The Elements of Statistical Learning - Chapter 3 Exercises

Exercise 3.1

Show that the F statistic (3.13) for dropping a single coefficient from a model is equal to the square of the corresponding z-score (3.12).

Solution

Without loss of generality, assume that the smaller model has had the final feature $\mathbf{x}_p$ removed. Let $\hat{\mathbf{y}}$ denote the least squares approximation for the larger model and $\hat{\mathbf{y}}'$ be that for the smaller. We need to show that

$$\frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\mathrm{RSS}_1/(N-p-1)} = \frac{\hat{\beta}_p^2}{\hat{\sigma}^2 v_p},$$

where v_j is the jth diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$.

First note that by definition $\hat{\sigma}^2 = \mathrm{RSS}_0/(N-p-1)$. Moreover, since $\hat{\mathbf{y}}$ is the projection of $\mathbf{y}$ onto the column space of $\mathbf{X}$, their difference is orthogonal to any element of the columns space. In particular, it is orthogonal to $\hat{\mathbf{y}} - \hat{\mathbf{y}}'$, so

$$\|\mathbf{y} - \hat{\mathbf{y}}'\|^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}} - \hat{\mathbf{y}}'\|^2$$
$$\Rightarrow \mathrm{RSS}_0 - \mathrm{RSS}_1 = \|\mathbf{y} - \hat{\mathbf{y}}'\|^2 - \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \|\hat{\mathbf{y}} - \hat{\mathbf{y}}'\|^2.$$

Now let $\mathbf{z}_0, \dots, \mathbf{z}_p$ denote the orthogonal basis of the column space of $\mathbf{X}$, obtained from $\mathbf{x}_0, \dots, \mathbf{x}_p$ using the Gram-Schmidt process (Algorithm 3.1). The least squares estimates $\hat{\mathbf{y}}$ and $\hat{\mathbf{y}}'$ are the projections of $\mathbf{y}$ onto the column space of $\mathbf{X}$ and

$$\mathrm{span}(\{\, \mathbf{x}_j \mid 0 \le j \le p-1 \,\}) = \mathrm{span}(\{\, \mathbf{z}_j \mid 0 \le j \le p-1 \,\})$$

respectively. Since the $\mathbf{z}_j$ are orthogonal, this implies that

$$\hat{\mathbf{y}} - \hat{\mathbf{y}}' = \frac{\langle \mathbf{z}_p, \mathbf{y}\rangle}{\langle \mathbf{z}_p, \mathbf{z}_p\rangle}\mathbf{z}_p = \hat{\beta}_p \mathbf{z}_p.$$

Putting these elements together, it just remains to show that $v_p = \|\mathbf{z}_p\|^{-2}$. But, if $\mathbf{X} = \mathbf{QR}$ is the QR-decomposition of $\mathbf{X}$ then $(\mathbf{X}^T\mathbf{X})^{-1} = \mathbf{R}^{-1}(\mathbf{R}^{-1})^T$. Since $\mathbf{R}$ is upper-triangular, the pth diagonal element of $\mathbf{R}^{-1}$ is $R_{pp}^{-1} = \|\mathbf{z}_p\|^{-1}$ and the claim follows.

Exercise 3.2

Given data on two variables X and Y, consider fitting a cubic polynomial regression model $f(X) = \sum_{j=0}^3 \beta_j X^j$. In addition to plotting the fitted curve, you would like a 95% confidence band about the curve. Consider the following two approaches:

1. At each point x_0, form a 95% confidence interval for the linear function $a^T\beta = \sum_{j=0}^3 \beta_j x_0^j$.
2. Form a 95% confidence set for β as in (3.15), which in turn generates confidence intervals for $f(x_0)$.

How do these approaches differ? Which band is likely to be wider? Conduct a small simulation experiment to compare the two methods.

Solution

Construction of Confidence Intervals

Let $p = 3$ and $\alpha = 0.05$ and write $x = (1, x_0, x_0^2, x_0^3)$.

1. We have $\beta \sim \mathcal{N}(\hat{\beta}, \sigma^2\mathbf{I})$, so

$$x^T(\hat{\beta} - \beta) \sim \mathcal{N}(0, x^T(\mathbf{X}^T\mathbf{X})^{-1}x\sigma^2).$$

Let $v = x^T(\mathbf{X}^T\mathbf{X})^{-1}x \in \mathbf{R}$. Then

$$\frac{x^T(\hat{\beta} - \beta)}{\hat{\sigma}\sqrt{v}} \sim t_{N-p-1},$$

where $\hat{\sigma}$ is the unbiased estimate for σ on p.47. Therefore, a $100(1-\alpha)\%$ confidence interval for $f(x_0) = x^T\beta$ has endpoints

$$\hat{f}(x_0) \pm \hat{\sigma}\sqrt{x^T(\mathbf{X}^T\mathbf{X})^{-1}x}\; t_{N-p-1,\alpha/2}$$

where $\hat{f}(x_0) = x^T\hat{\beta}$ and $t_{N-p-1,\alpha/2}$ is the $\frac{1}{2}\alpha$th percentile of a T distribution with $N-p-1$ degrees of freedom.

2. By the argument on p.49, an approximate 95% confidence set for $f(x_0)$ is the set of $x^T\beta$ such that β lies in

$$C = \left\{\, \beta \,\middle|\, \|\mathbf{X}(\beta - \hat{\beta})\|^2 \le \hat{\sigma}^2\chi^2 \,\right\},$$

where $\chi^2 = \chi^2_{p+1,\alpha}$ is the αth percentile of a chi-squared distribution with $p+1$ degrees of freedom.

First note that C is an ellipsoid in $\mathbb{R}^{p+1}$. This implies that the restriction of the linear function $x^T\beta$ to C achieves its maximum and minimum on the boundary ∂C of C and takes every value in between. In particular, $\{x^T\beta \mid \beta \in C\}$ is an interval and the endpoints of this interval are the maximum and minimum of $x^T\beta$ subject to the constraint $\beta \in \partial C$, or equivalently

$$\|\mathbf{X}(\beta - \hat{\beta})\|^2 = \hat{\sigma}^2\chi^2.$$

We solve this problem using Lagrange multipliers. Let

$$\mathcal{L}(\beta, \lambda) = x^T\beta - \lambda\left(\|\mathbf{X}(\beta - \hat{\beta})\|^2 - \hat{\sigma}^2\chi^2\right).$$

This has gradient $\nabla\mathcal{L} = (\frac{\partial\mathcal{L}}{\partial\beta}, \frac{\partial\mathcal{L}}{\partial\lambda})$ with

$$\frac{\partial\mathcal{L}}{\partial\beta} = x - 2\lambda\mathbf{X}^T(\mathbf{X}\beta - \mathbf{X}\hat{\beta}), \qquad \frac{\partial\mathcal{L}}{\partial\lambda} = \|\mathbf{X}(\beta - \hat{\beta})\|^2 - \hat{\sigma}^2\chi^2.$$

Any solution to our optimisation problem will have $\nabla\mathcal{L} = 0$. Setting the first partial derivative equation to zero gives

$$\mathbf{X}^T\mathbf{X}(\beta - \hat{\beta}) = \frac{1}{2\lambda}x \quad\Rightarrow\quad \beta = \hat{\beta} + \frac{1}{2\lambda}\mathbf{X}^T\mathbf{X}x.$$

Setting $\frac{\partial\mathcal{L}}{\partial\lambda} = 0$ and substituting this in give

$$\left\|\frac{1}{2\lambda}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x\right\|^2 = \hat{\sigma}^2\chi^2 \quad\Rightarrow\quad \frac{1}{2\lambda} = \pm\frac{\hat{\sigma}\chi}{\|\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x\|}.$$

So, since we know the optimisation problem has a maximuma and a minimum, they must occur at

$$\beta = \hat{\beta} \pm \hat{\sigma}\frac{(\mathbf{X}^T\mathbf{X})^{-1}x}{\|\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x\|}\chi.$$

Therefore, the confidence interval for $f(x_0)$ has endpoints

$$\hat{f}(x_0) \pm \hat{\sigma}\frac{x^T(\mathbf{X}^T\mathbf{X})^{-1}x}{\|\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x\|}\chi_{p+1,\alpha}.$$

Comparison

The key difference is that the first method gives a $100(1-\alpha)\%$ confidence interval at a *particular* x_0. That is, for a fixed x_0 the probability that $f(x_0)$ lies inside the confidence bands is $1-\alpha$. The second method starts with a $100(1-\alpha)\%$ confidence set for β and so is valid for all x_0 simultaneously. More precisely, there is a probability of $1-\alpha$ that β lies in the confidence set and thus $f(x_0)$ lies in the confidence interval *for all* x_0.

A less important distinction is that first method is exact (given the assumptions), whereas the second uses the approximation of t_{N-p-1} by the standard normal distribution $p+1$ times. The T distribution is less peaked than the standard normal, $t_{k,\alpha} > z_\alpha$ for $\alpha < 0.5$, but for N sufficiently large the difference will be insignificant.

Both of these differences imply that the chi-squared confidence interval should be wider.

