7.7 Variable Selection, Regularization, Ridging and the Lasso

7.7.1 Introduction

The following considerations and methods apply to both classification and regression in principle. For several reasons however we will focus on *regression* in this whole section.

In problems with many potential regressor or explanatory variables which may partly be highly correlated with each other, the classical least-squares regression approach can suffer from the fact that estimated regression coefficients can become quite ill-determined, i.e. have a high variance even when the fitted regression surface may be well determined. Also, in chapter 1's model equation (1.2) $Y_i = \mathbf{x}_i{}^\intercal \boldsymbol{\beta} + \varepsilon_i$, $(i = 1, \ldots, n)$, equivalently

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{7.4}$$

where $\boldsymbol{\beta}, \mathbf{x}_i \in \mathbb{R}^p$, and $\mathbf{X}$ is the $n \times p$ matrix of columns $\mathbf{x}_1$,$\mathbf{x}_2$,...,$\mathbf{x}_p$, we had assumed that the number of observations n was larger than the number of parameters p, i.e., $n > p$, in order e.g, for the matrix $X^\intercal X$ to be of full rank p.

One main motivation this topic is the advent of data where there are many more potential explanatory variables than observations, in short, where $\boxed{p \gg n}$. Today the most important class of such applications is the analysis of genomic data, where e.g., for so called microarrays, one gets gene expression levels for thousands of genes (p = 1000–50'000) for typically a few dozen cells (n = 20–100). The target variable Y here is often binary (e.g., “normal” vs “special”) which means a classification problem with $J = 2$ classes; we assume here however that treating it as a regression problem (e.g., using least squares or L_2 error) is useful.

7.7.2 Ridge Regression

Let's consider the regression model (7.4), or

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p} + \varepsilon_i, \quad i = 1, \ldots, n.$$

For least squares, one can see that $\hat{\beta}_0' = \overline{Y}(= 1/n \sum_{i=1}^n Y_i)$ when the model is re-written as

$$Y_i = \beta_0' + \beta_1(x_{i,1} - \overline{x_{\cdot,1}}) + \ldots + \beta_p(x_{i,p} - \overline{x_{\cdot,p}}) + \varepsilon_i.$$

Hence, in all of the following, we will assume that all the variables ($Y_\cdot$, and $x_{\cdot,j} =: \mathbf{x}^{(j)}$) have been centered, i.e., mean subtracted, such that no intercept is needed, and we use the model

$$Y_i = \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p} + \varepsilon_i, \quad i = 1, \ldots, n. \tag{7.5}$$

Let's now assume that the variables (column vectors) $\mathbf{x}^{(j)} = (x_{i,j})_{i=1}^n$ and $\mathbf{x}^{(k)}$ are highly correlated (positively or negatively), or equivalently, $\mathbf{x}^{(k)} \approx c + c'\mathbf{x}^{(j)}$ (for $c' \neq 0$) and $\mathbf{x}^{(j)} \approx -c/c' + 1/c'\mathbf{x}^{(k)}$. As a consequence, the following part of the model,

$$\begin{aligned}
\beta_0 + \beta_j \mathbf{x}^{(j)} + \beta_k \mathbf{x}^{(k)} &\approx \beta_1 + \beta_j \mathbf{x}^{(j)} + \beta_k \left(c + c'\mathbf{x}^{(j)}\right) \\
&= \underbrace{\beta_1 + c\beta_k}_{\beta_1'} + (\beta_j + \beta_k c')\mathbf{x}^{(j)} \\
&\approx \underbrace{\beta_1 + \beta_j/c'}_{\beta_1''} + (\beta_k + \beta_j/c')\mathbf{x}^{(k)}.
\end{aligned}$$

This means that the coefficients of $\hat\beta_j$ and $\hat\beta_k$ will themselves be highly correlated but not be well determined individually, i.e., have a large variance. In the extreme case of "perfect" correlation, the matrix $\mathbf{X}$ would have columns j and k collinear and hence only have rank $\le p-1$. When the correlation is less extreme, $\mathbf{X}$ is still of full rank p and $\mathbf{X}^\intercal\mathbf{X}$ is close to a singular matrix.[2] One approach to this problem is to *regularize* it by improving the condition of the matrix corresponding to $\mathbf{X}^\intercal\mathbf{X}$.

To give a numerical example, say $\mathbf{x}^{(2)} \approx -2\mathbf{x}^{(1)}$, then $1\mathbf{x}^{(1)}$ is close to $3\mathbf{x}^{(1)} + \mathbf{x}^{(2)}$ and hence to $5\mathbf{x}^{(1)} + 2\mathbf{x}^{(2)}$ or $51\mathbf{x}^{(1)} + 25\mathbf{x}^{(2)}$. One way to make the linear combination more clearly determined is to restrict the coefficients to small (absolute) values, or, more conveniently requiring that $\sum_j \beta_j^2$ be "small". This leads to the so called *ridge regression*

$$\widehat{\boldsymbol\beta}^*(s) = \underset{\|\boldsymbol\beta\|^2\le s}{\arg\min}\, \|\mathbf{Y} - X\boldsymbol\beta\|^2,$$

which can be shown (by way of a Lagrange multiplier) to be equivalent to

$$\widehat{\boldsymbol\beta}^*(\lambda) = \underset{\boldsymbol\beta}{\arg\min}\ \{\|\mathbf{Y} - X\boldsymbol\beta\|^2 + \lambda\,\|\boldsymbol\beta\|^2\}, \tag{7.6}$$

where there is a one-to-one relationship between λ and the bound s above, and the limit $\lambda \to 0$ corresponds to $s = max\,\|\boldsymbol\beta\| \to \infty$, namely the ordinary least squares of chapter 1. As there, by setting derivatives $\partial/\partial\beta$ to zero, this minimization problem is equivalent to the "normal equations"

$$(X^\intercal X + \lambda I)\widehat{\boldsymbol\beta}^* = X^\intercal\mathbf{Y}, \tag{7.7}$$

where the $p\times p$ matrix $(X^\intercal X + \lambda I)$ will be non-singular (and "well-conditioned") as soon as $\lambda > 0$ is large enough.

