde{\alpha}_\lambda, \tilde{\beta}_\lambda)$ of (α, β) is obtained by letting $r = 1$ in (7.2), that is $(\tilde{\alpha}_\lambda, \tilde{\beta}_\lambda) = (\hat{\alpha}_\lambda^{(1)}, \hat{\beta}_\lambda^{(1)})$.

For $r = 1$ the function defined in (7.3) is the sum of two convex functions, and is therefore convex. Consequently, all the local minima of this function are also global minima. In addition, as we will see below, the estimator $\tilde{\beta}_\lambda$ is (typically) sparse.

In this chapter, for $A \subseteq \{1, \ldots, p\}$ we let $\boldsymbol{X}_A$ be the $n \times |A|$ matrix having the vectors $\{x_{(j)}\}_{j \in A}$ as columns and, for $\beta \in \mathbb{R}^p$, we let β_A be the $|A|$-dimensional vector having elements $\{\beta_j\}_{j \in A}$.

Finally, we recall that for $z \in \mathbb{R}$ we have $\text{sign}(z) = 1$ if $z > 0$, $\text{sign}(z) = -1$ if $z < 0$ and $\text{sign}(z) = 0$ if $z = 0$. If $z \in \mathbb{R}^k$ for some integer $k > 1$ we abuse notation in what follows by using the shorthand $\text{sign}(z) = (\text{sign}(z_1), \ldots, \text{sign}(z_k))$.

The LASSO estimator

The following theorem characterizes the solution of the optimization problem (7.3) when $r = 1$.

Theorem 7.1 *Let $\lambda > 0$, $\tilde{\alpha} \in \mathbb{R}$ and $\tilde{\beta} \in \mathbb{R}^p$. Then, $(\tilde{\alpha}, \tilde{\beta})$ is a solution to (7.3) for $r = 1$ if and only if $\tilde{\alpha} = \bar{y}^0 - \tilde{\beta}^\top \bar{x}^0$ and if and only if, for all $j \in \{1, \dots, p\}$,*

$$\begin{aligned} x_{(j)}^\top (y - \boldsymbol{X}\tilde{\beta}) &= \lambda \operatorname{sign}(\tilde{\beta}_j) && \text{if } \tilde{\beta}_j \neq 0 \\ \left| x_{(j)}^\top (y - \boldsymbol{X}\tilde{\beta}) \right| &\leq \lambda && \text{if } \tilde{\beta}_j = 0. \end{aligned} \tag{7.4}$$

Moreover, letting $A = \{j \in \{1, \dots, p\} : \tilde{\beta}_j \neq 0\}$, if the matrix $(\boldsymbol{X}_A^\top \boldsymbol{X}_A)$ is invertible we have

$$\begin{aligned} \tilde{\beta}_A &= (\boldsymbol{X}_A \boldsymbol{X}_A^\top)^{-1} \Big(\boldsymbol{X}_A^\top y - \lambda \operatorname{sign}\big(\boldsymbol{X}_A^\top (y - \boldsymbol{X}\tilde{\beta}) \big) \Big) \\ &= (\boldsymbol{X}_A \boldsymbol{X}_A^\top)^{-1} \big(\boldsymbol{X}_A^\top y - \lambda \operatorname{sign}(\tilde{\beta}_A) \big). \end{aligned}$$

Remark: By Theorem 7.1, $\tilde{\beta}_\lambda = 0$ if $\lambda \geq \max_{j \in \{1,\dots,p\}} |x_{(j)}^\top y|$.

Proof of Theorem 7.1[a]*:* Following the computations done in the proof of Proposition 6.1, we have $\tilde{\alpha}_\lambda = \bar{y}^0 - \tilde{\beta}_\lambda^\top \bar{x}^0$ and

$$\tilde{\beta}_\lambda \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \|y - \boldsymbol{X}\beta\|_2^2 + 2\lambda \sum_{j=1}^p |\beta_j|.$$

We let

$$F_\lambda(\beta) = \frac{1}{2} \|y - \boldsymbol{X}\beta\|_2^2 + \lambda \sum_{j=1}^p |\beta_j|$$

and first show that if $\tilde{\beta} \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} F_\lambda(\beta)$ then (7.4) holds.

[a] A much shorter proof of this result, based on known necessary and sufficient subgradient conditions for finding the minimum of a non-differentiable convex function, can be found in [7].

Proof of Theorem 7.1 (end)

Remark that for all $j \in \{1, \dots, p\}$ and $\beta \in \mathbb{R}^p$ such that $\beta_j \neq 0$ we have

$$\frac{\partial}{\partial \beta_j} F_\lambda(\beta) = x_{(j)}^\top (\boldsymbol{X}\beta - y) + \lambda \operatorname{sign}(\beta_j). \tag{7.5}$$

We now let $j \in \{1, \dots, p\}$ be such that $\tilde{\beta}_j \neq 0$. Then, we must have

$$\frac{\partial}{\partial \beta_j} F_\lambda(\beta)\Big|_{\beta=\tilde{\beta}} = 0 \tag{7.6}$$

since otherwise by slightly increasing or decreasing $\tilde{\beta}_j$ we can reduce the value of $F_\lambda(\tilde{\beta})$, which would contradict the fact that $\tilde{\beta} \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} F_\lambda(\beta)$. Hence, using (7.5) it follows that if $\tilde{\beta}_j \neq 0$ then

$$x_{(j)}^\top (\boldsymbol{X}\tilde{\beta} - y) + \lambda \operatorname{sign}(\tilde{\beta}_j) = 0 \Leftrightarrow x_{(j)}^\top (y - \boldsymbol{X}\tilde{\beta}) = \lambda \operatorname{sign}(\tilde{\beta}_j).$$