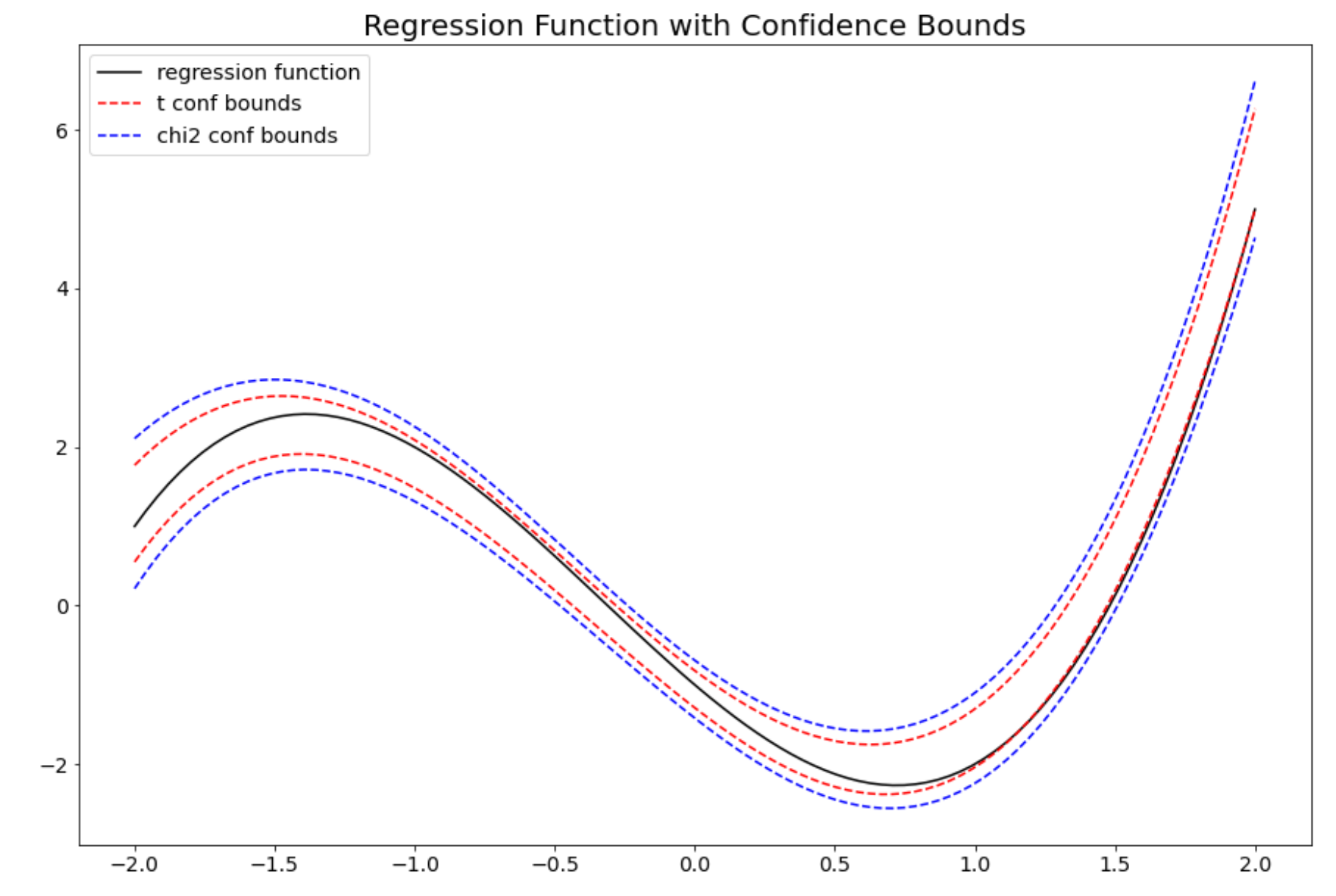
Simulation

The following function from a custom module calculates and prints the confidence bands from the two methods against the regression function $f(x_0)$. We use a simulated data set of size N with X sampled from a normal distribution with mean `xmean` and standard deviation `xstdev`, and show the $100(1-\alpha)\%$ confidence bands for x_0 in the range `plot_range` (= `xmean` ± 2 `xstdev` by default).

In [1]:

```
from confband import run_sim

beta = [-1, -3, 1, 1]
N = 100
sigma = 1
xmean = 0
xstdev = 1
alpha = 0.05
plot_range = None

run_sim(beta, alpha, N, sigma, xmean, xstdev, plot_range = plot_range)
```

```
Beta: [-1, -3, 1, 1]
Betahat: [-1.055 -2.678  1.113  0.949]


RSS:  90.9
Standard error:  0.973


Endpoints of 95.0% confidence interval at xmean:
T method:  [-1.29, -0.821]
Chi-square method:  [-1.419, -0.692]
```



Exercise 3.3

Gauss–Markov theorem:

(a) Prove the Gauss–Markov theorem: the least squares estimate of a parameter $a^T\beta$ has variance no bigger than that of any other linear unbiased estimate of $a^T\beta$ (Section 3.2.2).

(b) The matrix inequality $\mathbf{B} \preccurlyeq \mathbf{A}$ holds if $\mathbf{A} - \mathbf{B}$ is positive semidefinite. Show that if $\hat{\mathbf{V}}$ is the variance-covariance matrix of the least squares estimate of β and $\tilde{\mathbf{V}}$ is the variance-covariance matrix of any other linear unbiased estimate, then $\hat{\mathbf{V}} \preccurlyeq \tilde{\mathbf{V}}$.

Solution

(a) Let $\theta = a^T\beta$. The least squares estimator $\hat{\beta}$ of β satisfies $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$ so the least squares estimate $\hat{\theta} = a^T\hat{\beta}$ of θ has

$$\mathrm{Var}(\hat{\theta}) = \sigma^2 a^T(\mathbf{X}^{-1}\mathbf{X})^{-1}a.$$

Suppose $\tilde{\theta} = c^T\mathbf{y}$ is a linear unbiased estimator of θ. Then

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 I_N) \quad\Rightarrow\quad \tilde{\theta} \sim \mathcal{N}(c^T\mathbf{X}\beta, \sigma^2 c^T c).$$

Since $\tilde{\theta}$ is unbiased for θ,

$$\mathrm{E}(\tilde{\theta}) = \theta \quad\Rightarrow\quad c^T\mathbf{X}\beta = a^T\beta \quad\Rightarrow\quad (c^T\mathbf{X} - a^T)\beta = 0.$$

But this must hold for *every* $\beta \in \mathbb{R}^{p+1}$, so $c^T\mathbf{X} = a^T$.

We have reduced the problem to showing that if $c^T\mathbf{X} = a^T$ then $c^Tc \ge a^T(\mathbf{X}^T\mathbf{X})^{-1}a$. We will establish this by showing that $a^T(\mathbf{X}^T\mathbf{X})^{-1}a$ is a global minimum for $\|c\|^2$ subject to the constraint $c^T\mathbf{X} = a^T$.

First, observe that a minimum exists since the solution set is non-empty (consider $c = (\mathbf{X}^{-1}\mathbf{X})^{-1}\mathbf{X}^T$), bounded below by 0, and closed. We will find it using Lagrange multipliers.

Let $\lambda \in \mathbb{R}^{p+1}$ be variables and define

$$\mathcal{L}(c, \lambda) = \|c\|^2 = \lambda^T(\mathbf{X}^Tc - a).$$

This satisfies

$$\frac{\partial\mathcal{L}}{\partial c} = 2c - \mathbf{X}\lambda, \qquad \frac{\partial\mathcal{L}}{\partial c} = \mathbf{X}^Tc - a.$$

At an extremum of our optimisation problem both of these are zero, so

$$\begin{aligned} c = \frac{1}{2}\mathbf{X}\lambda \quad&\Rightarrow\quad \frac{1}{2}\mathbf{X}^T\mathbf{X}\lambda = a \\ &\Rightarrow\quad \lambda = 2(\mathbf{X}^T\mathbf{X})^{-1}a \\ &\Rightarrow\quad c = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}a \\ &\Rightarrow\quad \tilde{\theta} = c^T\mathbf{y} = \hat{\theta}. \end{aligned}$$

Since the Lagrange multiplier has a unique solution, this must be the global minimum of the constrained optimisation problem. In fact we have proved something slightly stronger: that $\hat{\theta}$ is the unique unbiased estimator with minimal variance.

(b) Our proof is analogous to that in part (a). The covariance matrix of $\hat{\beta}$ is $\mathrm{Var}(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. If $\tilde{\beta} = \mathbf{Ay}$ is another unbiased estimator for β then $\tilde{\beta} \sim \mathcal{N}(\mathbf{AX}\beta, \sigma^2\mathbf{AA}^T)$. Since $\tilde{\beta}$ is unbiased, $\mathbf{AX}\beta = \beta$ for all $\beta \in \mathbb{R}^{p+1}$ and so $\mathbf{AX} = \mathbf{I}_{P+1}$. Thus we have reduced the problem to showing that if $\mathbf{A}$ is a $(p+1) \times N$ matrix with $\mathbf{AX} = \mathbf{I}_{p+1}$ then $\mathbf{AA}^T - (\mathbf{X}^T\mathbf{X})^{-1}$ is positive semi-definite, or equivalently

$$v^T\mathbf{AA}v \ge v^T(\mathbf{X}^T\mathbf{X})^{-1}v \quad \text{for all } v \in \mathbb{R}^{p+1} \setminus \{0\}.$$

Again we treat this as a constrained optimisation problem and employ Lagrange multipliers. Fix $v \in \mathbb{R}^{p+1}$. The set $\{\|\mathbf{A}^Tv\|^2 \mid \mathbf{AX} = \mathbf{I}_{p+1}\}$ is non-empty (take $\mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$), bounded below by zero, and closed so has a minimum.

Let Λ be a $(p+1)\times(p+1)$ matrix of variables and define

$$\mathcal{L}(\mathbf{A}, \Lambda) = \|\mathbf{A}^Tv\|^2 - \Lambda \cdot (\mathbf{AX} - \mathbf{I}_{p+1},$$

where'$\cdot$' denotes the dot product. This has

$$\frac{\partial\mathcal{L}}{\partial A} = 2vv^T\mathbf{A} - \Lambda\mathbf{X}^T, \qquad \frac{\partial\mathcal{L}}{\partial\Lambda} = \mathbf{AX} - \mathbf{I}_{p+1}.$$

At an extremum of the constrained optimisation problem both of these are zero, so

$$\begin{aligned} 2vv^T\mathbf{AX} = \Lambda\mathbf{X}^T\mathbf{X} \quad&\Rightarrow\quad 2vv^T = \Lambda\mathbf{X}^T\mathbf{X} \\ &\Rightarrow\quad \Lambda = vv^T(\mathbf{X}^T\mathbf{X}^{-1}) \\ &\Rightarrow\quad vv^T\mathbf{A} = vv^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T. \end{aligned}$$

Multiplying each term on the right by its transpose yields

$$\begin{aligned} vv^T\mathbf{AA}^Tvv^T &= vv^T(\mathbf{X}^T\mathbf{X})^{-1}vv^T \\ \left[v^T\mathbf{AA}^Tv\right]vv^T &= \left[v^T(\mathbf{X}^T\mathbf{X})^{-1}v\right]vv^T \\ v^T\mathbf{AA}^Tv &= v^T(\mathbf{X}^T\mathbf{X})^{-1}v, \end{aligned}$$

where the second line holds as both terms in square brackets are scalars and so commute with everything. Since the constrained optimisation problem has a minimum this must be it. This establishes the claim.

Exercise 3.4

Show how the vector of least squares coefficients can be obtained from a single pass of the Gram–Schmidt procedure (Algorithm 3.1). Represent your solution in terms of the QR decomposition of $\mathbf{X}$.

Solution

To obtain $\hat{\beta}$ from the Gram-Schmidt process add an extra step $2'$. Take $j \in \{1, \dots, p\}$ and suppose that for $k < j$ we have written

$$\mathbf{z}_k = \mathbf{x}_k + \sum_{l=0}^{k-1}\hat{\delta}_{lk}\mathbf{x}_l.$$

Then we define coefficients $\hat{\delta}_{kj}$ for $k = 0, \dots, j-1$ by

$$\begin{aligned} \mathbf{z}_j &= \mathbf{x}_j - \sum_{k=0}^{j-1}\hat{\gamma}_{kj}\mathbf{z}_k \\ &= \mathbf{x}_j - \sum_{k=0}^{j-1}\hat{\gamma}_{kj}\left(\mathbf{x}_k + \sum_{l=0}^{k}\hat{\delta}_{lk}\mathbf{x}_l\right) \\ &= \mathbf{x}_j + \sum_{k=0}^{j-1}\hat{\delta}_{kj}\mathbf{x}_k. \end{aligned}$$

Let Δ be the matrix with $\Delta_{ij} = \hat{\delta}_{ij}$ for $i < j$, ones on the diagonal, and zeros elsewhere. By construction,

$$\mathbf{Z} = \mathbf{X}\Delta \quad\Rightarrow\quad \Delta = \Gamma^{-1}.$$

By (3.32), $\hat{\beta} = \Gamma^{-1}\mathbf{Z}^T\mathbf{y} = \Delta\mathbf{Z}^T\mathbf{y}$ so we can calculate $\hat{\beta}$ explicitly:

$$\hat{\beta}_j = \mathbf{z}_j + \sum_{k=j+1}^{p}\hat{\delta}_{jk}\mathbf{z}_k.$$

Exercise 3.5

Consider the ridge regression problem (3.41). Show that this problem is equivalent to the problem

$$\hat{\beta}^c = \underset{\beta^c}{\mathrm{argmin}}\left\{\sum_{i=1}^{N}\left[y_i - \beta_0^c - \sum_{j=1}^{p}(x_{ij} - \bar{x}_j)\beta_j^c\right]^2 + \lambda\sum_{j=1}^{p}(\beta_j^c)^2\right\}.$$

Give the correspondence between β^c and the original β in (3.41). Characterize the solution to this modified criterion. Show that a similar result holds for the lasso.

