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].

We consider below the fixed design setting, in which the covariates $\{x_i^0\}_{i=1}^n$ are fixed (i.e. non-random).

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

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

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

Letting

$$Y^0 = (Y_1^0, \dots, Y_p^0), \quad Y = Y^0 - \frac{1}{n} \sum_{i=1}^n Y_i^0,$$

the corresponding OLS estimate $\hat{\mu}$ of $\mathbb{E}[Y]$ is given by

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

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.

---

[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 $y_i^0 + c$.

Some properties of the estimator $\hat{\mu}$ under the model (6.1)

Under the model (6.1) the estimator[a] $\hat{\mu}$ is unbiased, i.e. $\mathbb{E}[\hat{\mu}] = \mathbb{E}[Y]$.

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

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

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

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

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

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 $p = n$[b].

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

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

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 $\{Y_i\}_{i=1}^n$.

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

[a] 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.

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

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. $\square$

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 $\boldsymbol{D} = \operatorname{diag}(s_1^2, \dots, s_p^2)$, $\tilde{\boldsymbol{X}}^0 = \boldsymbol{X}^0\boldsymbol{D}^{-1/2}$ 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 $\{\tilde{x}^0_{(j)}\}_{j=1}^p$. 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 $F(\alpha,\beta)=\sum_{i=1}^n \big(y_i^0-\alpha-\beta^\top x_i^0)^2+\lambda\|\beta\|_2^2$. 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. $\square$

$\hat{\beta}_\lambda$ 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 two propositions.

Proposition 6.2 *Assume that $\boldsymbol{X}^\top y \neq 0$. Then, the ridge estimate of β is such that have $\|\hat{\beta}_\lambda\| < \|\beta_{\lambda_0}\|$ for all $\lambda > \lambda_0 > 0$.*

Proof: Remark that for every $\lambda \geq 0$ we have

$$\hat{\beta}_\lambda = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top y \tag{6.5}$$

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

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

$$\|\hat{\beta}_{\lambda_0}\|^2 - \|\hat{\beta}_\lambda\|^2 = (\boldsymbol{X}^\top y)^\top (\boldsymbol{B}_{\lambda_0} - \boldsymbol{B}_\lambda) \boldsymbol{X}^\top y$$

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

Since the matrices $\boldsymbol{B}_{\lambda_0}$ and $\boldsymbol{B}_\lambda$ are invertible (see Lemma 6.1) we have

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

Therefore, noting that

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

the result follows from the fact that $\lambda > \lambda_0$ and the fact that the matrix $\boldsymbol{X}\boldsymbol{X}^\top$ is positive semi-definite. $\square$

Proposition 6.3 *Assume that* (6.1) *holds for some β such that $\boldsymbol{X}\beta \neq 0$. Then, $\|\mathbb{E}[\hat{\beta}_\lambda]\| < \|\mathbb{E}[\beta_{\lambda_0}]\|$ for all $\lambda > \lambda_0 > 0$. If in addition $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible then $\|\mathbb{E}[\hat{\beta}_\lambda]\| < \|\beta\|$ of for all $\lambda > 0$.*

Proof: The result follows from similar computations as in the proof of Proposition 6.3.

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

Proposition 6.3 implies that, 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 6.4 *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. $\square$

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

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

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*

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

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

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

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* $i \in \{1, \dots, n\}$. *Then,*

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

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. $\square$

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

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

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

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

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),

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

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

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

where $\beta_{*,j}$ is a random draw from the $\mathcal{U}(0,1)$ distribution for $j = 1, \dots, 10$ 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 $\mathbb{E}[Y^0]$.

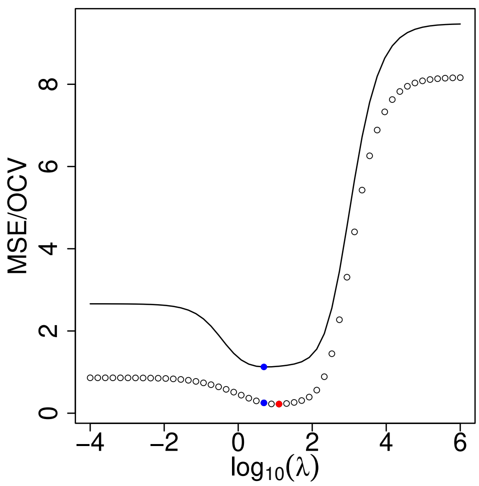

Figure 6.1: The dots show the mapping $\lambda \mapsto \mathrm{MSE}(\lambda) := \frac{1}{n}\|\hat{\mu}_\lambda - \boldsymbol{X}^0\beta^*\|^2$ and the solid line the mapping $\lambda \mapsto \mathrm{OCV}_{\mathrm{ridge}}(\lambda)$. The red dot is for $\lambda^* = \mathrm{argmin}_\lambda \mathrm{MSE}(\lambda)$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