We now let j be such that $\tilde{\beta}_j = 0$ and for every $\beta_j \neq 0$ let $\tilde{\beta}^{\beta_j}$ be the vector $\tilde{\beta}$ whose jth component has been replaced by β_j. Then,

- for all small enough $\beta_j > 0$ we must have $\frac{\partial}{\partial \beta_j} F_\lambda(\beta)\big|_{\beta=\tilde{\beta}^{\beta_j}} > 0$. By (7.5), this is equivalent to have $x_{(j)}^\top (y - \boldsymbol{X}\tilde{\beta}^{\beta_j}) < \lambda$ for all small enough $\beta_j > 0$.

- for all large enough $\beta_j < 0$ we must have $\frac{\partial}{\partial \beta_j} F_\lambda(\beta)\big|_{\beta=\tilde{\beta}^{\beta_j}} > 0$. By (7.5), this is equivalent to have $x_{(j)}^\top (y - \boldsymbol{X}\tilde{\beta}^{\beta_j}) > -\lambda$ for all large enough $\beta_j < 0$.

Therefore, for all β_j such that $|\beta_j|$ is small enough we have $|x_{(j)}^\top (y - \boldsymbol{X}\tilde{\beta}^{\beta_j})| < \lambda$ and therefore

$$|x_{(j)}^\top (y - \boldsymbol{X}\tilde{\beta})| = \lim_{\beta_j \to 0} |x_{(j)}^\top (y - \boldsymbol{X}\tilde{\beta}^{\beta_j})| \leq \lambda.$$

This shows that if $\tilde{\beta}_j = 0$ then $|x_{(j)}^\top (y - \boldsymbol{X}\tilde{\beta})| \leq \lambda$, which concludes to show that (7.4) holds if $\tilde{\beta} \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} F_\lambda(\beta)$.

We now let $\tilde{\beta} \in \mathbb{R}^p$ be such that (7.4) holds and show that $\tilde{\beta} \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} F_\lambda(\beta)$. Let

$$F_A(\beta) = \frac{1}{2} \sum_{i=1}^n \Big(y_i - \sum_{j \in A} x_{ij} \beta_j\Big)^2 + \lambda \sum_{j \in A} |\beta_j|, \quad \forall \beta \in \mathbb{R}^{|A|}.$$

Then, using (7.5) and under (7.4), we have $\frac{\partial}{\partial \beta} F_A(\beta)\big|_{\beta=\tilde{\beta}_A} = 0$ and since F_A is convex it follows that $\tilde{\beta}_A \in \operatorname{argmin}_{\beta \in \mathbb{R}^{|A|}} F_A(\beta)$.

Proof of Theorem 7.1 (end)

Consequently, for all $\beta \in \mathbb{R}^p$ we have

$$\begin{aligned} F_\lambda(\beta) - F_\lambda(\tilde{\beta}) &= F_A(\beta_A) - F_A(\tilde{\beta}_A) \\ &\quad + \frac{1}{2}\sum_{i=1}^n \Big(\sum_{j\notin A} \beta_j x_{ij}\Big)^2 - \sum_{j\notin A}\Big(\beta_j x_{(j)}^\top (y - \boldsymbol{X}\tilde{\beta}) - \lambda|\beta_j|\Big) \\ &\geq -\sum_{j\notin A}\Big(\beta_j x_{(j)}^\top (y - \boldsymbol{X}\tilde{\beta}) - \lambda|\beta_j|\Big) \end{aligned} \tag{7.7}$$

and we know show that the term after the inequality sign is non-negative under (7.4).

To this aim let $j \notin A$ and assume first that $\beta_j > 0$. In this case, we have

$$\beta_j x_{(j)}^\top (y - \boldsymbol{X}\tilde{\beta}) - \lambda|\beta_j| = \beta_j\big(x_{(j)}^\top (y - \boldsymbol{X}\tilde{\beta}) - \lambda\big) \tag{7.8}$$

Under (7.4) we have $|x_{(j)}^\top (y - \boldsymbol{X}\tilde{\beta})| \leq \lambda$ and thus $x_{(j)}^\top (y - \boldsymbol{X}\tilde{\beta}) \leq \lambda$. Together with (7.8), this shows that if $\beta_j > 0$ then $\beta_j x_{(j)}^\top (y - \boldsymbol{X}\tilde{\beta}) - \lambda|\beta_j| \leq 0$.

Assume now that $\beta_j < 0$ so that

$$\beta_j x_{(j)}^\top (y - \boldsymbol{X}\tilde{\beta}) - \lambda|\beta_j| = -|\beta_j|\big(x_{(j)}^\top (y - \boldsymbol{X}\tilde{\beta}) + \lambda\big). \tag{7.9}$$

Under (7.4) we have $|x_{(j)}^\top (y - \boldsymbol{X}\tilde{\beta})| \leq \lambda$ and thus $x_{(j)}^\top (y - \boldsymbol{X}\tilde{\beta}) \geq -\lambda$. Together with (7.9), this shows that if $\beta_j < 0$ then $\beta_j x_{(j)}^\top (y - \boldsymbol{X}\tilde{\beta}) - \lambda|\beta_j| \leq 0$.