Solution

Our solution is the same for both ridge regresssion and lasso. Since

$$\beta_0^c + \sum_{j=1}^{p}(x_{ij} - \bar{x}_j)\beta_j^c = \left(\beta_0^c - \sum_{j=1}^{p}\bar{x}_j\beta_j^c\right) + \sum_{j=1}^{p}x_{ij}\beta_j^c,$$

the two minimisation problems are equivalent with

$$\beta_0^c = \beta_0 + \sum_{j=1}^{p}\bar{x}_j\beta_j = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i$$

and $\beta_j^c = \beta_j$ for $j \ne 0$.

At the minimum, the derivative with respect to β_0^c of the expression in braces is zero, so

$$\sum_{i=1}^{N}\left[y_i - \beta_0^c - \sum_{j=1}^{p}(x_{ij} - \bar{x}_j)\beta_j^c\right] = 0 \quad\Rightarrow\quad \left(\sum_{i=1}^{N}y_i\right) - N\beta_0^c - 0 = 0$$

and thus the solution to the modified criterion is $\beta_0^c = \bar{y}$, $\beta_j^c = \beta_j$ for $j > 0$.

Exercise 3.6

Show that the ridge regression estimate is the mean (and mode) of the posterior distribution, under a Gaussian prior $\beta \sim \mathcal{N}(0, \tau^2\mathbf{I})$, and Gaussian sampling model $y \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbf{I})$. Find the relationship between the regularization parameter λ in the ridge formula, and the variances τ^2 and σ^2.

Solution

The posterior distribution for β has density function

$$\begin{aligned} f_{\beta|Y_1,\dots,Y_N}(\beta) &\propto f_{Y_1,\dots,Y_N|\beta}(y_1, \dots, y_N)f_\beta(\beta) \\ &= \left(\prod_{i=1}^{N}f_{Y|\beta}(y_i)\right)f_\beta(\beta) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i^T\beta)^2 - \frac{1}{\tau^2}\|\beta\|^2\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}\left(\|\mathbf{y} - \mathbf{X}\beta\|^2 + \frac{\sigma^2}{\tau^2}\|\beta\|^2\right)\right). \end{aligned}$$

So the mode of this distribution is

$$\underset{\beta}{\mathrm{argmax}}\left(f_{\beta|Y_1,\dots,Y_N}(\beta)\right) = \underset{\beta}{\mathrm{argmin}}\left(\|\mathbf{y} - \mathbf{X}\beta\|^2 + \frac{\sigma^2}{\tau^2}\|\beta\|^2\right).$$

This is the ridge regression estimate with $\lambda = \sigma^2/\tau^2$. Moreover, since the posterior is Gaussian (p.64) the mean equals the mode.

Exercise 3.7

Assume $y_i \sim \mathcal{N}(\beta_0 + x_i^T\beta, \sigma^2)$, $i = 1, 2, \dots, N$, and the parameters β_j, $j = 1, \dots, p$ are each distributed as $\mathcal{N}(0, \tau^2)$, independently of one another. Assuming σ^2 and τ^2 are known, and β_0 is not governed by a prior (or has a flat improper prior), show that the (minus) log-posterior density of β is proportional to $\sum_{i=1}^{N}(y_i - \beta_0 - \sum_j x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$ where $\lambda = \sigma^2/\tau^2$.

Solution

This follows from the solution to exercise 3.6.

Exercise 3.8

Consider the QR decomposition of the uncentered $N \times (p+1)$ matrix $\mathbf{X}$ (whose first column is all ones), and the SVD of the $N \times p$ centered matrix $\tilde{\mathbf{X}}$. Show that Q_2 and U span the same subspace, where Q_2 is the sub-matrix of Q with the first column removed. Under what circumstances will they be the same, up to sign flips?

Solution

For $j = 1, \dots, p$, the Gram-Schmidt algorithm implies

$$\mathbf{x}_j = \mathbf{z}_j + \sum_{k=0}^{j-1}\hat{\gamma}_{kj}\mathbf{z}_k$$
$$\Rightarrow \mathbf{x}_j - \hat{\gamma}_{0j}\mathbf{z}_0 = \mathbf{z}_j + \sum_{k=1}^{j-1}\hat{\gamma}_{kj}\mathbf{z}_k.$$

By definition,

$$\hat{\gamma}_{0j} = \frac{\langle\mathbf{z}_0, \mathbf{x}_j\rangle}{\langle\mathbf{z}_0, \mathbf{z}_0\rangle} = \frac{1}{N}\sum_{k=1}^{N}x_{jk}$$

equals the mean of $\mathbf{x}_j$; that is, the jth column $\tilde{\mathbf{x}}_j$ of $\tilde{\mathbf{X}}$. Since $\mathbf{z}_1, \dots, \mathbf{z}_p$ are scalar multiples of the columns of Q_2 this implies that the column space of Q_2 contains the column space of $\tilde{\mathbf{X}}$. By dimensions they are equal. The column space of U equals that of $\tilde{\mathbf{X}}$ by definition and the first claim follows.

Moving on to the second part, let R_2 be the $p \times p$ submatrix of R obtained by removing its first row and column. Observe that $\tilde{\mathbf{X}} = Q_2R_2$ is the QR decomposition of $\tilde{\mathbf{X}}$, we can forget $\mathbf{X}$ and work entirely with the SVD and QR decompositions of $\tilde{\mathbf{X}}$.

We claim that the columns of Q_2 equal those of U up to sign flips if and only if $\tilde{\mathbf{X}}$ (equivalently $\mathbf{X}$) has orthogonal columns.

First assume that $\tilde{\mathbf{X}}$ has orthogonal columns. Take a permutation $\sigma \in S_p$ such that $\|\tilde{\mathbf{x}}_{\sigma(1)}\| \ge \dots \ge \|\tilde{\mathbf{x}}_{\sigma(p)}\|$. Let $P = P_\sigma$ be the corresponding permutation matrix whose jth column is the standard basis vector $e_{\sigma(j)}$ so that $\tilde{\mathbf{X}}P$ has columns $\tilde{\mathbf{x}}_{\sigma(1)}, \dots, \tilde{\mathbf{x}}_{\sigma(p)}$. Let D be the diagonal matrix with jth diagonal entry $\|\tilde{\mathbf{x}}_{\sigma(j)}\|$, let $U = \tilde{\mathbf{X}}PD^{-1}$, and let $V = P$.

We claim that $\tilde{\mathbf{X}} = UDV^T$ is the SVD of $\tilde{\mathbf{X}}$. Indeed, since the columns $\tilde{\mathbf{X}}$ are orthogonal, the columns of U are orthonormal and so U and V are orthogonal matrices. Moreover, if $j \in \{1, \dots, p\}$ then $\tilde{\mathbf{X}}v_j$ is the jth column of $\tilde{\mathbf{X}}P = (\tilde{\mathbf{X}}PD^{-1})D$ so equals d_ju_j. Finally, $\tilde{\mathbf{X}}^Tu_j$ equals the jth column of $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}PD^{-1}$. Since U is orthogonal,

$$U^TU = I \quad\Rightarrow\quad D^{-1}P^T\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}PD^{-1} = I \quad\Rightarrow\quad \tilde{\mathbf{X}}^T\tilde{\mathbf{X}}PD^{-1} = PD,$$

so $\tilde{\mathbf{X}}^Tu_j = d_jv_j$. Thus U and V consist of left- and right-singular vectors of $\tilde{\mathbf{X}}$ respectively and $\tilde{\mathbf{X}} = UDV^T$ is its SVD.

It just remains to note that since the columns of $\tilde{\mathbf{X}}$ are orthogonal its QR-decomposition consists of $Q_2 = \tilde{\mathbf{X}}D^{-1}$ and $R_2 = D$. Observe that Q_2 and U have the same columns up to permutation.

For the converse, assume that Q_2 and U have the same columns up to permutation and sign. Then there exists a permutation matrix $P = P_\sigma$ and a diagonal matrix S with ±1 on the diagonal such that $U = Q_2PS$. So

$$Q_2R_2 = UDV^T \quad\Rightarrow\quad Q_2R_2 = Q_2PSDV^T \quad\Rightarrow\quad R_2 = PSDV^T,$$

since Q_2 is orthogonal so has a left-inverse. This implies that

$$R_2^TR_2 = (SPDV^T)(VDP^TS) = SPD^2P^TS = D^2,$$

so R_2 is upper triangular with orthogonal columns and hence diagonal. Therefore $\tilde{\mathbf{X}} = Q_2R_2$ has orthogonal columns.

A corollary of our proof is the following: the columns of Q_2 and U are equal *with the same ordering* if and only if $\tilde{\mathbf{X}}$ is orthogonal with $\|\tilde{\mathbf{x}}_1\| \ge \dots \ge \|\tilde{\mathbf{x}}_p\|$.

Exercise 3.9

Forward stepwise regression. Suppose we have the QR decomposition for the $N \times q$ matrix $\mathbf{X}_1$ in a multiple regression problem with response $\mathbf{y}$, and we have an additional $p - q$ predictors in the matrix $\mathbf{X}_2$. Denote the current residual by $\mathbf{r}$. We wish to establish which one of these additional variables will reduce the residual-sum-of squares the most when included with those in $\mathbf{X}_1$. Describe an efficient procedure for doing this.

Solution

Let $\mathbf{X}_1 = \mathbf{Q}_1\mathbf{R}_1$ be the QR decomposition of $\mathbf{X}_1$ and let $\mathbf{q}_0, \dots, \mathbf{q}_q$ denote the columns of $\mathbf{X}_1$. The current residual is $\mathbf{r} = \mathbf{y} - \mathbf{Q}_1\mathbf{Q}_1^T\mathbf{y}$ so the residual sum of squares is

$$\begin{aligned} \|\mathbf{y} - \mathbf{Q}_1\mathbf{Q}_1^T\mathbf{y}\|^2 &= \|\mathbf{y} - \sum_{j=0}^{q}\mathbf{q}_j\mathbf{q}_j^T\mathbf{y}\|^2 \\ &= \|\mathbf{y}\|^2 - \sum_{j=0}^{q}2\langle\mathbf{y}, \mathbf{q}_j\mathbf{q}_j^T\mathbf{y}\rangle + \sum_{j=0}^{q}\sum_{k=0}^{q}\langle\mathbf{q}_j\mathbf{q}_j^T\mathbf{y}, \mathbf{q}_k\mathbf{q}_k^T\mathbf{y}\rangle \\ &= \|\mathbf{y}\|^2 - \sum_{j=0}^{q}2\langle\mathbf{q}_j^T\mathbf{y}, \mathbf{q}_j^T\mathbf{y}\rangle + \sum_{j=0}^{q}\sum_{k=0}^{q}\langle\mathbf{q}_j^T\mathbf{y}, \mathbf{q}_j^T\mathbf{q}_k\mathbf{q}_k^T\mathbf{y}\rangle \\ &= \|\mathbf{y}\|^2 - \sum_{j=0}^{q}\langle\mathbf{q}_j^T\mathbf{y}, \mathbf{q}_j^T\mathbf{y}\rangle \\ &= \|\mathbf{y}\|^2 - \sum_{j=0}^{q}(\mathbf{q}_j^T\mathbf{y})^2. \end{aligned}$$

Adding another variable corresponds to adding another column $\mathbf{q}$ to $\mathbf{Q}_1$. Since $\mathbf{q}$ is a normalised version of the Gram-Schmidt vector $\mathbf{z}$, we can choose which variable to include as follows: for each column $\mathbf{x}$ of $\mathbf{X}_2$, calculate

$$\mathbf{z} = \mathbf{x} - \sum_{k=0}^{q}\langle\mathbf{q}_k, \mathbf{x}\rangle\mathbf{q}_k$$

and pick the variable for which $|\mathbf{z}^T\mathbf{y}|\big/\|\mathbf{z}\|$ is maximal.