The ridge penalty entails that $\hat\beta_j(\lambda) \to 0$ (*"shrinking"*) when $\lambda \to \infty$, and also, in general, $\hat\beta_j \to \hat\beta'_j$ ("shrinking together") for two different coefficients.

Therefore, it's intutitive that $\widehat{\boldsymbol\beta}^*$ will have some bias ($\mathbb{E}[\widehat{\boldsymbol\beta}^*] \ne \boldsymbol\beta$), but that its variance(s) can be considerably smaller than $\widehat{\boldsymbol\beta}^{LS}$ such that mean squared errors are smaller. As for smoothing (in particular, spline smoothing, sect. 3.4) we have a regularization parameter λ which determines the trade-off between bias and variance, and as there, we'd use something like cross validation to determine an approximately optimal value for λ.

In the literature (and the R function `lm.ridge()` from the package `MASS`) there are "cheaper" approaches like GCV (see ch. 4) for determining an approximately optimal λ. In practice, one often wants to look at the *ridge traces*, i.e., a plot of the coefficients $\hat\beta_j(\lambda)$ vs λ. As an example we consider the longley macroeconomical data, for once modelling $y =$ `GNP.deflator` as function of the other six variables. The ridge traces $\hat\beta_j(\lambda)$ are shown in Figure 7.10. We have used a "relevant" interval for λ where the shrinking towards zero is visible (but still somewhat distant from the limit).

7.7.3 The Lasso

In some sense, the "Lasso" (Tibshirani, 1996) regression is just a simple variant of ridge regression. However with the goal of *variable selection* in mind, the lasso brings a major improvement:

[2] such that $\mathrm{Cov}(\widehat{\boldsymbol\beta}) = \sigma^2(X^\intercal X)^{-1}$ (section 1.4.1) will have very large entries corresponding to the high variance (and correlation) of $\hat\beta_j$ and $\hat\beta_k$.

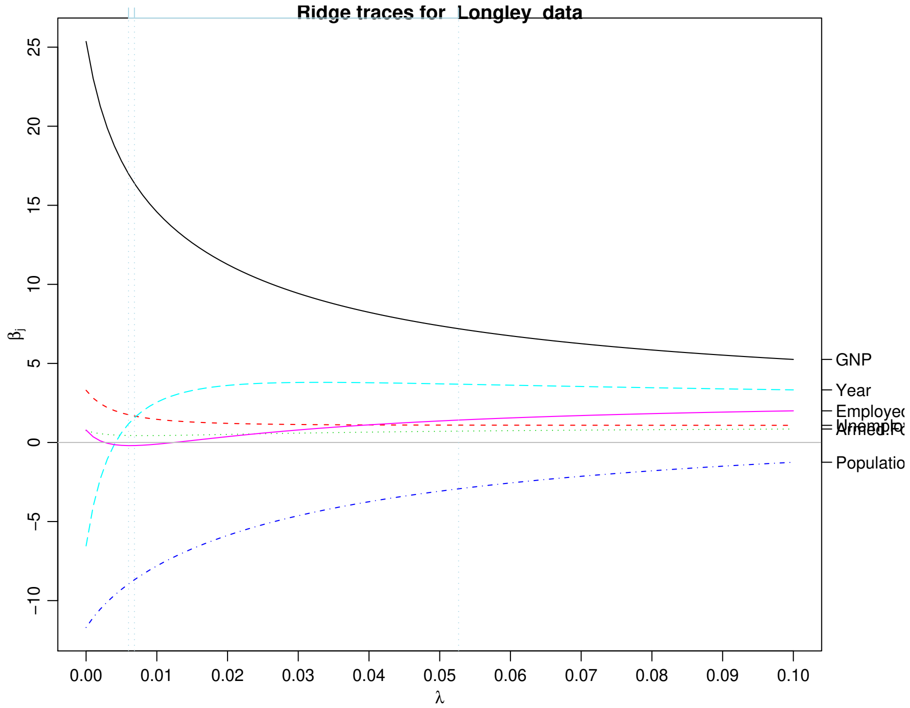

Figure 7.10: Ridge traces for the six coefficients $\beta_j(\lambda)$ $(j = 1, \ldots, 6)$ for the Longley data. The vertical lines indicate traditional estimates of the optimal ridge parameter λ.

The lasso can be defined by restricting the *absolute* instead of the squared values of the coefficients, i.e.,

$$\widehat{\boldsymbol{\beta}}^*(s) = \underset{\sum_j |\beta_j| \le s}{\arg\min} \|\mathbf{Y} - X\boldsymbol{\beta}\|^2 ,$$

or

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^*(\lambda) &= \underset{\boldsymbol{\beta}}{\arg\min} \{\|\mathbf{Y} - X\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|\}, \\ &= \underset{\boldsymbol{\beta}}{\arg\min} \{\|\mathbf{Y} - X\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1\}. \end{aligned} \tag{7.8}$$

As opposed to the ridge regression case above, this problem is not solvable by simple linear algebra but rather needs quadratic programming or related algorithms.

On the other hand, the solution is much more interesting, because it will be frequent that $\hat{\beta}_j$ will become *exactly 0* as soon as $\lambda > \lambda_j$, in other words, choosing λ here, automatically means model selection, namely only choosing regressor variables $\mathbf{x}^{(j)}$ with $\beta_j! = 0$.

This can be visualized considering the "lasso traces".