Therefore, using (7.7) it follows that

$$F_\lambda(\beta) - F_\lambda(\tilde{\beta}) \geq -\sum_{j\notin A}\Big(\beta_j x_{(j)}^\top (y - \boldsymbol{X}\tilde{\beta}) - \lambda|\beta_j|\Big) \geq 0, \quad \forall \beta \in \mathbb{R}^p$$

showing that $\tilde{\beta} \in \operatorname{argmin}_{\beta\in\mathbb{R}^p} F_\lambda(\beta)$. This concludes to show that (7.4) holds if and only if $\tilde{\beta} \in \operatorname{argmin}_{\beta\in\mathbb{R}^p} F_\lambda(\beta)$.

To prove the last part of the theorem note that for all $j \in A$ we have $x_{(j)}^\top (y - \boldsymbol{X}\tilde{\beta}) = \lambda \operatorname{sign}(\tilde{\beta}_j)$, and thus $\operatorname{sign}\big(x_{(j)}^\top (y - \boldsymbol{X}\tilde{\beta})\big) = \operatorname{sign}(\tilde{\beta}_j)$ for all $j \in A$. Hence,

$$\lambda \operatorname{sign}(\tilde{\beta}_A) = \boldsymbol{X}_A^\top (y - \boldsymbol{X}\tilde{\beta}) = \boldsymbol{X}_A^\top (y - \boldsymbol{X}_A\tilde{\beta}_A) = \lambda \operatorname{sign}\Big(\boldsymbol{X}_A^\top (y - \boldsymbol{X}\tilde{\beta})\Big)$$

and the proof is complete. □

Computation of $\tilde{\beta}_\lambda$: A first preliminary result

For every $\lambda > 0$ let $A_\lambda \subseteq \{1, \ldots, p\}$ be such that $\tilde{\beta}_{\lambda,j} \neq 0$ if and only if $j \in A_\lambda$.

Proposition 7.1 *Let $\lambda_0 > 0$, $\lambda_1 = \sup\{\lambda \in [0, \lambda_0] : A_\lambda \neq A_{\lambda_0}\}$ and assume that the matrix $\boldsymbol{X}_{A_{\lambda_0}}^\top \boldsymbol{X}_{A_{\lambda_0}}$ is invertible. Then, for all $\lambda \in (\lambda_1, \lambda_0]$ we have*

$$\tilde{\beta}_{A_\lambda} = (\boldsymbol{X}_{A_{\lambda_0}}^\top \boldsymbol{X}_{A_{\lambda_0}})^{-1}\Big(\boldsymbol{X}_{A_{\lambda_0}} y - \lambda \operatorname{sign}(\tilde{\beta}_{A_{\lambda_0}})\Big).$$

Proof: Remark that $\operatorname{sign}(\tilde{\beta}_{A_\lambda}) = \operatorname{sign}(\tilde{\beta}_{A_{\lambda_0}})$ for all $\lambda \in (\lambda_1, \lambda_0]$. Then, by Theorem 7.1, for all $\lambda \in (\lambda_1, \lambda_0]$ we have

$$\begin{aligned}\tilde{\beta}_{A_\lambda} &= (\boldsymbol{X}_{A_\lambda}^\top \boldsymbol{X}_{A_\lambda})^{-1}\Big(\boldsymbol{X}_{A_\lambda} y - \lambda \operatorname{sign}(\tilde{\beta}_{A_\lambda})\Big) \\ &= (\boldsymbol{X}_{A_{\lambda_0}}^\top \boldsymbol{X}_{A_{\lambda_0}})^{-1}\Big(\boldsymbol{X}_{A_{\lambda_0}} y - \lambda \operatorname{sign}(\tilde{\beta}_{A_{\lambda_0}})\Big)\end{aligned}$$

and the proof is complete. □

Remark: Proposition 7.1 shows that $\tilde{\beta}_{A_\lambda}$ (and thus $\tilde{\beta}_\lambda$) is a linear function of λ as long as the set A_λ remains unchanged.

Computation of $\tilde{\beta}_\lambda$: A second preliminary result

Proposition 7.2 *Let $\lambda_0 > 0$, $\lambda_1 = \sup\{\lambda \in [0, \lambda_0] : A_\lambda \neq A_{\lambda_0}\}$ and assume that the matrix $\boldsymbol{X}_{A_{\lambda_0}}^\top \boldsymbol{X}_{A_{\lambda_0}}$ is invertible. Then,*

$$\lambda_1 = \max\left\{0, \lambda_1^{(1)}, \lambda_1^{(2)}\right\}$$

where

$$\lambda_1^{(1)} = \max_{j \notin A_{\lambda_0}} \left\{ \max\left\{ -\frac{w_j^0}{\tilde{w}_j^0 + 1}, -\frac{w_j^0}{\tilde{w}_j^0 - 1} \right\} \right\}, \quad \lambda_1^{(2)} = \max_{j \in \{1,\dots,|A_{\lambda_0}|\}} \left\{ \frac{\tilde{v}_j^0}{v_j^0} \right\}$$

with

$$\begin{aligned}
\tilde{v}^0 &= (\boldsymbol{X}_{A_{\lambda_0}}^\top \boldsymbol{X}_{A_{\lambda_0}})^{-1} \boldsymbol{X}_{A_{\lambda_0}} y \\
v^0 &= (\boldsymbol{X}_{A_{\lambda_0}}^\top \boldsymbol{X}_{A_{\lambda_0}})^{-1} \operatorname{sign}(\tilde{\beta}_{A_{\lambda_0}}) \\
w_j^0 &= x_{(j)}^\top (y - \boldsymbol{X}_{A_{\lambda_0}} \tilde{v}^0), && \forall j \notin A_{\lambda_0} \\
\tilde{w}_j^0 &= x_{(j)}^\top \boldsymbol{X}_{A_{\lambda_0}} v^0, && \forall j \notin A_{\lambda_0}.
\end{aligned}$$