Because it follows relatively easy, we deduce a full procedure for forward-stepwise regression. This is implemented in `stepsel.py`.

We run a loop for $k = 0, \dots, p$. At the start of step k assume that we have:

- $\mathcal{A}$: a set of k active variables
- RSS: the residual sum-of-squares at the end of the previous step
- $\mathbf{B}$: a $(p+1) \times k$ matrix whose jth column gives the approximation to β at the end of the jth step
- $\mathbf{Q}$: an $N \times k$ matrix that gives the QR-decomposition for $\mathbf{X}$ restricted to the active columns
- $\mathbf{R}_{\mathrm{inv}}$: a $(p+1) \times k$ matrix whose restriction to its k non-zero rows gives the $\mathbf{R}^{-1}$ in the QR-decomposition described above.

In the kth step let $\mathbf{M} = \mathbf{I}_{p+1} - \mathbf{R}_{\mathrm{inv}}\mathbf{Q}^T\mathbf{X}$ and consider

$$\mathbf{XM} = \mathbf{X} - \mathbf{XR}_{\mathrm{inv}}\mathbf{Q}^T\mathbf{X} = \mathbf{X} - \mathbf{QQ}^T\mathbf{X}.$$

The jth column of this is

$$\mathbf{x}_j - \sum_{a\in\mathcal{A}}\langle\mathbf{q}_a, \mathbf{x}_j\rangle\mathbf{q}_a$$

and so right multiplication by $\mathbf{M}$ gives the orthogonal projection of $\mathbf{X}$ onto the space spanned by its active columns. In particular, the columns of $\mathbf{XM}$ indexed by the active set are zero.

Rescale the columns of $\mathbf{M}$ such that the columns of $\mathbf{XM}$ have unit norm. Let j be the index at which the element-wise square of $(\mathbf{XM})^T\mathbf{y}$ is greatest and let m denote this maximal value. By the solution to Exercise 3.9 described above, j is the index that will decrease the RSS the most and m is the amount it will decrease by. So, at the end of this step we add j to the active set, subtract m from RSS, and append the jth columns of $\mathbf{M}$ and $\mathbf{XM}$ to $\mathbf{R}_{\mathrm{inv}}$ and $\mathbf{Q}$ respectively. The new approximation to β is $\mathbf{R}_{\mathrm{inv}}\mathbf{Q}^T\mathbf{y}$.

Exercise 3.10

Backward stepwise regression. Suppose we have the multiple regression fit of $\mathbf{y}$ on $\mathbf{X}$, along with the standard errors and Z-scores as in Table 3.2. We wish to establish which variable, when dropped, will increase the residual sum-of-squares the least. How would you do this?

Solution

As in the defintion of the F statistic (3.13), let RSS_1 denote the residual sum of squares for the existing model and RSS_0 the same quantity for the model with one variable removed. We wish to choose the variable which minimises RSS_0. This is equivalent to minimising the F statistic

$$F = \frac{RSS_0 - RSS_1}{RSS_1\big/(N-p-1)}$$

or its square root. But by exercise 3.1 $\sqrt{F}$ equals the Z-score. Thus eliminating the variable with the lowest Z-score will increase the residual sum of squares the least.

We will now describe an efficient full procedure for backward-stepwise selection. This is implemented in `stepsel.py`.

We run a loop for $k = 0, \dots, p$. At the start of the kth step we have:

- $\mathcal{A}$: a set of $p+1-k$ active variables
- RSS: the residual sum-of-squares at the end of the previous step
- $\mathbf{B}$: a $(p+1) \times k$ matrix whose jth column gives the approximation to β at the start of the jth step
- $\mathbf{Q}$: an $N \times (p+1-k)$ matrix that gives the QR-decomposition for $\mathbf{X}$ restricted to the active columns
- $\mathbf{R}_{\mathrm{inv}}$: a $(p+1-k) \times (p+1-k)$ matrix that gives the $\mathbf{R}^{-1}$ in the QR-decomposition described above.

At the start, $\mathcal{A} = \{0, \dots, p\}$, $\mathbf{B}$ is empty, $\mathbf{Q}$ and $\mathbf{R}_{\mathrm{inv}}$ come from the QR decomposition of $\mathbf{X}$, and

$$RSS = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \|\mathbf{y} - \mathbf{QQ}^T\mathbf{y}\|^2 = \mathbf{y}^T(\mathbf{I}_N - \mathbf{QQ}^T)^2\mathbf{y}.$$

In the kth step we calculate the current approximation $\mathbf{R}_{\backslash\mathrm{inv}}\mathbf{Q}^T\mathbf{y}$ to β and add it as the last column of $\mathbf{B}$. Then we examine the Z-scores to determine which feature to remove. We can ignore the constant $\hat{\sigma}$. Moreover, v_j is the jth diagonal element of $\mathbf{R}_{\backslash\mathrm{inv}}\mathbf{R}_{\backslash\mathrm{inv}}^T$, which equals the square norm of the jth row of $\mathbf{R}_{\backslash\mathrm{inv}}$. Thus it suffices to examine $\hat{\beta}$ (the last column of $\mathbf{B}$) divided element-wise by the row norms of $\mathbf{R}_{\backslash\mathrm{inv}}$ and remove the index j with the smallest value.

We then calculate the new residual sum-of-squares $RSS + (\mathbf{q}_j^T\mathbf{y})^2$, remove the jth column and the jth row and column from $\mathbf{Q}$ and $\mathbf{R}_{\backslash\mathrm{inv}}$ respectively, and remove j from $\mathcal{A}$.

By keeping track of things as we go we end with the list of variables in the order they were removed, the approximation to β at each step, and the RSS at each step.

Exercise 3.11

Show that the solution to the multivariate linear regression problem (3.40) is given by (3.39). What happens if the covariance matrices Σ_i are different for each observation?

Solution

We wish the find the minimum of

$$RSS(\mathbf{B}; \Sigma) = \sum_{i=1}^{N}(y_i - f(x_i))^T\Sigma^{-1}(y_i - f(x_i)).$$

Let a_{ij} denote the (i, j) entry of Σ^{-1}. Note that since Σ is a covariance matrix, both it and its inverse are symmetric. We can expand RSS into

$$RSS = \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{l=1}^{K}(y_{ik} - f_k(x_i))a_{kl}(y_{il} - f_l(x_i)),$$

where $f_k(x_i) = \sum_{j=0}^{p}x_{ij}\beta_{jk}$ is the (i, k) entry of $\mathbf{XB}$. This implies that

$$\frac{\partial RSS}{\partial\beta_{mn}} = -2\sum_{i=1}^{N}x_{im}\left(\sum_{k=1}^{K}(y_{ik} - f_k(x_i))a_{kn}\right)$$
$$\Rightarrow \frac{\partial RSS}{\partial\mathbf{B}} = -\mathbf{X}^T(\mathbf{Y} - \mathbf{XB})\Sigma^{-1}.$$

Since RSS is bounded below and quadratic in $\mathbf{B}$ it has a global minimum. Setting $\frac{\partial RSS}{\partial\mathbf{B}} = 0$ gives the unique solution $\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ (assuming $\mathbf{X}$ has full column rank). So the solution is the same as K uncorrelated regressions.

If the covariance matrices Σ_i are different for different observations then we need to minimise

$$RSS(\mathbf{B}; \Sigma) = \sum_{i=1}^{N}(y_i - f(x_i))^T\Sigma_i^{-1}(y_i - f(x_i)).$$

Let $a_{jk}^{(i)}$ denote the (j, k) entry of Σ_i^{-1}. Then

$$\frac{\partial RSS}{\partial\beta_{mn}} = -2\sum_{i=1}^{N}x_{im}\left(\sum_{k=1}^{K}(y_{ik} - f_k(x_i))a_{kn}^{(i)}\right).$$

If $\mathbf{Z}$ denotes the matrix whose ith row is the ith row of $(\mathbf{Y} - \mathbf{XB})\Sigma^{-1}$ then $\frac{\partial RSS}{\partial\mathbf{B}} = -\mathbf{X}^T\mathbf{Z}$. Again, RSS must have a global minimum at which $\mathbf{X}^T\mathbf{Z} = 0$. Since this is a system of linear equations in the β_{ij} it's easy to find the solutions in practice but I cannot find a closed-form solution.

Exercise 3.12

Show that the ridge regression estimates can be obtained by ordinary least squares regression on an augmented data set. We augment the centered matrix $\mathbf{X}$ with p additional rows $\sqrt{\lambda}\mathbf{I}$, and augment $\mathbf{y}$ with p zeros. By introducing artificial data having response value zero, the fitting procedure is forced to shrink the coefficients toward zero. This is related to the idea of hints due to Abu-Mostafa (1995), where model constraints are implemented by adding artificial data examples that satisfy them.

Solution

Let $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$ denote the augmented matrix and vector, respectively. Without loss of generality we assume that the new rows have been added at the bottom. Then

$$\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = \mathbf{X}^T\mathbf{X} + (\sqrt{\lambda}\mathbf{I}_p)^T(\sqrt{\lambda}\mathbf{I}_p) = \mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p$$

and $\tilde{\mathbf{X}}^T\tilde{\mathbf{y}} = \mathbf{X}^T\mathbf{y}$, so the least squares estimate using $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$ is

$$(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{y}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y};$$

the ridge regression estimate for the non-augmented data.

Exercise 3.13

Derive the expression (3.62), and show that $\hat{\beta}^{\mathrm{pcr}}(p) = \hat{\beta}^{\mathrm{ls}}$.