Moreover,

$$A_{\lambda_1} = \begin{cases}
0, & \lambda_1 = 0 \\
A_{\lambda_0} \cup \left\{ \operatorname{argmax}_{j \notin A_{\lambda_0}} \left\{ \max\left\{ -\frac{\tilde{w}_j^0}{w_j^0 + 1}, -\frac{\tilde{w}_j^0}{w_j^0 - 1} \right\} \right\} \right\}, & \lambda_1 = \lambda_1^{(1)} \\
A_{\lambda_0} \setminus \left\{ j \in A_{\lambda_0} : \frac{\tilde{v}_j^0}{v_j^0} = \max_{k \in \{1,\dots,|A_{\lambda_0}|\}} \left\{ \frac{\tilde{v}_k^0}{v_k^0} \right\}, \right. & \lambda_1 = \lambda_1^{(2)}.
\end{cases}$$

Proof of Proposition 7.2

We first remark that we have either $A_{\lambda_0} \subsetneq A_{\lambda_1}$ (active set addition) or $A_{\lambda_1} \subsetneq A_{\lambda_0}$ (active set deletion).

We first compute λ_1 assuming that $A_{\lambda_0} \subsetneq A_{\lambda_1}$. In this case, by Theorem 7.1 and noting that, by Proposition 7.1, $\hat{\beta}_{A_\lambda} = \tilde{v}^0 - \lambda v^0$ for all $\lambda \in (\lambda_1, \lambda_0)$, we have

$$\begin{aligned}
\lambda_1 &= \sup\Big\{\lambda \in [0, \lambda_0] : \ |x_{(j)}^\top (y - \boldsymbol{X}_{A_{\lambda_0}} \tilde{\beta}_{A_\lambda})| \geq \lambda \text{ for a } j \notin A_{\lambda_0}\Big\} \\
&= \sup\Big\{\lambda \in [0, \lambda_0] : \ x_{(j)}^\top (y - \boldsymbol{X}_{A_{\lambda_0}} \tilde{\beta}_{A_\lambda}) \pm \lambda \text{ for a } j \notin A_{\lambda_0}\Big\} \\
&= \sup\Big\{\lambda \in [0, \lambda_0] : \ x_{(j)}^\top (y - \boldsymbol{X}_{A_{\lambda_0}} (\tilde{v}^0 - \lambda v^0)) \pm \lambda \text{ for a } j \notin A_{\lambda_0}\Big\} \\
&= \sup\Big\{\lambda \in [0, \lambda_0) : \ w_j^0 + \lambda(\tilde{w}_j^0 \pm 1) = 0 \text{ for a } j \notin A_{\lambda_0}\Big\} \\
&= \max\Big\{0, \max_{j \notin A_{\lambda_0}} \Big\{ - \frac{w_j^0}{\tilde{w}_j^0 + 1}\Big\}, \max_{j \notin A_{\lambda_0}} \Big\{ - \frac{w_j^0}{\tilde{w}_j^0 - 1}\Big\}\Big\}.
\end{aligned}$$

We now compute λ_1 assuming that $A_{\lambda_1} \subsetneq A_{\lambda_0}$. Since $\hat{\beta}_{A_\lambda} = \tilde{v}^0 - \lambda v^0$ as long as $A_\lambda = A_{\lambda_0}$, as λ decreases from λ_0 an element will be removed from A_{λ_0} as soon λ is such that $\tilde{v}_j^0 - \lambda v_j^0 = 0$ for at least one $j \in A_{\lambda_0}$. Therefore, if as λ decreases from λ_0 an active set deletion occurs before and active set addition we have

$$\lambda_1 = 0 \vee \max_{j \in \{1, \ldots, |A_{\lambda_0}|\}} \Big\{\frac{\tilde{v}_j^0}{v_j^0}\Big\}.$$

The proof is complete. □

Computation of $\tilde{\beta}_\lambda$: An algorithm

Theorem 7.1 and Propositions 7.1-7.2 lead to the following algorithm for computing, for all $\lambda > 0$, a solution $\tilde{\beta}_\lambda$ to optimization problem (7.3) when $r = 1$.

LASSO path algorithm

(i) Let j_0 be such that $|x_{(j_0)}^\top y| = \max_{j\in\{1,\dots,p\}} |x_{(j)}^\top y|$ and $\bar{\lambda} = |x_{(j_0)}^\top y|$.