Solution

For the first part observe that

$$\hat{\mathbf{y}}^{\mathrm{pcr}} = \sum_{m=1}^{M}\hat{\theta}_m\mathbf{z}_m = \sum_{m=1}^{M}\hat{\theta}_m\mathbf{X}v_m = \mathbf{X}\left(\sum_{m=1}^{M}\hat{\theta}_mv_m\right),$$

so $\hat{\beta}^{\mathrm{pcr}}(M) = \sum_{m=1}^{M}\hat{\theta}_mv_m$. The second part follows from the fact that $\mathbf{X}$ and $\mathbf{Z} = \mathbf{UD}$ have the same column space, but we can also show it explicitly. Using the SVD decomposition $\mathbf{X} = \mathbf{UDV}^T$,

$$\begin{aligned} \hat{\beta}^{\mathrm{ls}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= (\mathbf{VD}^2\mathbf{V}^T)^{-1}\mathbf{VDU}^T\mathbf{y} \\ &= \mathbf{VD}^{-1}\mathbf{U}^T\mathbf{y} \\ &= \mathbf{VD}^{-2}\mathbf{Z}^T\mathbf{y} \\ &= \sum_{m=1}^{p}\frac{\langle\mathbf{z}_m, \mathbf{y}\rangle}{d_m^2}v_m. \end{aligned}$$

We're done since $\|\mathbf{z}_m\| = d_m^2$.

Exercise 3.14

Show that in the orthogonal case, PLS stops after $m = 1$ steps, because subsequent φ_{mj} in step 2 in Algorithm 3.3 are zero.

Solution

It suffices to show that $\langle\mathbf{x}_j^{(1)}, \mathbf{y}\rangle = 0$ for all j, or equivalently that

$$\frac{\langle\mathbf{z}_0, \mathbf{x}_j\rangle}{\langle\mathbf{z}_0, \mathbf{z}_0\rangle}\langle\mathbf{z}_0, \mathbf{y}\rangle = \langle\mathbf{x}_j, \mathbf{y}\rangle.$$

Using the fact that the $\mathbf{x}_j$ are orthonormal (they have norm one since they are standardised with mean zero and unit variance) we have

$$\langle\mathbf{z}_0, \mathbf{x}_j\rangle = \langle\mathbf{x}_j, \mathbf{y}\rangle\|\mathbf{x}_j\|^2 = \langle\mathbf{x}_j, \mathbf{y}\rangle,$$
$$\langle\mathbf{z}_0, \mathbf{z}_0\rangle = \sum_{k=1}^{p}\langle\mathbf{x}_k, \mathbf{y}\rangle^2\|\mathbf{x}_k\|^2 = \sum_{k=1}^{p}\langle\mathbf{x}_k, \mathbf{y}\rangle^2,$$
$$\langle\mathbf{z}_0, \mathbf{y}\rangle = \sum_{k=1}^{p}\langle\mathbf{x}_k, \mathbf{y}\rangle^2.$$

The claim follows.

Exercise 3.15

Verify expression (3.64), and hence show that the partial least squares directions are a compromise between the ordinary regression coefficient and the principal component directions.

Solution

We start with $m = 1$. Note that

$$\mathrm{Corr}^2(\mathbf{y}, \mathbf{X}\alpha)\mathrm{Var}(\mathbf{X}\alpha) = \frac{\mathrm{Cov}^2(\mathbf{y}, \mathbf{X}\alpha)}{\mathrm{Var}(\mathbf{y})\mathrm{Var}(\mathbf{X}\alpha)}\mathrm{Var}(\mathbf{X}\alpha),$$

so maximising this over $\|\alpha\| = 1$ is equivalent to maximising $\mathbf{y}^T\mathbf{X}\alpha$. We have

$$\mathbf{y}^T\mathbf{X}\alpha = \langle\mathbf{X}^T\mathbf{y}, \alpha\rangle = \|\mathbf{X}^T\mathbf{y}\|\|\alpha\|\cos(\theta),$$

so this is maximised at $\alpha = \mathbf{X}^T\mathbf{y}\big/\|\mathbf{X}^T\mathbf{y}\|$, which equals $\hat{\varphi}_1$ up to a scalar.

Now take $m > 1$. The sample covariance matrix is $\mathbf{S} = \frac{1}{N}\mathbf{X}^T\mathbf{X}$ so we need to maximise $\mathbf{y}^T\mathbf{X}\alpha$ subject to $\|\alpha\| = 1$ and $\langle\mathbf{X}\alpha, \mathbf{z}_l\rangle = 0$ for $l < m$. Let π denote projection from $\mathbb{R}^p$ onto the subspace spanned by $\{\mathbf{z}_1, \dots, \mathbf{z}_{m-1}\}$, where $\mathbf{z}_l = \mathbf{X}\hat{\varphi}_l$. Note that $(1-\pi)(\mathbf{x}_j) = \mathbf{x}_j^{(m-1)}$ for all j.

The condition $\langle\mathbf{X}\alpha, \mathbf{z}_l\rangle = 0$ for $l < m$ is equivalent to $(1-\pi)(\mathbf{X}\alpha) = \mathbf{X}\alpha$. However,

$$(1-\pi)(\mathbf{X}\alpha) = (1-\pi)\left(\sum_{j=1}^{p}\alpha_j\mathbf{x}_j\right) = \sum_{j=1}^{p}\alpha_j\mathbf{x}_j^{(m-1)} = \mathbf{X}^{(m-1)}\alpha,$$

where $X^{(m-1)}$ is the matrix whose jth column is $\mathbf{x}_j^{(m-1)}$, so the orthogonality condition holds iff $\mathbf{X}\alpha = \mathbf{X}^{(m-1)}\alpha$. If so then

$$\mathbf{y}^T\mathbf{X}\alpha = \mathbf{y}^T\mathbf{X}^{(m-1)}\alpha = \langle(\mathbf{X}^{(m-1)})^T\mathbf{y}, \alpha\rangle = \|(\mathbf{X}^{(m-1)})^T\mathbf{y}\|\|\alpha\|\cos(\theta).$$

Clearly this is maximised at $\alpha = (\mathbf{X}^{(m-1)})^T\mathbf{y}\big/\left\|(\mathbf{X}^{(m-1)})^T\mathbf{y}\right\|$, which equals $\hat{\varphi}_m$ up to a scalar.

Note that as a corollary we have $\mathbf{z}_m = \sum_j\hat{\varphi}_{mj}\mathbf{x}_j^{(m-1)} = \sum_j\hat{\varphi}_{mj}\mathbf{x}_j$.

For the second part, note that the principal component directions maximise $\mathrm{Var}(\mathbf{X}\alpha)$ and the partial least squares directions maximise $\mathrm{Corr}^2(\mathbf{y}, \mathbf{X}\alpha)\mathrm{Var}(\mathbf{X}\alpha)$. We claim that the ordinary regression coefficient $\hat{\beta}$ maximises $\mathrm{Corr}^2(\mathbf{y}, \mathbf{X}\beta)$ and hence PLS is a compromise between the other two.

It suffices to maximise the log

$$C = \ln\big(\mathrm{Corr}^2(\mathbf{y}, \mathbf{X}\beta)\big) = 2\ln\langle\mathbf{y}, \mathbf{X}\beta\rangle - \ln\langle y, y\rangle - \ln\langle\mathbf{X}\beta, \mathbf{X}\beta\rangle.$$

This has partial derivative

$$\frac{\partial C}{\partial\beta} = 2\left[\frac{\mathbf{X}^T\mathbf{y}}{\langle\mathbf{y}, \mathbf{X}\beta\rangle} - \frac{\mathbf{X}^T\mathbf{X}\beta}{\langle\mathbf{X}\beta, \mathbf{X}\beta\rangle}\right].$$

Setting this to zero yields

$$\frac{\langle\mathbf{y}, \mathbf{X}\beta\rangle}{\langle\mathbf{X}\beta, \mathbf{X}\beta\rangle}\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \hat{\beta},$$

so β must be a scalar multiple of $\hat{\beta}$. In fact any non-zero multiple satisfies the equation.

It remains to show that this is a maximum. However, note that $\mathrm{Corr}^2(\mathbf{y}, \mathbf{X}\beta)$ is invariant under scaling of β. On the closed, bounded set $\{\beta \mid \|\beta\| = 1\}$ the function must have both a maximum and a minimum. Thus the extremum $\beta = \hat{\beta}$ we found must be a global maximum (the minimum - zero - didn't show up in our analysis as C is undefined when the correlation is zero).

Exercise 3.16

Derive the entries in Table 3.4, the explicit forms for estimators in the orthogonal case.

Solution

Assume that the columns of $\mathbf{X}$ are orthonormal. We go through the different methods in order.

Best subset: First note that if a variable $\mathbf{x}_j$ is included in the subset then its coefficient $\hat{\beta}_j = \langle\mathbf{x}_j, \mathbf{y}\rangle$ is independent of what other variables are included. Take $S \subseteq \{0, \dots, p\}$ and let $\beta = (\beta_j)$ where $\beta_j = \hat{\beta}_j$ if $j \in S$ and zero otherwise. Then

$$\begin{aligned} \|\mathbf{y} - \mathbf{X}^T\beta\|^2 &= \left\|\mathbf{y} - \sum_{j\in S}\beta_j\mathbf{x}_j\right\|^2 \\ &= \|\mathbf{y}\|^2 - 2\sum_{j\in S}\langle\mathbf{x}_j, \mathbf{y}\rangle^2 + \sum_{j,k\in S}\langle\mathbf{x}_j, \mathbf{y}\rangle\langle\mathbf{x}_k, \mathbf{y}\rangle\langle\mathbf{x}_j, \mathbf{x}_k\rangle \\ &= \|\mathbf{y}\|^2 - \sum_{j\in S}\langle\mathbf{x}_j, \mathbf{y}\rangle^2. \end{aligned}$$

So the best subset of size M consists of the M indices with the greatest $|\hat{\beta}_j|$. This implies the claim.

Ridge: If X has orthonormal columns then $\mathbf{X}^T\mathbf{X} = \mathbf{I}$ so the ridge estimate is

$$\beta = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} = \frac{1}{1+\lambda}\hat{\beta}$$

as required.

Lasso: Consider the LAR path. Take the variable $\mathbf{x}_j$ with the greatest absolute correlation with $\mathbf{y}$. We begin moving along the path $\hat{\mathbf{f}}(\alpha) = \alpha\langle\mathbf{x}_j, \mathbf{y}\rangle\mathbf{x}_j$ for $\alpha \in [0, 1]$. Along this path the correlations with residuals are given by

$$|\langle\mathbf{x}_k, \mathbf{y} - \hat{\mathbf{f}}(\alpha)\rangle| = \begin{cases}(1-\alpha)|\langle\mathbf{x}_j, \mathbf{y}\rangle| & \text{if } j = k, \\ |\langle\mathbf{x}_k, \mathbf{y}\rangle| & \text{otherwise.}\end{cases}$$

Thus we continue along the current path until $(1-\alpha)|\langle\mathbf{x}_j, \mathbf{y}\rangle| = |\langle\mathbf{x}_k, \mathbf{y}\rangle|$ for some $k \ne j$ at which point we include k in the active set. As above, the coefficient of $\mathbf{x}_k$ will only affect the correlation of $\mathbf{x}_k$ with the residual. Thus

Thus the coefficient β_l of $\mathbf{x}_l$ is zero until the shared absolute covariance λ is greater than $|\langle\mathbf{x}_l, \mathbf{y}\rangle|$, after which it increases at the same rate that λ increases. More precisely, when the shared covariance is λ, the coefficient of $\mathbf{x}_l$ is

$$\beta_l = \begin{cases}0 & \text{if } \lambda > |\langle\mathbf{x}_l, \mathbf{y}\rangle|, \\ \mathrm{sign}(\langle\mathbf{x}_l, \mathbf{y}\rangle\,(|\langle\mathbf{x}_l, \mathbf{y}\rangle| - \lambda) & \text{otherwise.}\end{cases}$$

This agrees with the entry in Table 3.4. Indeed, since no coefficient in the LAR path ever hits zero after being added to the active set, it is identical to the Lasso path.

Exercise 3.17

Repeat the analysis of Table 3.3 on the spam data discussed in Chapter 1.

Solution

This analysis can be found in `spam-data/spam-regression-analysis.ipynb`. The test errors were as follows:

Least-Squares	Forward-Stepwise	Ridge	Lasso	PCR	PLS
0.1160	0.1162	0.1141	0.1163	0.1196	0.1163

Note that best subset couldn't be performed because there were too many featuresm, so forward-stepwise selection was substituted.

Exercise 3.18

Read about conjugate gradient algorithms (Murray et al., 1981, for example), and establish a connection between these algorithms and partial least squares.

Solution

TODO

Exercise 3.19

Show that $\|\hat{\beta}^{\text{ridge}}\|$ increases as its tuning parameter $\lambda \to 0$. Does the same property hold for the lasso and partial least squares estimates? For the latter, consider the "tuning parameter" to be the successive steps in the algorithm.

Solution

Ridge: Let

$$R(\beta; \lambda) = \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda\|\beta\|^2.$$

Take $\lambda > \lambda' > 0$ and let $\hat{\beta}$ and $\hat{\beta}'$ be the ridge approximations with tuning parameters λ and λ', respectively (so they minimise $R(\beta; \lambda)$ and $R(\beta; \lambda')$). Then by definition

$$R(\hat{\beta}; \lambda) \leq R(\hat{\beta}'; \lambda) \qquad \text{and} \qquad R(\hat{\beta}'; \lambda') \leq R(\hat{\beta}; \lambda').$$

Subtract the second inequality form the first yields

$$\begin{aligned} R(\hat{\beta}; \lambda) - R(\hat{\beta}; \lambda') &\leq R(\hat{\beta}'; \lambda) - R(\hat{\beta}'; \lambda') \\ (\lambda - \lambda')\|\hat{\beta}\|^2 &\leq (\lambda - \lambda')\|\hat{\beta}'\|^2 \\ \|\hat{\beta}\| &\leq \|\hat{\beta}'\|, \end{aligned}$$

so the ridge approximation increases in L^2 norm as λ decreases.

Lasso: The same proof as above shows that the lasso approximation to β increases in L^1 *norm* as λ decreases. However, this doesn't necessarily mean that it increases in L^2 norm. For example suppose that $p = 2$ and assume the least squares approximation has $\hat{\beta}_1, \hat{\beta}_2 > 0$. If lasso modification of the LAR path went from $(0, 0)$ to $(\hat{\beta}_1 + \hat{\beta}_2, 0)$ to $(\hat{\beta}_1, \hat{\beta}_2)$ then the L^1 norm of the coefficient vector would always be increasing as λ decreases, but the L^2 norm would decrease on the last segment.

It remains to give an example to show this behaviour is possible. Take $p = 2$ and $N = 3$ - the minimal values for an example. Take

$$\mathbf{x}_1 = \begin{pmatrix} -\sqrt{\frac{3}{2}} \\ 0 \\ \sqrt{\frac{3}{2}} \end{pmatrix}, \qquad \mathbf{x}_2 = \begin{pmatrix} -\sqrt{\frac{1}{2}} \\ -\sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} \end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}.$$

Note that all three vectors have mean zero and the first two have unit variance. Since

$$\langle \mathbf{x}_1, \mathbf{y} \rangle = \sqrt{6} > \sqrt{2} = \langle \mathbf{x}_2, \mathbf{y} \rangle,$$

the first step is in the direction of $\mathbf{x}_1$. This ends at $\frac{\sqrt{2}}{11}(\sqrt{3} + 5)\mathbf{x}_1$ since at this point the residual $\mathbf{r}$ has equal correlation with $\mathbf{x}_1$ and $\mathbf{x}_2$ ($\langle \mathbf{x}_j, \mathbf{r} \rangle = 5\sqrt{6}$ for $j = 1, 2$). We then continue in a straight line in direction $\begin{pmatrix} -1 \\ 1 \end{pmatrix}$ to the least squares solution

$$\hat{\beta} = \frac{\sqrt{2}}{11}\begin{pmatrix} \sqrt{3} \\ 5 \end{pmatrix}.$$

So although the coefficient path increases in L^1 norm it decreases in L^2 norm along the final stretch.

PLS: Let $\hat{\beta}^{(m)}$ denote the PLS approximation after m steps. We claim that $\|\hat{\beta}^{(m)}\|$ increases as m increases. For $m = 0, \ldots, p$ let $\mathbf{X}^{(m)}$ denote the matrix whose jth column is $\mathbf{x}_j^{(m)}$. The proof relies heavily on the following three facts:

- $\mathbf{z}_m = \mathbf{X}^{(m-1)}\hat{\varphi}_m$;
- $\hat{\varphi}_m = \left(\mathbf{X}^{(m-1)}\right)^T\mathbf{y}$;
- $\mathbf{X}\hat{\varphi}_m = \mathbf{X}^{(l-1)}\hat{\varphi}_m$ for any $l \leq m$.

The first two of these are by definition. The last was essentially proved in the course of Exercise 3.15, but we quickly recall the proof here. By the equivalent characterisation of $\hat{\varphi}_m$ in Exercise 3.15, $\mathbf{X}\hat{\varphi}_m$ is orthogonal to $\mathbf{z}_1, \ldots, \mathbf{z}_{l-1}$. So if π is the projection onto the subspace they span then $(1 - \pi)\left(\mathbf{X}\hat{\varphi}_m\right) = \mathbf{X}\hat{\varphi}_m$. But by construction $(1 - \pi)(\mathbf{x}_j) = \mathbf{x}_j^{(l-1)}$ so

$$\mathbf{X}\hat{\varphi}_m = (1 - \pi)\left(\mathbf{X}\hat{\varphi}_m\right) = (1 - \pi)\left(\sum_{j=1}^{m} \hat{\varphi}_{mj}\mathbf{x}_j\right) = \sum_{j=1}^{m} \hat{\varphi}_{mj}\mathbf{x}_j^{(l-1)} = \mathbf{X}^{(l-1)}\hat{\varphi}_m.$$

Now we continue to the proof. First note that

$$\hat{\mathbf{y}}^{(m)} = \sum_{l=1}^{m} \hat{\theta}_l \mathbf{z}_l = \sum_{l=1}^{m} \hat{\theta}_l \mathbf{X}^{(l-1)}\hat{\varphi}_l = \sum_{l=1}^{m} \hat{\theta}_l \mathbf{X}\hat{\varphi}_l = \mathbf{X}\left(\sum_{l=1}^{m} \hat{\theta}_l \hat{\varphi}_l\right),$$

so $\hat{\beta}^{(m)} = \sum_{l=1}^{m} \hat{\theta}_l \hat{\varphi}_l$. We claim that $\hat{\theta}_l \geq 0$ and $\langle \hat{\varphi}_l, \hat{\varphi}_m \rangle \geq 0$ for all $l \leq m$. Since

$$\left\|\hat{\beta}^{(m)}\right\|^2 = \left\|\hat{\beta}^{(m-1)}\right\|^2 + \sum_{l=1}^{m-1} \hat{\theta}_l\hat{\theta}_m\langle \hat{\varphi}_l, \hat{\varphi}_m \rangle + \hat{\theta}_m^2\|\hat{\varphi}_m\|^2$$

this will establish that $\left\|\hat{\beta}^{(m)}\right\|$ increases with m.

The first claim holds because by definition $\hat{\theta}_l = \langle \mathbf{y}, \mathbf{z}_l \rangle / \|\mathbf{z}_l\|^2$ and

$$\begin{aligned} \langle \mathbf{y}, \mathbf{z}_l \rangle &= \langle \mathbf{y}, \mathbf{X}^{(l-1)}\hat{\varphi}_l \rangle \\ &= \langle \mathbf{y}, \mathbf{X}^{(l-1)}\left(\mathbf{X}^{(l-1)}\right)^T\mathbf{y} \rangle \\ &= \langle \left(\mathbf{X}^{(l-1)}\right)^T\mathbf{y}, \left(\mathbf{X}^{(l-1)}\right)^T\mathbf{y} \rangle \\ &= \left\|\hat{\varphi}_l\right\|^2 \geq 0. \end{aligned}$$

For the second claim observe that if $l \leq m$ then

$$\begin{aligned} \langle \hat{\varphi}_l, \hat{\varphi}_m \rangle &= \langle \left(\mathbf{X}^{(l-1)}\right)^T\mathbf{y}, \hat{\varphi}_m \rangle \\ &= \langle \mathbf{y}, \mathbf{X}^{(l-1)}\hat{\varphi}_m \rangle \\ &= \langle \mathbf{y}, \mathbf{X}^{(m-1)}\hat{\varphi}_m \rangle \\ &= \langle \left(\mathbf{X}^{(m-1)}\right)^T\mathbf{y}, \hat{\varphi}_m \rangle \\ &= \left\|\hat{\varphi}_m\right\|^2 \geq 0 \end{aligned}$$

as required.

Exercise 3.20

Consider the canonical-correlation problem (3.67). Show that the leading pair of canonical variates u_1 and v_1 solve the problem

$$\max_{\substack{v^T(\mathbf{X}^T\mathbf{X})v=1 \\ u^T(\mathbf{Y}^T\mathbf{Y})u=1}} u^T(\mathbf{Y}^T\mathbf{X})v,$$

a generalised SVD problem. Show that the solution is given by $u_1 = (\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}u_1^*$, and $v_1 = (\mathbf{X}^T\mathbf{X})^{-\frac{1}{2}}v_1^*$, where u_1^* and v_1^* are the leading left and right singular vectors in

$$(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}(\mathbf{Y}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-\frac{1}{2}} = \mathbf{U}^*\mathbf{D}^*\mathbf{V}^{*T}.$$

Show that the entire sequence $u_m, v_m, m = 1, \ldots, \min(K, p)$ is also given by (3.87).

Solution

The leading pair of canonical variates u_1 and v_1 must maximise:

$$\text{Corr}^2(\mathbf{Y}u, \mathbf{X}v) = \frac{u^T\mathbf{Y}^T\mathbf{X}v}{(u^T\mathbf{Y}^T\mathbf{Y}u)(v^T\mathbf{X}^T\mathbf{X}v)}.$$

Since any solution vectors u and v can be scaled so that the two quantities in the denominator equal one, the problem is equivalent to the generalised SVD problem given.

Take Cholesky decompositions $\mathbf{X}^T\mathbf{X} = \mathbf{K}^T\mathbf{K}$ and $\mathbf{Y}^T\mathbf{Y} = \mathbf{L}^T\mathbf{L}$ where $\mathbf{K}$ and $\mathbf{L}$ are square and invertible. We will prove a slightly modified version of the problem: $u_1 = \mathbf{L}^{-1}u_1^*$ and $v_1 = \mathbf{K}^{-1}v_1^*$, where the starred vectors are left and right singular vectors of $\mathbf{M} = (\mathbf{L}^T)^{-1}\mathbf{Y}^T\mathbf{X}\mathbf{K}^{-1}$. One can oneself that these two statements are essentially the same.