(ii) Let $\lambda = \bar{\lambda}$, $\lambda_0 = \bar{\lambda}$ and $\tilde{\beta}_\lambda$ be such that

$$\tilde{\beta}_{A_{\lambda_0}} = (x_{(j_0)}^\top x_{(j_0)})^{-1}\big(x_{(j_0)}^\top y - \lambda_0 \operatorname{sign}\big(\tilde{\beta}_{A_{\lambda_0}}\big)\big).$$

while $\lambda > 0$ **do**

(iii) Let λ_1 and A_{λ_1} be as in Proposition 7.2.

(iv) For all $\lambda \in (\lambda_1, \lambda_0]$ let $\tilde{\beta}_\lambda$ be such that

$$\tilde{\beta}_{A_\lambda} = (\boldsymbol{X}_{A_{\lambda_0}}^\top \boldsymbol{X}_{A_{\lambda_0}})^{-1}\big(\boldsymbol{X}_{A_{\lambda_0}} y - \lambda \operatorname{sign}(\tilde{\beta}_{A_{\lambda_0}})\big)$$

if $|A_{\lambda_1}| > \min(p, n)$ **then**

(v) Break

else

(vi) Let $\tilde{\beta}_\lambda$ be such that

$$\tilde{\beta}_{A_{\lambda_1}} = (\boldsymbol{X}_{A_{\lambda_1}}^\top \boldsymbol{X}_{A_{\lambda_1}})^{-1}\big(\boldsymbol{X}_{A_{\lambda_1}} y - \lambda_1 \operatorname{sign}\big(\tilde{\beta}_{A_{\lambda_1}}\big)\big).$$

(vii) Let $\lambda_0 = \lambda_1$ and $\lambda = \lambda_1$.

end if

end while

Return: $(\lambda, \tilde{\beta}_\lambda)_{\lambda\in(\lambda_1,\bar{\lambda}]}$

Practical comments

1. The above algorithm assumes that the matrix $(\boldsymbol{X}_A^\top \boldsymbol{X}_A)$ is invertible for any $A \subseteq \{1, \ldots, n\}$ such that $|A| \leq \min(p, n)$.

2. The above algorithm requires to compute $(\boldsymbol{X}_{A_{\lambda_1}}^\top \boldsymbol{X}_{A_{\lambda_1}})^{-1}$ for all values of λ_1. Using the fact that the matrix is $\boldsymbol{X}_{A_{\lambda_1}}$ is obtained from the matrix $\boldsymbol{X}_{A_{\lambda_0}}$ either by adding or by removing one column, it is possible to compute $(\boldsymbol{X}_{A_{\lambda_1}}^\top \boldsymbol{X}_{A_{\lambda_1}})^{-1}$ from $(\boldsymbol{X}_{A_{\lambda_0}}^\top \boldsymbol{X}_{A_{\lambda_0}})^{-1}$ in $\mathcal{O}(|A_{\lambda_1}|^2)$ operations. By doing this, the complexity of computing the LASSO path is $\mathcal{O}(I^3)$, where I is the number of different values for λ_1 computed in the above algorithm. In practice, it is often the case that $I = \min(n, p)$ but the worst case complexity of the LASSO path algorithm is exponential in p [7].

   $\Longrightarrow$ The exact computation of $\tilde{\beta}_\lambda$ can be expensive when p and n is large and, in this case, $\tilde{\beta}_\lambda$ is often approximated using a coordinate descent algorithm (see below).

3. Due to the finite precision of computers, in practice the value $\lambda_1^{(2)}$ defined in Proposition 7.2, and the set A_{λ_1} when $\lambda_1 = \lambda_1^{(2)}$, are often computed by excluding the last variable added in the path. (This is to avoid that, due to numerical errors, a variable is removed from the active set just after it has been added.)

4. As in ridge regression, and for the same reason, the LASSO estimator is usually computed from the standardized variables $x_{(1)}, \ldots, x_{(p)}$. The resulting estimate of the original model parameter is then obtained as explained in Chapter 6 (see page 113).

The coordinate descent algorithm

Let $f : \mathbb{R}^p \to \mathbb{R}$ be a convex function and $x^* = \text{argmin}_{x \in \mathbb{R}^p} f(x)$.

Then, assuming $p \geq 2$, the coordinate descent algorithm for computing x^* is as follows:

Coordinate descent

Input: Starting value $x_0 \in \mathbb{R}^p$.

for $k \geq 1$ **do**

for $j = 1, \ldots, p$ **do**

(i) With obvious convention when $j \in \{1, p\}$, let

$$x_{k,j} = \underset{x_j \in \mathbb{R}}{\text{argmin}}\, f(x_{k,1}, \ldots, x_{k,j-1}, x_j, x_{k-1,j+1}, \ldots, x_{k-1,p})$$

end for

(ii): Set $x_k = (x_{k,1}, \ldots, x_{k,p})$

end for

Remark that each update of the approximation x of x^* performed at step (i) either leaves $f(x)$ unchanged or increases $f(x)$.

Under suitable conditions on f[a] we have $x_k \to x^*$ as $k \to \infty$ (see [4], Section 5.4).

Remark: Instead of updating one component of x at a time we can of course update several of them at the same time.

[a] e.g. f is strictly convex in each of its components.

LASSO with coordinate descent

We apply coordinate descent to the function $F_\lambda \mathbb{R}^p \to \mathbb{R}$ defined by

$$F_\lambda(\beta) = \|y - \boldsymbol{X}\beta\|_2^2 + 2\lambda \sum_{j=1}^{p} |\beta_j|, \quad \beta \in \mathbb{R}^p$$

which, for $j \in \{1, \dots, p\}$ and $b \in \mathbb{R}^p$, requires to compute (with obvious convention when $j \in \{1, p\}$)

$$\beta_j^{(b)} \in \underset{\beta_j \in \mathbb{R}}{\operatorname{argmin}} F_\lambda(b_1, \dots, b_{j-1}, \beta_j, b_{j+1}, \dots, b_p). \tag{7.10}$$

To this aim, for every $c \in \mathbb{R}$ we let $\mathcal{S}_c(z) = \operatorname{sign}(z)\big(|z| - c\big)_+$ be the soft-thresholding operator, recalling that for $z \in \mathbb{R}$ we have $z_+ = \max(0, z)$.

Using Theorem 7.1 we readily obtain the following result for the expression of $\beta_j^{(b)}$ defined in (7.10).

Theorem 7.2 *Assume that the variables $x_{(1)}, \dots, x_{(p)}$ are normalized, so that $\sum_{i=1}^n x_{ij}^2 = 1$ for all $j \in \{1, \dots, p\}$. Let $j \in \{1, \dots, p\}$, $b \in \mathbb{R}^p$ and $\beta_j^{(b)}$ be as defined in (7.10). Then,*

$$\beta_j^{(b)} = \mathcal{S}_\lambda\bigg(\sum_{i=1}^{n} x_{ij}\big(y_i - \sum_{m \neq j} b_m x_{im}\big) \bigg).$$

Proof of Theorem 7.2

Let $y_i^{(b)} = y_i - \sum_{m \neq j} b_m x_{im}$ for all $i \in \{1, \ldots, n\}$ and assume first that $\beta_j^{(b)} \neq 0$. Then, applying Theorem 7.1 with $p = 1$, y replaced by $y^{(b)}$ and $\boldsymbol{X}$ replaced by $x_{(j)}$, it follows that $\beta_j^{(b)}$ solves

$$\sum_{i=1}^{n} x_{ij} \Big(y_i - \sum_{m \neq j} b_m x_{im} - \beta_j^{(b)} x_{ij} \Big) - \lambda \, \text{sign}(\beta_j^{(b)}) = 0$$

and thus, recalling that the variables $x_{(1)}, \ldots, x_{(p)}$ are assumed to be normalized

$$\begin{aligned} \beta_j^{(b)} &= \text{sign} \Big(\sum_{i=1}^{n} x_{ij} \big(y_i - \sum_{m \neq j} b_m x_{im} \big) \Big) \Big(\Big| \sum_{i=1}^{n} x_{ij} \big(y_i - \sum_{m \neq j} b_m x_{im} \big) \Big| - \lambda \Big) \\ &= \mathcal{S}_\lambda \Big(\sum_{i=1}^{n} x_{ij} \big(y_i - \sum_{m \neq j} b_m x_{im} \big) \Big). \end{aligned}$$

Assume now that $\beta_j^{(b)} = 0$. Then, applying again Theorem 7.1 with $p = 1$, y replaced by $y^{(b)}$ and $\boldsymbol{X}$ replaced by $x_{(j)}$, we must have

$$\sum_{i=1}^{n} x_{ij} \Big(y_i - \sum_{m \neq j} b_m x_{im} \Big) - \lambda \leq 0. \tag{7.11}$$

We remark that (7.11) holds if and only if

$$\mathcal{S}_\lambda \Big(\sum_{i=1}^{n} x_{ij} \big(y_i - \sum_{m \neq j} b_m x_{im} \big) \Big) = 0$$

and thus when $\beta_j^{(b)} = 0$ we also have

$$\beta_j^{(b)} = \mathcal{S}_\lambda \Big(\sum_{i=1}^{n} x_{ij} \big(y_i - \sum_{m \neq j} b_m x_{im} \big) \Big).$$

The proof is complete. □

Computing $\tilde{\beta}_\lambda$ using coordinate descent

Using Theorem 7.2 we obtain the following coordinate descent algorithm for computing, for a given $\lambda > 0$, a solution $\tilde{\beta}_\lambda$ of the optimization problem (7.3) when $r = 1$.

Coordinate descent algorithm for computing $\tilde{\beta}_\lambda$

Input: Starting value $\beta_0 \in \mathbb{R}^p$.
(i) Let $r_i = y_i - \sum_{j=2}^p \beta_{0,j} x_{ij}$ for $i = 1, \ldots, n$
for $k \geq 1$ **do**
(ii): Let $\beta_{k,1} = \mathcal{S}_\lambda\big(\sum_{i=1}^n x_{i1} r_i\big)$
for $j = 2, \ldots, p$ **do**
(iii) $r_i \leftarrow r_i - \beta_{k,j-1} x_{i(j-1)} + \beta_{k-1,j} x_{ij}$ for $i = 1, \ldots, n$
(iv) Let $\beta_{k,j} = \mathcal{S}_\lambda\big(\sum_{i=1}^n x_{ij} r_i\big)$
end for
(v) $r_i \leftarrow r_i - \beta_{k,p} x_{ip} + \beta_{k,1} x_{i1}$ for $i = 1, \ldots, n$
(vi) Set $\beta_k = (\beta_{k,1}, \ldots, \beta_{k,p})$.
if Convergence=TRUE **then**
(vii) **return** β_k.
(viii) **break**
end if
end for

Remark: It can be shown that the above algorithm is indeed valid, in the sense that for a $\tilde{\beta}_\lambda \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|y - \boldsymbol{X}\beta\|_2^2 + 2\lambda \sum_{j=1}^p |\beta_j|$ we have $\beta_k \to \tilde{\beta}_\lambda$ as $k \to \infty$ (see [4], Section 5.4).

Remark: Each iteration of the algorithm (i.e. computing β_k from β_{k+1}) costs only $\mathcal{O}(p\,n)$ operations, making is suitable for large p and/or large n problems.

Computing the LASSO path with coordinate descent

LASSO path Coordinate descent

Input: Starting value $\beta \in \mathbb{R}^p$, integer $M \in \mathbb{N}$ and $\epsilon \in (0,1)$.

(i) Let $\lambda_1 = \max_{j\in\{1,\dots,p\}} |x_{(j)}^\top y|$

(ii) Compute an approximation β'_{λ_1} of $\tilde{\beta}_{\lambda_1}$ using coordinate descent with starting value $\beta_0 = \beta$.

for $m = 2, \dots M$ **do**

(iii): Let $\lambda_m = \exp\Big(\log \lambda_{m-1} - \log(-\epsilon)/(M-1)\Big)$.

(iv) Compute an approximation β'_{λ_m} of $\tilde{\beta}_{\lambda_m}$ using coordinate descent with starting value $\beta_0 = \beta'_{\lambda_{m-1}}$.

end for

Return: $\{\lambda_m, \beta'_{\lambda_m}\}_{m=1}^M$

In the above algorithm λ decreases from λ_1 to $\lambda_M = \epsilon\lambda_1$ linearly on a log scale (other choices for the sequence $\{\lambda_m\}_{m=1}^M$ are of course possible).

The tuning parameter ϵ is usually a small number (for instance $\epsilon = 0.0001$ if $n > p$ and $\epsilon = 0.01$ if $n < p$) while M is a large number (e.g. $M = 100$).

Remark: The definition of $\{\lambda_m\}_{m=1}^M$ used in the above algorithm, as well as the aforementioned proposed default values for ϵ and M, are as in the `R` package `glmnet`.

Choice of the parameter λ

K-fold cross-validation is often use in LASSO to choose λ.

In this case, we randomly divide the data sets into $K > 1$ groups of equal size, where typically $K = 5$ or $K = 10$.

Then, we successively treat each group as the test set and compute the (approximate) LASSO path using the observations from the remaining $K - 1$ groups. For each λ we record the mean squared prediction error on the test set (i.e. on the group excluded from the estimation step).

At the end of the process we obtain K estimates of the prediction error for a range of λ, which are averaged to produce the cross validation curve. We finally retain the value of λ at which this curve reach its minimum.

Remark: K-fold cross-validation with $K = n$ reduces to the leave-one-out cross validation procedure discussed in Chapter 6.

Remark: We saw in Chapter 6 that, for ridge regression, cross-validation can be implemented in such a way that only one ridge regression needs to be performed for each value of the penalty parameter $\lambda > 0$[a]. Such a simplification of the cross-validation procedure does not exist for LASSO, and therefore for each value of λ it is necessary to compute K LASSO estimators.

[a] We saw this result for $K = n$ but it generalizes to any K.

Bayesian perspective of LASSO

Consider the following Bayesian linear regression model

$$\beta_j \overset{\text{iid}}{\sim} \text{Laplace}\big(0, 2\sigma^2/\lambda\big), \quad Y_i \sim \mathcal{N}_1(x_i^\top \beta, \sigma^2), \quad i = 1, \dots, n$$

for some $\sigma^2 > 0$.

Recalling that the posterior distribution of β given the observations y is $\pi(\beta|y) \propto \pi(y|\beta)\pi(\beta)$, we have

$$\log \pi(\beta|y) = c - \frac{1}{2\sigma^2}\|y - \boldsymbol{X}\beta\|^2 - \frac{\lambda}{2\sigma^2}\sum_{j=1}^{p}|\beta_j|$$

for some constant $c \in \mathbb{R}$ (i.e. c is independent of β).

Therefore, the posterior mode β_{mode} of β satisfies

$$\begin{aligned}
\beta_{\text{mode}} &\in \operatorname*{argmax}_{\beta\in\mathbb{R}^p} \pi(\beta|y) \\
&= \operatorname*{argmax}_{\beta\in\mathbb{R}^p} \log \pi(\beta|y) \\
&= \operatorname*{argmax}_{\beta\in\mathbb{R}^p} -\|y - \boldsymbol{X}\beta\|^2 - \lambda\sum_{j=1}^{p}|\beta_j| \\
&= \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \|y - \boldsymbol{X}\beta\|^2 + \lambda\sum_{j=1}^{p}|\beta_j|.
\end{aligned}$$

Consequently, while the posterior mode of β in the Bayesian linear regression model with a Gaussian prior is the ridge regression estimator, it is the LASSO estimator when a Laplace prior is used.

Remark: Unlike for ridge regression, the posterior distribution for β in the Bayesian model associated to LASSO regression is not tractable.

Example: The prostate dataset[a]

The objective of this example is to examine the impact on the level of a prostate specific antigen of $p = 8$ clinical measures in $n = 67$ men who were about to receive a radical prostatectomy.

The variables $\{x_{(j)}\}_{j=1}^{p}$ are centred before estimating the regression parameter β and K-fold cross-validation with $K = 10$ is used to choose λ.

From Figure 7.1 we observe the coordinate descent provides a very good estimate of the lasso path $\{\tilde{\beta}_\lambda\}_{\lambda>0}$. In addition, the value of λ chosen by cross-validation is $\lambda \approx 0.89$, in which case one variable is not selected (i.e. one component of $\tilde{\beta}_\lambda$ is zero).

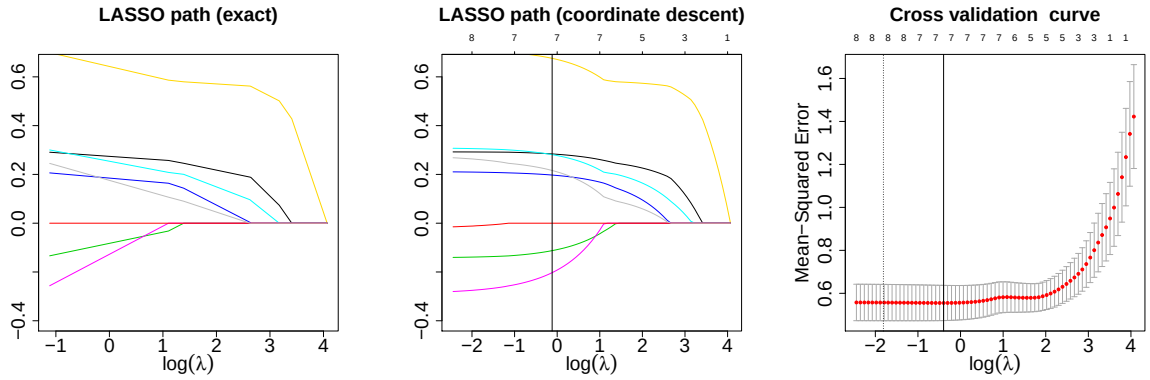

Figure 7.1: LASSO path for the prostate cancer dataset and cross-validation plot. In the middle and right plot the (black) vertical line represents the value of λ which minimizes the cross-validation error.

[a]This example is taken from [4] and the dataset is available at `https://hastie.su.domains/ElemStatLearn/`

Ridge v.s. LASSO regression: The prostate dataset

Figure 7.2 below compares the LASSO path and ridge path for the prostate cancer dataset.

As expected, and unlike in LASSO regression, we observe that none of the component of the ridge regression estimate $\hat{\beta}_\lambda$ is shrink to zero when λ is large enough. It is also interesting to see that the ordering of the $\{\tilde{\beta}_{\lambda,j}\}_{j=1}^p$ is similar to that of $\{\hat{\beta}_{\lambda,j}\}_{j=1}^p$ and thus that, on this example, LASSO shrinks to zero the elements of $\{\hat{\beta}_{\lambda,j}\}_{j=1}^p$ which are close to zero.

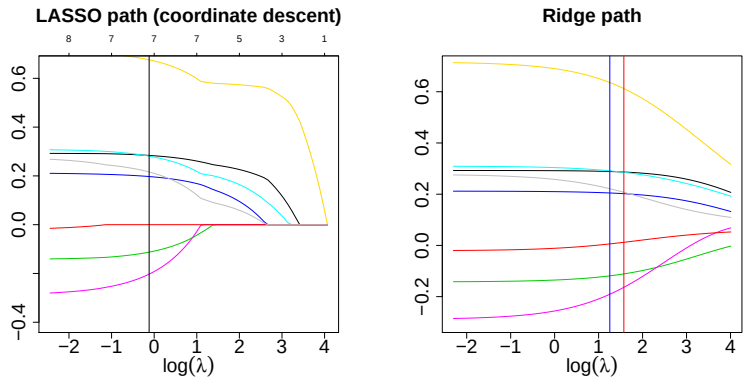

Figure 7.2: Ridge and LASSO paths for the prostate cancer dataset. For ridge regression, the horizontal blue line shows the optimal λ according to the OCV criterion and the red line the optimal λ according to the GCV criterion.

Remark: This example also illustrates the fact that, in ridge regression, ordinary cross-validation and generalized cross validation may lead to different choices of λ.

References

[1] Bishop, C. M. (2006). Pattern recognition. *Machine learning*, 128(9).

[2] Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, 36(3):287–314.

[3] Friedman, J., Hastie, T., Tibshirani, R., et al. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

[4] Hastie, T., Tibshirani, R., and Wainwright, M. (2019). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

[5] Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430.

[6] Inaba, M., Katoh, N., and Imai, H. (1994). Applications of weighted voronoi diagrams and randomization to variance-based k-clustering. In *Proceedings of the tenth annual symposium on Computational geometry*, pages 332–339.

[7] Mairal, J. and Yu, B. (2012). Complexity analysis of the lasso regularization path. *arXiv preprint arXiv:1205.0079*.

[8] Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate analysis*. Probability and mathematical statistics. Academic Press Inc.

[9] Ng, A., Jordan, M., and Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14.

[10] Paulsen, V. I. and Raghupathi, M. (2016). *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge Studies in Advanced Mathematics. Cambridge University Press.

[11] van Wieringen, W. N. (2015). Lecture notes on ridge regression. *arXiv preprint arXiv:1509.09169*.

[12] Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

[13] Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.