We begin by performing a change of basis by writing $u^* = \mathbf{L}u$ and $v^* = \mathbf{K}v$. In this basis our goal is to maximise

$$u^T\mathbf{Y}^T\mathbf{X}\mathbf{v} = (\mathbf{L}u)^T(\mathbf{L}^T)^{-1}\mathbf{Y}\mathbf{X}\mathbf{K}^{-1}(\mathbf{K}v) = u^{*T}\mathbf{M}v^*$$

subject to $\|u^*\| = \|v^*\| = 1$.

First note that this problem has a maximum since $u^{*T}\mathbf{M}v^*$ is a continuous function on the compact set $\{u^*, v^* \mid \|u^*\| = \|v^*\| = 1\}$. We will find it using Lagrange multipliers. Let

$$\mathcal{L}(u^*, v^*, \lambda, \mu) = u^{*T}\mathbf{M}v^T - \lambda(u^{*T}u - 1) - \mu(v^{*T}v - 1).$$

At a maximum of the constrained optimisation problem,

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial u^*} &= \mathbf{M}v^* - 2\lambda u^* = 0 \\ \frac{\partial \mathcal{L}}{\partial v^*} &= \mathbf{M}^T u^* - 2\mu v^* = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda} &= u^{*T}u - 1 = 0 \\ \frac{\partial \mathcal{L}}{\partial \mu} &= v^{*T}v - 1 = 0. \end{aligned}$$

Multiplying the first and second equations on the left by u^{*T} and v^{*T} respectively yields

$$u^{*T}\mathbf{M}v^* = 2\lambda \qquad \text{and} \qquad v^{*T}\mathbf{M}^T u^* = 2\mu,$$

which implies that $\lambda = \mu$. Moreover, this implies that u^* and v^* are a pair of singular vectors for $\mathbf{M}$ with singular value 2λ. To attain the global maximum of $u^{*T}\mathbf{M}v^* = 2\lambda$ they are must be the first singular vectors.

Now take $1 < m \leq \min(K, p)$ and suppose that we have shown that $u_1^*, \ldots, u_{m-1}^*$ and $v_1^*, \ldots, v_{m-1}^*$ are the $m - 1$ first singular vectors of $\mathbf{M}$. The pair (u_m^*, v_m^*) are the solution to the following constrained optimisation problem: maximise $u_m^{*T}\mathbf{M}v_m^*$ subject to $\|u_m^*\| = \|v_m^*\| = 1$ and

$$u_m^{*T}u_k^* = u_m^{*T}v_k^* = v_m^{*T}v_k^* = v_m^{*T}u_k^* = 0$$

for $k < m$.

We approach this problem much the same as above. First note that there is a maximum since we have a continuous function on a compact set. Now consider the function

$$\begin{aligned} \mathcal{L} = &\ u_m^{*T}\mathbf{M}v_m^T - \lambda(u_m^{*T}u_m - 1) - \mu(v_m^{*T}v_m - 1) \\ &+ \sum_{k=1}^{m-1}\left(\alpha_k^{YX}u_m^{*T}\mathbf{M}v_k^* - \alpha_k^{YY}u_m^{*T}u_k^* + \alpha_k^{XY}v_m^{*T}\mathbf{M}u_k^* - \alpha_k^{XX}v_m^{*T}v_k^*\right) \end{aligned}$$

Setting $\frac{\partial \mathcal{L}}{\partial u_m^*} = \frac{\partial \mathcal{L}}{\partial v_m^*} = 0$ yields

$$\sum_{k=1}^{m}\alpha_k^{YX}\mathbf{M}v_k^* = \sum_{k=1}^{m}\alpha_k^{YY}u_k^* \qquad \text{and} \qquad \sum_{k=1}^{m}\alpha_k^{XY}\mathbf{M}u_k^* = \sum_{k=1}^{m}\alpha_k^{XX}v_k^*,$$

where we have written $\alpha_m^{YX} = \alpha_m^{XY} = 1$, $\alpha_m^{YY} = 2\lambda$, and $\alpha_m^{XX} = 2\mu$ for notational convenience. Setting the derivatives with respect to the parameters equal to zero yields the original constraints of the problem.

Multiplying the two equations on the left by u_m^{*T} and v_m^{*T} respectively yields

$$u_m^{*T}\mathbf{M}v_m^* = 2\lambda \qquad \text{and} \qquad v_m^{*T}\mathbf{M}^T u_m^* = 2\mu,$$

so $\lambda = \mu$. On the other hand, multiplying the equations by u_k^{*T} and v_k^{*T} yields

$$\alpha_k^{YX}u_k^{*T}\mathbf{M}v_k^* = \alpha_k^{YY} \qquad \text{and} \qquad \alpha_k^{XY}v_k^{*T}\mathbf{M}^T u_k^* = \alpha_k^{XX}$$

so $\alpha_k^{YY} = d_k^*\alpha_k^{YX}$ and $\alpha_k^{XX} = d_k^*\alpha_k^{XY}$, where d_k^* is the kth singular value of $\mathbf{M}$.

Using these two sets of equations and the fact that $(u_1^*, v_1^*) \ldots, (u_{m-1}^*, v_{m-1}^*)$ are pairs of singular vectors, the equations from setting the partial derivatives of $\mathcal{L}$ equal to zero yield

$$\mathbf{M}v_m^* = 2\lambda u_m^* \qquad \text{and} \qquad \mathbf{M}^T u_m^* = 2\lambda v_m^*.$$

So (u_m^*, v_m^*) are a pair of singular vectors for $\mathbf{M}$. To maximise $u_m^{*T}\mathbf{M}v_m^* = 2\lambda$ subject to the orthogonality conditions, the must by the mth pair of singular vectors.

Exercise 3.21

Show that the solution to the reduced-rank regression problem (3.68), with Σ estimated by $\mathbf{Y}^T\mathbf{Y}/N$, is given by (3.69). *Hint:* Transform $\mathbf{Y}$ to $\mathbf{Y}^* = \mathbf{Y}\Sigma^{-\frac{1}{2}}$, and solve in terms of the canonical vectors u^*. Show that $\mathbf{U}_m = \Sigma^{-\frac{1}{2}}\mathbf{U}_m^*$, and a generalized inverse is $\mathbf{U}_m^- = \mathbf{U}_m^{*T}\Sigma^{\frac{1}{2}}$.

Exercise 3.22

Exercise 3.23

Consider a regression problem with all variables and response having mean zero and standard deviation one. Suppose also that each variable has identical absolute correlation with the response:

$$\frac{1}{N}|\langle \mathbf{x}_j, \mathbf{y} \rangle| = \lambda, \quad j = 1, \ldots, p.$$

Let $\hat{\beta}$ be the least-squares coefficient of $\mathbf{y}$ on $\mathbf{X}$, and let $\mathbf{u}(\alpha) = \alpha\mathbf{X}\hat{\beta}$ for $\alpha \in [0, 1]$ be the vector that moves a fraction α toward the least squares fit $\mathbf{u}$. Let RSS be the residual sum-of-squares from the full least squares fit.

(a) Show that

$$\frac{1}{N}|\langle \mathbf{x}_j, \mathbf{y} - \mathbf{u}(\alpha) \rangle| = (1 - \alpha)\lambda, \quad j = 1, \ldots, p$$

and hence the correlations of each $\mathbf{x}_j$ with the residuals remain equal in magnitude as we progress toward $\mathbf{u}$.

(b) Show that these correlations are all equal to

$$\lambda(\alpha) = \frac{1 - \alpha}{\sqrt{(1 - \alpha)^2 + \frac{\alpha(2 - \alpha)}{N} \cdot RSS}} \cdot \lambda,$$

and hence they decrease monotonically to zero.

(c) Use these results to show that the LAR algorithm in Section 3.4.4 keeps the correlations tied and monotonically decreasing, as claimed in (3.55).

Solution

(a) Take $j \in \{0, \ldots, p\}$ and let $\mathbf{e}_j$ denote the coordinate vector with a 1 in the jth position and 0s elsewhere. Then

$$\begin{aligned} \langle \mathbf{x}_j, \mathbf{u}(\alpha) \rangle &= \langle \mathbf{x}_j, \alpha\mathbf{X}\hat{\beta} \rangle \\ &= \alpha\langle \mathbf{X}\mathbf{e}_j, \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \rangle \\ &= \alpha\langle \mathbf{e}_j, \mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \rangle \\ &= \alpha\langle \mathbf{e}_j, \mathbf{X}^T\mathbf{y} \rangle \\ &= \alpha\langle \mathbf{X}\mathbf{e}_j, \mathbf{y} \rangle \\ &= \alpha\langle \mathbf{x}_j, \mathbf{y} \rangle \end{aligned}$$

so

$$\frac{1}{N}|\langle \mathbf{x}_j, \mathbf{y} - \mathbf{u}(\alpha) \rangle| = \frac{1}{N}|(1 - \alpha)\langle \mathbf{x}_j, \mathbf{y} \rangle| = (1 - \alpha)\lambda$$

as required.

(b) It will be useful in part (c) to drop the assumption that $\mathbf{y}$ has unit standard deviation and show that $\mathbf{x}_j$ and $\mathbf{y}$ have absolute correlation

$$\lambda(\alpha) = \frac{1 - \alpha}{\sqrt{\frac{(1-\alpha)^2}{N}\|\mathbf{y}\|^2 + \frac{\alpha(2-\alpha)}{N} \cdot RSS}} \cdot \lambda.$$

Since $\bar{\mathbf{y}} = 0$, the variance of $\mathbf{y}$ is $\|\mathbf{y}\|^2/N$ and so reduces to the formula in the question if $\mathbf{y}$ has standard deviation one.

Recall that the correlation between two variables is equal to their covariance over the product of their standard deviations. The standard deviation of each $\mathbf{x}_j$ is 1 so it suffices to show that the sample standard deviation of $\mathbf{y} - \mathbf{u}(\alpha)$ is equal to the denominator of the expression. Since $\bar{\mathbf{y}} = 0$ by assumption and $\overline{\mathbf{u}(\alpha)} = \alpha\sum\hat{\beta}_j\bar{x}_j = 0$, this is equivalent to showing

$$\frac{1}{N}\|\mathbf{y} - \alpha\hat{\mathbf{y}}\|^2 = (1 - \alpha)^2 + \frac{\alpha(2 - \alpha)}{N}RSS.$$

First note that $\mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to the column space of $\mathbf{X}$ so

$$\|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \langle \mathbf{y} - \hat{\mathbf{y}}, \mathbf{y} \rangle - \langle \mathbf{y} - \hat{\mathbf{y}}, \hat{\mathbf{y}} \rangle = \langle \mathbf{y} - \hat{\mathbf{y}}, \mathbf{y} \rangle.$$

Using this, we have

$$\begin{aligned} \|\mathbf{y} - \alpha\hat{\mathbf{y}}\|^2 &= \|\alpha(\mathbf{y} - \hat{\mathbf{y}}) + (1 - \alpha)\mathbf{y}\|^2 \\ &= \alpha^2\|\mathbf{y} - \hat{\mathbf{y}}\|^2 + 2\alpha(1 - \alpha)\langle \mathbf{y} - \hat{\mathbf{y}}, \mathbf{y} \rangle + (1 - \alpha)^2\|\mathbf{y}\|^2 \\ &= \alpha^2\|\mathbf{y} - \hat{\mathbf{y}}\|^2 + 2\alpha(1 - \alpha)\|\mathbf{y} - \hat{\mathbf{y}}\|^2 + (1 - \alpha)^2\|\mathbf{y}\|^2, \end{aligned}$$

which establishes the formula for $\lambda(\alpha)$.

Finally, one can show that

$$\lambda'(\alpha) = \frac{-RSS\lambda}{N\left(\frac{(1-\alpha)^2}{N}\|\mathbf{y}\|^2 + \frac{\alpha(2-\alpha)}{N} \cdot RSS\right)^{3/2}} < 0$$

so that $\lambda(\alpha)$ decreases monotonically to zero.

(c) Start with standardised predictors with mean zero and unit variance.

Take $j \in \{1, \ldots, p\}$ and assume that at the start of step j we have a residual vector $\mathbf{r}_j$ which has mean zero and identical absolute correlation λ_j with $\mathbf{x}_1, \ldots, \mathbf{x}_j$. Let $\mathbf{u}_j(\alpha)$ for $\alpha \in [0, 1]$ be the vector that moves a fraction α towards the least squares fit of $\mathbf{x}_1, \ldots, \mathbf{x}_j$ to $\mathbf{r}_j$ (for the first step we set $\mathbf{r}_1 = \mathbf{y} - \bar{\mathbf{y}}$ and assume wlog that $\mathbf{x}_1$ has the greatest absolute correlation with $\mathbf{r}_1$).

Take $\alpha_j \in [0, 1]$ minimal such that a predictor $\mathbf{x}_k$ with $k > j$ has correlation with $\mathbf{r}_j - \mathbf{u}_j(\alpha_j)$ equal to

$$\lambda_j(\alpha_j) = \frac{(1 - \alpha_j)\lambda_j}{\sqrt{\frac{1}{N}\left((1 - \alpha_j)^2\|\mathbf{r}_j\|^2 + \alpha_j(2 - \alpha_j)RSS\right)}},$$

where RSS denotes the residual sum of squares between $\mathbf{r}_j$ and its least squares fit (if there is no such k then we're done). Wlog assume $k = j + 1$ and set $\mathbf{r}_{j+1} = \mathbf{r}_j - \mathbf{u}_j(\alpha_j)$ and $\lambda_{j+1} = \lambda_j(\alpha_j)$. By (b) we are now in the right position to start step $j + 1$.

Assuming that $\mathbf{X}$ has full column rank this process will terminate at the pth step where it reaches the least squares solution. This gives us a piecewise linear path $\mathbf{u}(\alpha)$ for $\alpha \in [0, \sum_0^p \alpha_j]$ with $\mathbf{u}(0) = \mathbf{y}$, $\mathbf{u}(\sum_0^p \alpha_j) = \hat{\beta}$ and

$$\mathbf{u}\left(\sum_{j=1}^{k}\alpha_j + t\right) = \mathbf{r}_k + \mathbf{u}_j(t)$$

for $t \in [0, \alpha_{j+1}]$. By part (a) the correlations stay tied and decreasing along this path

Exercise 3.24

LAR directions. Using the notation around equation (3.55) on page 74, show that the LAR direction makes an equal angle with each of the predictors in $\mathcal{A}_k$.

Solution

The angle between $\mathbf{u}_k$ and $\mathbf{x}_j$ is

$$\arccos\left(\frac{|\langle \mathbf{x}_j, \mathbf{u}_k \rangle|}{\|\mathbf{x}_j\|\|\mathbf{u}_k\|}\right)$$

so the smallest angle corresponds to the greatest value of $\frac{|\langle \mathbf{x}_j, \mathbf{u}_k \rangle|}{\|\mathbf{x}_j\|\|\mathbf{u}_k\|}$, that is the greatest absolute correlation. But by exercise 3.23 elements of $\mathcal{A}_k$ have equal absolute correlation with $\mathbf{u}_k$, which is maximal among $\mathbf{x}_j$.

Exercise 3.25

LAR look-ahead (Efron et al., 2004, Sec. 2). Starting at the beginning of the kth step of the LAR algorithm, derive expressions to identify the next variable to enter the active set at step $k + 1$, and the value of α at which this occurs (using the notation around equation (3.55) on page 74).

Solution

Take $\mathbf{x}_a \in \mathcal{A}_k$ and $\mathbf{x}_b \notin \mathcal{A}_k$. We wish to identify the values of α such that the two vectors have the same absolute correlation with $\mathbf{r}_k - \alpha\mathbf{u}_k$. This is the case if and only if

$$\frac{|\langle \mathbf{x}_a, \mathbf{r}_k - \alpha\mathbf{u}_k \rangle|}{\|\mathbf{r}_k - \alpha\mathbf{u}_k\|} = \frac{|\langle \mathbf{x}_b, \mathbf{r}_k - \alpha\mathbf{u}_k|}{\|\mathbf{r}_k - \alpha\mathbf{u}_k\rangle\|} \quad \Longleftrightarrow \quad |\langle \mathbf{x}_a, \mathbf{r}_k - \alpha\mathbf{u}_k \rangle| = |\langle \mathbf{x}_b, \mathbf{r}_k - \alpha\mathbf{u}_k \rangle|.$$

This has two solutions:

$$\alpha = \frac{\langle \mathbf{x}_a \pm \mathbf{x}_b, \mathbf{r}_k \rangle}{\langle \mathbf{x}_a \pm \mathbf{x}_b, \mathbf{u}_k \rangle}.$$

So we can identify the next variable to enter the active set as follows. Fix $\mathbf{x}_a \in \mathcal{A}_k$ and calculate the two α values above for each $\mathbf{x}_b \notin \mathcal{A}_k$. The $\mathbf{x}_b$ with the greatest α value lying in $[0, 1]$ is the next to join the active set and it will occur at that value of α.

Exercise 3.26

Forward stepwise regression enters the variable at each step that most reduces the residual sum-of-squares. LAR adjusts variables that have the most (absolute) correlation with the current residuals. Show that these two entry criteria are not necessarily the same. [Hint: let $\mathbf{x}_{j.\mathcal{A}}$ be the jth variable, linearly adjusted for all the variables currently in the model. Show that the first criterion amounts to identifying the j for which $\text{Cor}(\mathbf{x}_{j.\mathcal{A}}, \mathbf{r})$ is largest in magnitude.]

Solution

I believe that the statement in the hint was more-or-less shown in Exercise 3.9. Regress $\mathbf{x}_j$ on the variables currently in the model and let $\mathbf{x}_{j.\mathcal{A}}$ denote the residual. In Exercise 3.9 we showed that the j with the greatest $|\mathbf{x}_{j.\mathcal{A}}^T\mathbf{y}|/\|\mathbf{x}_{j.\mathcal{A}}\|$ will reduce the residual sum-of-squares the most. Since by construction $\mathbf{x}_{j.\mathcal{A}}$ is orthogonal the variables in the active set, $\mathbf{x}_{j.\mathcal{A}}^T\mathbf{y} = \mathbf{x}_{j.\mathcal{A}}^T\mathbf{r}$ and so this is the same as identifying the j with the greatest value of $|\text{Cor}(\mathbf{x}_{j.\mathcal{A}}, \mathbf{r})|$.

To illustrate how these criteria could lead to different results consider the situation where the current active set consists of a single variable $\mathbf{x}_k$. Then

$$\mathbf{r} = \mathbf{y} - \frac{\langle \mathbf{x}_k, \mathbf{y} \rangle}{\langle \mathbf{x}_k, \mathbf{x}_k \rangle}\mathbf{x}_k, \qquad \text{and} \qquad \mathbf{x}_{j.\mathcal{A}} = \mathbf{y} - \frac{\langle \mathbf{x}_k, \mathbf{x}_j \rangle}{\langle \mathbf{x}_k, \mathbf{x}_k \rangle}\mathbf{x}_k \qquad \text{for all } j \neq k.$$

A little calculation shows that

$$\text{Corr}(\mathbf{x}_{j.\mathcal{A}}, \mathbf{r}) = \left(1 - \text{Corr}^2(\mathbf{x}_j, \mathbf{x}_k)\right)\text{Corr}(\mathbf{x}_j, \mathbf{r}).$$

So if $\mathbf{x}_j$ is highly correlated with $\mathbf{x}_k$ then it may be included in the active set under the LAR criterion but not the forward stepwise criterion.

Exercise 3.27

Lasso and LAR: Consider the lasso problem in Lagrange multiplier form: with $L(\beta) = \frac{1}{2}\sum_i(y_i - \sum_j x_{ij}\beta_j)^2$, we minimise

$$L(\beta) + \lambda\sum_j|\beta_j|$$

for fixed $\lambda > 0$.

(a) Setting $\beta_j = \beta_j^+ - \beta_j^-$ with $\beta_j^+, \beta_j^- \geq 0$, expression (3.88) [above] becomes $L(\beta) + \lambda\sum_j(\beta_j^+ + \beta_j^-)$. Show that the Lagrange dual function is

$$L(\beta) + \lambda\sum_j(\beta_j^+ + \beta_j^-) - \sum_j\lambda_j^+\beta_j^+ - - \sum_j\lambda_j^-\beta_j^-$$

and the Karush-Kuhn-Tucker optimality conditions are

$$\begin{aligned} \nabla L(\beta)_j + \lambda - \lambda_j^+ &= 0 \\ -\nabla L(\beta)_j + \lambda - \lambda_j^- &= 0 \\ \lambda_j^+\beta_j^+ &= 0 \\ \lambda_j^-\beta_j^- &= 0, \end{aligned}$$

along with the non-negativity constraints on the parameters and all the Lagrange multipliers.

(b) Show that $|\nabla L(\beta)_j| \leq \lambda\ \forall j$, and that the KKT conditions imply one of the following three scenarios:

$$\begin{aligned} \lambda = 0 \quad &\Rightarrow \quad \nabla L(\beta)_j = 0\ \forall j \\ \beta_j^+ > 0,\ \lambda > 0 \quad &\Rightarrow \quad \lambda_j^+ = 0,\ \nabla L(\beta)_j = -\lambda < 0,\ \beta_j^- = 0 \\ \beta_j^- > 0,\ \lambda > 0 \quad &\Rightarrow \quad \lambda_j^- = 0,\ \nabla L(\beta)_j = \lambda > 0,\ \beta_j^+ = 0. \end{aligned}$$

Hence show that for any "active" predictor having $\beta_k \neq 0$, we must have $\nabla L(\beta)_j = -\lambda$ if $\beta_j > 0$, and $\nabla L(\beta)_j = \lambda$ if $\beta_j < 0$. Assuming the predictors are standardized, relate λ to the correlation between the jth predictor and the current residuals.

(c) Suppose that the set of active predictors is unchanged for $\lambda_0 \geq \lambda \geq \lambda_1$. Show that there is a vector γ_0 such that

$$\hat{\beta}(\lambda) = \hat{\beta}(\lambda_0) - (\lambda - \lambda_0)\gamma_0.$$

Thus the lasso solution path is linear as λ ranges from λ_0 to λ_1 (Efron et al., 2004; Rosset and Zhu, 2007).

In [ ]: