SOLUTIONS TO ELEMENTS OF STATISTICAL LEARNING

J. WILSON PEOPLES

Contents

1. Introduction

2. Solutions to Chapter 2

Exercise 2.1. *Suppose each of K classes has an associated target t_k, which is a vector of all zeros, except a 1 in the k-th position. Show that classifying to the largest element of $\hat{y}$ amounts to choosing the closest target $\min_k \|t_k - \hat{y}\|$, if the elements of $\hat{y}$ sum to 1.*

Solution. The elements of $\hat{y}$ must be nonnegative and sum to 1. Hence, each entry of $\hat{y}$ is less than or equal to 1. It follows immediately that the largest entry of $\hat{y}$ will be the closest entry to 1. ■

Exercise 2.2. *Show how to compute the Bayes Decision boundary for the simulation example in Figure 2.5.*

Comment. The sampling procedure used to obtain the points in Figure 2.5 is described in the following sentences (summarized from paragraph 3 of Section 2.3.3 of [2]). First, 10 means (points, to be used as means later) $b_1, \ldots, b_k$ were drawn randomly from $\mathcal{N}\left((1,0)^\top, \mathbf{I}\right)$ and labeled BLUE. Then, 10 means $o_1, \ldots, o_{10}$ were drawn from $\mathcal{N}\left((0,1)^\top, \mathbf{I}\right)$ and were labeled $ORANGE$. Then, 100 blue points were generated as follows. Choose a b_k uniformly from $\{b_1, \ldots, b_{10}\}$, then choose a point from $\mathcal{N}(b_k, \mathbf{I}/5)$. 100 orange points were obtained similarly, according to the orange means $o_1, \ldots, o_k$, with the same variance. ■

Solution. By definition, the Bayes decision boundary for this example is

$$\left\{(x,y) \in \mathbb{R}^2 : \mathbb{P}\left((x,y) \in BLUE\right) = \mathbb{P}\left((x,y) \in ORANGE\right)\right\}.$$

Date: April 26, 2023.

Hence, this is given by the set of points $\mathbf{x} = (x, y)^\top$ satisfying

$$\frac{1}{10}\sum_{k=1}^{10}\frac{1}{2\pi\sqrt{\det(\mathbf{I}/5)}}\exp\left(\frac{1}{2}\left(\langle\mathbf{x}-b_k, 5\mathbf{I}(\mathbf{x}-b_k)\rangle\right)\right)$$
$$=\frac{1}{10}\sum_{k=1}^{10}\frac{1}{2\pi\sqrt{\det(\mathbf{I}/5)}}\exp\left(\frac{1}{2}\left(\langle\mathbf{x}-o_k, 5\mathbf{I}(\mathbf{x}-o_k)\rangle\right)\right).$$

We can cancel all constants outside of exp and use that the inside is simply $\frac{5}{2}(\|\mathbf{x}-b_k\|^2)$ to obtain that the decision boundary can be calculated as the solution to

$$\sum_{k=1}^{10}\exp\left(\frac{5}{2}\|\mathbf{x}-b_k\|^2\right)=\sum_{k=1}^{10}\exp\left(\frac{5}{2}\|\mathbf{x}-o_k\|^2\right).$$

Given $\{b_k\}, \{o_k\}$, one can numerically compute the solutions to the above equation. ■

Exercise 2.3. *Derive equation (2.24)*

Comment. (2.24) describes the median distance from the origin to the closest data point, where N points are drawn uniformly from a p-dimensional ball centered at the origin. (2.24) is given by

$$d(p,N)=\left(1-\left(\frac{1}{2}\right)^{1/N}\right)^{1/p}.$$

■

Solution. Let X_N denote the random variable representing the distance to closest point in this setting. By definition of median, we wish to find d such that

$$\mathbb{P}\left(X_N \le d\right), \mathbb{P}\left(X_N \ge d\right) = \frac{1}{2}.$$

If the closest point were a distance d away, then N points were drawn from a volume of $C_p - C_p d^p$, where C_p denotes the constant such that $C_p r^p$ is the volume of a p-ball with radius r. Since the trials are independent, this occurs with probability

$$\left(1-\frac{C_p d^p}{C_p 1^p}\right)^N = (1-d^p)^N.$$

Setting this equal to $1/2$ and solving yields the desired result. ■

Exercise 2.4. *The edge affect problem discussed on page 23 is not peculiar to uniform sampling from bounded domains. Consider inputs drawn from a spherical multinormal distribution $X \sim \mathcal{N}(0, \mathbf{I}_p)$. The squared distance from any sample point to the origin has χ^2_p distribution with mean p. Consider a prediction point x_0 drawn from this distribution, and let $a = x_0/\|x_0\|$ be an associated unit vector. Let $z_i = a^\top x_i$ be the projection of each training point along a.*
Show that the z_i are distributed $\mathcal{N}(0,1)$ with expected squared distance from the origin 1, while the target point has expected squared distance p from the origin.
Hence for $p = 10$, a randomly drawn test point is about 3.1 standard deviations from the origin, while all other training points are on average one standard deviation along a. So most prediction points see themselves as lying on the edge of the training data.

Solution. After dividing by $\|x_0\|$, it is WLOG assume $\|x_0\|^2 = 1$. Since $x_i = ((x_i)_1, \ldots, (x_i)_p)^\top \sim \mathcal{N}(0, \mathbf{I}_p)$, we have each $(x_i)_j \sim \mathcal{N}(0,1)$. Hence,

$$z_i = \sum_{j=1}^{p} (x_0)_j (x_i)_j$$

is normal, with mean $\mathbb{E}(z_i) = \sum_{j=1}^{p} (x_0)_j \mathbb{E}(x_i)_j = 0$, and variance $\sum_{j=1}^{p} (x_0)_j^2 \cdot 1^2 = 1$ (since $\|x_0\| = 1$). This shows $z_i \sim \mathcal{N}(0,1)$. But as is stated, the expected squared distance of x_i is p. Plugging in for $p = 10$, a random point is a distance of 10 away from the origin, which is about $3.1 \times \sqrt{10} \approx 10$. I.e., a random point is about 3.1 standard deviations from the origin, but its projection along x_0 has an expected squared distance of 1. ■

Exercise 2.5. *(a) Derive equation (2.27). The last line makes use of (3.8) through a conditioning argument.*
(b) Derive (2.28), making use of the cyclic property of the trace operator, and its linearity.

Comment. Note that the setup for these equations is the following. We are told that the true solution is linear:

$$Y = X^\top \beta + \epsilon,$$

where X is a random variable in $\mathbb{R}^p$, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The model we fit by is linear, minimizing square distance. We denote by $\hat{y}_0$ our prediction point given an arbitrary tst point x_0. We organize all the training data-points $\{x_1, \ldots, x_N\}$ into an $N \times p$ matrix $\mathbf{X}$. Note that $\hat{y}_0$ is given by $\hat{y}_0 = x_0^T \hat{\beta} = x_0^\top \beta + \sum_{i=1}^{N} \ell_i(x_0)\epsilon_i$, where ϵ_i is the ϵ associated to the i-th data point and output y_i, and $\ell_i(x_0)$ denotes the i-th component of the vector $\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} x_0$. This can also be written

$$\hat{y}_0 = x_0^\top \beta + x_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \vec{\epsilon},$$

where $\vec{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)^\top$. *EPE* at a datapoint x_0 is defined to be $\mathbb{E}_{\mathcal{T}}(\mathbb{E}((y_0 - \hat{y}_0)^2 | x_0, \mathcal{T}))$, where $\mathcal{T}$ denotes the set of training data. ■

Solution.
(a) Notice that

$$(y_0 - \hat{y}_0)^2 = \left((y_0 - x_0^\top \beta)^2 - (x_0^\top \beta - \hat{y}_0)\right)^2.$$

Consider the conditional expectation of the above given $x_0, \mathcal{T}$. Expanding, we obtain

$$\begin{aligned} \mathbb{E}((y_0 - \hat{y}_0)^2 | \mathcal{T}, x_0) &= \mathbb{E}((y_0 - x_0^\top \beta)^2 | \mathcal{T}, x_0) \\ &+ 2\mathbb{E}(y_0 - x_0^\top \beta | \mathcal{T}, x_0)\mathbb{E}(x_0^\top \beta - \hat{y}_0 | \mathcal{T}, x_0) \\ &+ \mathbb{E}((x_0^\top \beta - \hat{y}_0)^2 | \mathcal{T}, x_0). \end{aligned}$$

Notice that $(y_0 - x_0^\top \beta)^2 = (x_0^\top \beta + \epsilon - x_0^\top \beta)^2 = \epsilon^2$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. After taking the expectation w.r.t $\mathcal{T}$, we see that the first term is by definition σ^2. Similarly, the second term, after taking the expectation w.r.t. $\mathcal{T}$, is 0, since $\mathbb{E}_{\mathcal{T}}\mathbb{E}(y_0 - x_0^\top \beta | \mathcal{T}, x_0) = \mathbb{E}_{\mathcal{T}}\mathbb{E}(\epsilon | \mathcal{T}, x_0) = 0$, and the two cross terms are independent (Indeed, ϵ, the noise from new observation y_0, is independent from $\sum_{i=1}^{N} \ell_i(x_0)\epsilon_i$, errors on the training data). For the final term, $\hat{y}_0$ is an unbiased estimator of $x_0^\top \beta$. Indeed, $\left(\mathbb{E}_{\mathcal{T}}(\hat{y}_0) - x_0^\top \beta\right)^2 = \left(\mathbb{E}_{\mathcal{T}} \sum_{i=1}^{N} \ell_i(x_0)\epsilon_i + \mathbb{E}_{\mathcal{T}}(x_0^\top \beta) - x_0^\top \beta\right)^2 = (0 + x_0^\top \beta - x_0^\top \beta)^2 = 0$. Hence, using the standard variance-bias decomposition, we see that

the final term is $\mathrm{Var}(\hat{y}_0|x_0,\mathcal{T})$. Taking the epxectation w.r.t. $\mathcal{T}$ and putting everything together, we finally see that

$$EPE(x_0) = \sigma^2 + \mathrm{Var}_{\mathcal{T}}(\hat{y}_0). \tag{2.1}$$

It remains to compute the variance $\mathrm{Var}_{\mathcal{T}}(\hat{y}_0)$. For convenience, denote $\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}$ by

$$\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} := A_{\mathbf{X}} \in \mathbb{R}^{N\times p}.$$

Notice that $\mathbb{E}_{\mathcal{T}}(\hat{y}_0) = x_0^\top\beta + \mathbb{E}_{\mathcal{T}}\left((A_{\mathbf{X}}x_0)^\top\vec{\epsilon}\right) = x_0^\top\beta + 0$ (since the expectation w.r.t. $\mathcal{T}$ contains an expectation over ϵ, which vanishes, as well as one over $\mathbf{X}$). Hence,

$$\mathrm{Var}_{\mathcal{T}}(\hat{y}_0) = \mathbb{E}_{\mathcal{T}}\left(x_0^\top\beta + (A_{\mathbf{X}}x_0)^\top\vec{\epsilon} - x_0^\top\beta\right)^2 = \mathbb{E}_{\mathcal{T}}\left((A_{\mathbf{X}}x_0)^\top\vec{\epsilon}\right)^2 = \mathbb{E}_{\mathcal{T}}\left(x_0^\top A_{\mathbf{X}}^\top\vec{\epsilon}\vec{\epsilon}^\top A_{\mathbf{X}}x_0\right).$$

Taking the expectation associated to the noise first, we get $\sigma^2\mathbf{I}_{N\times N}$. Plugging the above into Equation (2.1), we obtain

$$EPE(x_0) = \sigma^2 + \mathbb{E}_{\mathcal{T}}\left(x_0^\top(\mathbf{X}^\top\mathbf{X})^{-1}x_0\right)\sigma^2,$$

which was to be demonstrated.

(b) Using that $\mathbf{X}^\top\mathbf{X} \to N\mathrm{Cov}(X)$ for large N, we have that

$$EPE(x_0) \approx \sigma^2 + \frac{x_0^\top\mathrm{Cov}(X)x_0}{N}\sigma^2 = \sigma^2 + \frac{\sigma^2}{N}\mathrm{tr}\left(x_0^\top Cov(X)^{-1}x_0\right),$$

where we have inserted trace onto a 1×1 quantity. Using that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, we have $\mathrm{tr}(x_0^\top(Cov(X))^{-1}x_0) = \mathrm{tr}(x_0x_0^\top Cov(X)^{-1})$. By linearity of the trace and expectation, we can exchange $\mathbb{E}_{x_0}$ with tr. Hence, taking the expectation, since $\mathbb{E}_{x_0}x_0 = 0$, we have

$$\mathbb{E}_{x_0}EPE(x_0) \approx \sigma^2 + \frac{\sigma^2}{N}\mathrm{tr}\left(Cov(x_0)Cov(X)^{-1}\right).$$

Using that $Cov(x_0)(Cov(X)^{-1}) = \mathbf{I}_p$, the $p\times p$ identity matrix, we see that

$$\mathbb{E}_{x_0}EPE(x_0) \approx \sigma^2 + \frac{p\sigma^2}{N}.$$

■

Exercise 2.6. *Consider a regression problem with inputs x_i and outputs y_i, and a parameterized model $f_\theta(x)$ to be fit by least squares. Show that if there are observations with tied or identical values of x, then the fit can be obtained from a reduced weighted least squares problem.*

Solution. Given N data points, fitting by least squares amounts to minimizing

$$\min_\theta \sum_{i=1}^{N}(y_i - f_\theta(x_i))^2$$

Suppose out of the N inputs $\{x_i\}$, there are $N' < N$ unique inputs. For each unique $j = 1,\dots,N'$, there are k_j possible different outputs, where $\sum_{j=1}^{N'}k_j = N$. Denote by x'_{ij}

the j-th copy of the i-th unique input (i.e., $1 \leq i \leq N'$, $1 \leq j \leq k_i$). Similarly, denote by y'_{ij} the true output for x_{ij}. Then we can rewrite the above minimization problem as

$$\min_\theta \sum_{i=1}^{N'} \sum_{j=1}^{k_i} (y'_{ij} - f_\theta(x_{ij}))^2 = \min_\theta \sum_{i=1}^{N'} \left(k_i f_\theta^2(x_i) + \sum_{j=1}^{k_i} (y'_{ij})^2 - \sum_{j=1}^{k_i} 2 y'_{ij} f_\theta(x_i) \right).$$

Notice we can factor out k_i to obtain that the above is equal to

$$\min_\theta \sum_{i=1}^{N'} k_i \left(f_\theta^2(x_i) + \frac{1}{k_i} \sum_{j=1}^{k_i} (y'_{ij})^2 - 2 f_\theta(x_i) \cdot \frac{1}{k_i} \sum_{j=1}^{k_i} y'_{ij} \right).$$

Define $\hat{y}_i = \frac{1}{k_i} \sum_{j=1}^{k_i} y'_{ij}$. Adding and subtracting $\hat{y}_i^2$, we obtain the above is equal to

$$\min_\theta \left(\sum_{i=1}^{N'} k_i \left(f_\theta(x_i) - \hat{y}_i \right)^2 + \sum_{i=1}^{N'} k_i \left(\frac{1}{k_i} \sum_{j=1}^{k_i} (y'_{ij})^2 - \hat{y}_i^2 \right) \right).$$

The second term does not depend on θ, and so minimizing the above is equivalent to minimizing

$$\min_\theta \sum_{i=1}^{N'} k_i \left(f_\theta(x_i) - \hat{y}_i \right)^2.$$

This is a least squares error from the average of the outputs over the unique inputs, but it is weighted by the number of times each input occurs. This completes the problem. ■

Exercise 2.7. *Suppose we have a sample of N pairs x_i, y_i drawn i.i.d. from the distribution characterized as follows.*

$$x_i \sim h(x), \text{ the design density}$$
$$y_i = f(x_i) + \epsilon_i, f \text{ is the regression function}$$
$$\epsilon_i \sim (0, \sigma^2)$$

We construct an estimator for f linear in the y_i:

$$\hat{f}(x_0) = \sum_{i=1}^{N} \ell_i(x_0; \mathcal{X}) y_i$$

where the weights $\ell_i(x_0; \mathcal{X})$ do not depend on the y_i, but do depend on the entire training sequence of x_i, denoted here by $\mathcal{X}$.

(a) Show that linear regression and k-nearest neighbor regression are members of this class of estimators. Describe explicitly the weights $\ell_i(x_0; \mathcal{X})$ in each of these cases.

(b) Decompose the conditional mean squared error

$$\mathbb{E}_{\mathcal{Y}|\mathcal{X}} (f(x_0) - \hat{f}(x_0))^2$$

into a conditional bias squared and conditional variance component. Like $\mathcal{X}$, here $\mathcal{Y}$ represents the entire training sequence y_i.

(c) Decompose the (unconditional) MSE

$$\mathbb{E}_{\mathcal{X},\mathcal{Y}}(f(x_0) - \hat{f}(x_0))^2$$

into a squared bias and variance component.

(d) Establish a relationship between the squared biases and variances in the above two cases.

Solution.
(a) Linear regression, given training data $\mathcal{X}, \mathcal{Y}$, by definition estimates f with a function $\hat{f}(x_0) = x_0^\top \hat{\beta}$, where

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

where $\mathbf{X}$ is an $N \times p$ matrix with i-th row $x_i \in \mathbb{R}^p$, and $\mathbf{y} \in \mathbb{R}^N$ with i-th entry y_i. Note that $x_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ is a $1 \times N$. Rewriting the matrix notation as a sum, we see that

$$\hat{f}(x_0) = \sum_{i=1}^{N} \left(x_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \right)_{1,i} y_i.$$

This shows that linear regression is of the above form, with $\ell_i(x_0; \mathcal{X}) = \left(x_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \right)_{1,i}$. We emphasize that $\ell_i(x_0; \mathcal{X})$ does not depend on y_i, as required. For $k-NN$ regression, the approximator $\hat{f}$ is chosen as

$$\hat{f}(x_0) = \frac{1}{k} \sum_{x_i \in N_k(x_0)} y_i,$$

where $N_k(x_0)$ denotes the set of k elements of $\mathcal{X}$ closest to x_0. Rewriting this, we see that

$$\hat{f}(x_0) = \sum_{i=1}^{N} \frac{1}{k} \chi_{N_k(x_0)}(x_i) y_i.$$

where $\chi_A(x)$ denotes the characteristic function of set A. This way, we see $k-NN$ regression is of the above form as well, where $\ell_i(x_0; \mathcal{X}) = \frac{1}{k} \chi_{N_k(x_0)}(x_i)$, which again depends only on x_0 and $\mathcal{X}$.

(b) Adding and subtracting the conditional expectation $\mathbb{E}[\hat{f}(x_0)|\mathcal{X}]$,

$$\mathbb{E}\left[(f(x_0) - \hat{f}(x_0))^2 \middle| \mathcal{X} \right] = \mathbb{E}\left[(f(x_0) - \mathbb{E}[\hat{f}(x_0)|\mathcal{X}])^2 \middle| \mathcal{X} \right] + \mathbb{E}\left[(\mathbb{E}[\hat{f}(x_0)|\mathcal{X}] - \hat{f}(x_0))^2 \middle| \mathcal{X} \right]$$
$$+2\mathbb{E}\left[(f(x_0) - \mathbb{E}[\hat{f}(x_0)|\mathcal{X}])(\mathbb{E}[\hat{f}(x_0)|\mathcal{X}] - \hat{f}(x_0) \middle| \mathcal{X} \right].$$

Looking at the first term, we see that these are constants and can be taken out of the conditional expectation. The first term becomes $(f(x_0) - \mathbb{E}[\hat{f}(x_0)|\mathcal{X}])^2 = \text{Bias}(\hat{f}(x_0)|\mathcal{X})^2$. The second term already reads as conditional variance. For the final term, since $\mathbb{E}[\hat{f}(x_0)|\mathcal{X}]$ is measurable w.r.t. $\mathcal{X}$, and $f(x_0)$ is fixed, we use the taking out what is known property of conditional expectation to obtain

$$2\mathbb{E}\left[(f(x_0) - \mathbb{E}[\hat{f}(x_0)|\mathcal{X}])(\mathbb{E}[\hat{f}(x_0)|\mathcal{X}] - \hat{f}(x_0) \middle| \mathcal{X} \right]$$

$$= 2(f(x_0) - \mathbb{E}[\hat{f}(x_0)|\mathcal{X}])\mathbb{E}\left[\mathbb{E}[\hat{f}(x_0)|\mathcal{X}] - \hat{f}(x_0)\middle|\mathcal{X}\right].$$

Then, by using linearity of conditional expectation, we see that the above term vanishes. We are left with

$$\text{Bias}(\hat{f}(x_0)|\mathcal{X})^2 + \text{Var}(\hat{f}(x_0)|\mathcal{X}),$$

which was to be demonstrated.

(c) Let us to the same computation as before, only this time we are taking the expectation over $\mathcal{X}, \mathcal{Y}$. We have

$$MSE(f)(x_0) = \mathbb{E}\left[(f(x_0) - \mathbb{E}[\hat{f}(x_0)])^2\right] + \mathbb{E}\left[(\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))^2\right]$$

$$+2\mathbb{E}\left[(f(x_0) - \mathbb{E}[\hat{f}(x_0)])(\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))\right].$$

where the expectations are taken over $\mathcal{X}, \mathcal{Y}$. Notice that $f(x_0) - \mathbb{E}_{\mathcal{X},\mathcal{Y}}[\hat{f}(x_0)]$ is a constant, and can be pulled out of the expectation. Linearity then shows that the cross term vanishes. Similarly, since $(f(x_0) - \mathbb{E}_{\mathcal{X},\mathcal{Y}}[\hat{f}(x_0)])^2$ is constant, we can pull it out of the expectation. This is the Bias^2 term. The second term again is already written as variance. This completes the proof.

(d) Taking the expectation over $\mathcal{X}$ of the conditioned MSE yields the (unconditional) MSE, by the expectation property of conditional expectation. But in fact, comparing individual terms, it is easy to see that

$$\mathbb{E}_{\mathcal{X}}\text{Bias}(\hat{f}(x_0)|\mathcal{X})^2 = \text{Bias}^2(\hat{f}(x_0)),$$

which in turn shows that

$$\mathbb{E}_{\mathcal{X}}\text{Var}(\hat{f}(x_0)|\mathcal{X}) = \text{Var}(\hat{f}(x_0)).$$

■

Exercise 2.8. *Coding exercise*

Solution. ■

Exercise 2.9. *Consider a linear regression model with p parameters, fit by least squares to a set of training data $(x_1, y_1), \ldots, (x_N, y_N)$ drawn at random from a population. Let $\hat{\beta}$ be the least squares estimate. Suppose we have some test data $(\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_M, \tilde{y}_M)$ drawn at random from the same population as the training data. If $R_{tr}(\hat{\beta}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\beta}^\top x_i)^2$, and $R_{te}(\hat{\beta}) = \frac{1}{M}\sum_{i=1}^{M}(\tilde{y}_i - \hat{\beta}^\top \tilde{x}_i)^2$, show that*

$$\mathbb{E}\left[R_{tr}(\hat{\beta})\right] \leq \mathbb{E}\left[R_{te}(\hat{\beta})\right],$$

where the expectations are over all that is random in each expression.

Solution. Notice that $\hat{\beta}$ can be viewed as a function of the test data $\hat{\beta}((x_1, y_1), \ldots, (x_N, y_N)) \in \mathbb{R}^p$ such that

$$R_{\text{tr}}(\hat{\beta}) \leq R_{\text{tr}}(\beta)$$

almost surely for any set of data points $(x_1, y_1), \dots (x_N, y_N)$ drawn i.i.d., and *any* vector β. Hence, the inequality holds taking the expectation over all testing data:

$$\mathbb{E}\left[R_{\text{tr}}(\hat{\beta})\right] \leq \mathbb{E}\left[R_{\text{tr}}(\beta)\right].$$

By the i.i.d. assumption, note that $\mathbb{E}\left[R_{\text{tr}}(\beta)\right] = \mathbb{E}\left[(y_i - \beta^\top x_i)^2\right]$, for any $i = 1, \dots, N$. Particular choice of β yields

$$\mathbb{E}\left[R_{\text{tr}}(\hat{\beta})\right] \leq \mathbb{E}\left[(\tilde{y}_j - \hat{\beta}^\top \tilde{x}_j)^2\right].$$

Indeed, one can check that $\beta = \frac{1}{\|x_i\|^2}\left(y_i x_i - \tilde{y}_j x_i + x_i \tilde{x}_j^\top \hat{\beta}\right)$ works. Since the testing data is also i.i.d., we have

$$\mathbb{E}\left[R_{\text{tr}}(\hat{\beta})\right] \leq \mathbb{E}\left[(\tilde{y}_j - \hat{\beta}^\top \tilde{x}_j)^2\right] = \frac{1}{M}\sum_{j=1}^{M} \mathbb{E}\left[(\tilde{y}_j - \hat{\beta}^\top \tilde{x}_j)^2\right] = \mathbb{E}\left[R_{\text{te}}(\hat{\beta})\right].$$

This completes the problem. ■

3. Solutions to Chapter 3

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).*

Comment. The equations referred to above are given by the following. The z-score is given by

$$z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}, \tag{3.12}$$

where $\hat{\beta}_j$ is the least-squares approximation to β_j, $\hat{\sigma}$ is an estimation of the variance of the noise, and v_j is the j-th diagonal element of $(\mathbf{X}^\top \mathbf{X})^{-1}$. This is used to test the hypothesis that the coefficient β_j is 0. A z-score with large absolute value implies that this hypothesis is not true. Similarly, the F-statistic tests for significance of dropping $p_1 - p_0$ parameters from a model with $p_1 + 1$ parameters simultaneously:

$$F = \frac{(\text{RSS}_0 - \text{RSS}_1)/(p_1 - p_0)}{\text{RSS}_1/(N - p_1 - 1)}, \tag{3.13}$$

where RSS_1 is the residual sum-of-squares fit for the model with $p_1 + 1$ parameters, and RSS_0 is the same for the smaller model with $p_0 + 1$ parameters. ■

Solution. Consider first the denominator in Equation (3.13):

$$\text{RSS}_1/(N - p_1 - 1) = \frac{1}{N - p_1 - 1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \hat{\sigma}^2, \tag{3.1}$$

using the definition of $\hat{\sigma}$ on p. 47 of [2]. Since we are only dropping one parameter from the model, $p_1 - p_0 = 1$. It remains to show that the numerator of Equation (3.13) is $\hat{\beta}_j^2/v_j$. For this, it is most useful to appeal to Algorithm 3.1 in [2]. First note that, upon reordering the inputs, it is without loss of generality to assume that the final $(p_1 + 1)$-th input is dropped. Let

$$\mathbf{z}_0, \dots, \mathbf{z}_{p_1 - 1}$$

be the orthogonal basis for the column space of $\widetilde{\mathbf{X}}$ obtained from Algorirthm 3.1, where $\widetilde{\mathbf{X}}$ is the matrix obtained from $\mathbf{X}$ by removing the (p_1+1)-th column. Doing this a final time for the removed column, we have that

$$\mathbf{z}_0, \dots, \mathbf{z}_{p_1-1}, \mathbf{z}_{p_1}$$

is a basis for the column space of $\mathbf{X}$. Extend this basis to an orthonormal basis $\left\{\frac{\mathbf{z}_j}{\|\mathbf{z}_j\|}\right\}_{j=1}^{N}$ for $\mathbb{R}^N$. Write $\mathbf{y}$ in terms of this basis:

$$\mathbf{y} = \sum_{j=0}^{N} a_j \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|}.$$

Notice that $\mathrm{RSS}_0 = \sum_{j=p_1-1}^{N} |a_j|^2$, while $\mathrm{RSS}_1 = \sum_{j=p_1}^{N} |a_j|^2$. Hence,

$$\mathrm{RSS}_0 - \mathrm{RSS}_1 = \left| \left\langle \mathbf{y}, \frac{\mathbf{z}_{p_1}}{\|\mathbf{z}_{p_1}\|} \right\rangle \right|^2 = \frac{|\langle \mathbf{y}, \mathbf{z}_{p_1} \rangle|^2}{\langle \mathbf{z}_{p_1}, \mathbf{z}_{p_1} \rangle} = \left| \frac{\langle \mathbf{y}, \mathbf{z}_{p_1} \rangle}{\langle \mathbf{z}_{p_1}, \mathbf{z}_{p_1} \rangle} \right|^2 \langle \mathbf{z}_{p_1}, \mathbf{z}_{p_1} \rangle = \hat{\beta}_{p_1}^2 \|\mathbf{z}_{p_1}\|^2. \tag{3.2}$$

To see that $\|z_{p_1}\|^2$ is indeed the (p_1+1)-th diagonal entry of $(\mathbf{X}^\top \mathbf{X})^{-1}$, consider the QR decomposition. We have that

$$X = QR,$$

where Q is an orthogonal matrix, and R is an upper triangular matrix with $R = D\Gamma$. Here, Γ is upper triangular with entries $(\Gamma)_{ij} = \frac{\langle \mathbf{z}_i, \mathbf{x}_j \rangle}{\langle \mathbf{z}_i, \mathbf{z}_i \rangle}$, and D is diagonal with entries $(D)_{ii} = \|\mathbf{z}_i\|$. Notice that

$$\mathbf{X}^\top \mathbf{X} = R^\top Q^\top Q R = R^\top R = \Gamma^\top D^2 \Gamma.$$

Inverting, we have that $(\mathbf{X}^\top \mathbf{X})^{-1} = \Gamma^{-1} D^{-2} (\Gamma^{-1})^\top$. Using the above to investigate $(\mathbf{X}^\top \mathbf{X})^{-1}$ entry by entry, we have that

$$\left((\mathbf{X}^\top \mathbf{X})^{-1}\right)_{(p_1+1),(p_1+1)} = \left(\text{last row of } \Gamma^{-1}\right) \cdot \left(\text{last column of } D^{-2}(\Gamma^{-1})^\top\right).$$

Since Γ is upper triangular, it is easy to see that Γ^{-1} is upper triangular. Moreover, the diagonal entries of Γ^{-1} are given by the reciprocal of the corresponding diagonal entries in Γ. Using this, the final column of $D^{-2}(\Gamma^{-1})^\top$ is easily computed to be

$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ \frac{1}{\|\mathbf{z}_{p_1}\|^2} \frac{\|\mathbf{z}_{p_1}\|^2}{\langle \mathbf{x}_{p_1}, \mathbf{z}_{p_1} \rangle} \end{bmatrix},$$

since D^{-2} is diagonal and $(\Gamma^{-1})^\top$ is *lower* triangular. Since $\mathbf{x}_{p_1} = \mathbf{z}_{p_1} + \sum_{j<p_1} \hat{\gamma}_{jp_1} \mathbf{z}_j$, the above vector simplifies to

$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}.$$

Similarly, the last row of Γ^{-1} is given by

$$\begin{bmatrix} 0 & \dots & 0 & \frac{\|\mathbf{z}_{p_1}\|^2}{\langle \mathbf{x}_{p_1}, \mathbf{z}_{p_1} \rangle} \end{bmatrix}.$$

Again since $\mathbf{x}_{p_1} = \mathbf{z}_{p_1} + \sum_{j<p_1} \hat{\gamma}_{jp_1}\mathbf{z}_j$, the final row simplifies to

$$\begin{bmatrix} 0 & \dots & 0 & \|\mathbf{z}_{p_1}\|^2 \end{bmatrix},$$

and hence the (p_1+1)-th diagonal entry of $(\mathbf{X}^\top\mathbf{X})^{-1}$ is $\|\mathbf{z}_{p_1}\|^2$, which was to be demonstrated. Putting together the above discussion with Equations (3.1) and (3.2), we have that

$$F = \frac{\hat{\beta}_j^2}{\hat{\sigma}^2 v_j}.$$

This completes the problem. ■

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 fitting the 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 $a^\top\beta = \sum_{j=0}^3 \beta_j x_0^j$.
(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.

Comment. The equation referenced above in the text is given by

$$C_\beta = \{\beta : (\hat{\beta}-\beta)^\top\mathbf{X}^\top\mathbf{X}(\hat{\beta}-\beta) \le \hat{\sigma}(\chi^2_{p+1})^{1-\alpha}\}. \tag{3.15}$$

The corresponding confidence band generated for f is

$$\{f_\beta(x) : \beta \in C_\beta\}.$$

■

Solution. We will discuss for each item separately. For the following discussion we will assume the model is $Y = f(X) = \sum_{j=0}^3 \beta_j X^j + \epsilon$ where $\epsilon \sim \mathcal{N}(0,\sigma^2)$, and given N data points $\mathbf{X}$ with observations $\mathbf{Y}$, we estimate with a cubic model $\hat{f}$ obtained by least squares on the data.

(1) Given a point x_0, we form the estimated output $\hat{f}(x_0)$. The confidence band for this point is given by

$$\hat{f}(x_0) \pm z\sqrt{\mathrm{Var}(\hat{f}(x_0))}$$

where z is the desired confidence level associated with the distribution of $\frac{\hat{f}(x_0)-f(x_0)}{\sqrt{\mathrm{Var}(\hat{f})}}$. For convenience, denote by $\mathbf{x}_0$ the vector $\mathbf{x}_0 = [1, x_0, x_0^2, x_0^3]^\top$. Notice that

$$\mathbb{E}\hat{f}(x_0) = \mathbb{E}\mathbf{x}_0^\top\hat{\beta} = \mathbf{x}_0^\top\beta,$$

since $\mathbb{E}\hat{\beta} = \beta$ (see Equation 3.10 in [2]). Similarly,

$$\mathrm{Var}(\hat{f}(x_0)) = \mathbb{E}\left(\mathbf{x}_0^\top(\hat{\beta}-\beta)(\hat{\beta}-\beta)^\top\mathbf{x}_0\right)$$

$$= \mathbf{x}_0^\top\mathbb{E}\left[(\hat{\beta}-\beta)(\hat{\beta}-\beta)^\top\right]\mathbf{x}_0 = \sigma^2\mathbf{x}_0^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_0,$$

where we again used Equation (3.10) in [2]. Moreover, $\hat{f}(x_0)$ follows a normal distribution, and hence z corresponds to the quantiles of a standard normal distribution. It follows that

$$\hat{f}(x_0) - 1.96\sigma\sqrt{\mathbf{x}_0^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_0} \leq f(x_0) \leq \hat{f}(x_0) + 1.96\sigma\sqrt{\mathbf{x}_0^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_0}.$$

with probability at least .95. Note that if σ is not known, then we estimate σ with $\hat{\sigma}$, in which case the above follows a t-distribution. For large N, the difference is negligible, and we can still estimate the above using 1.96. See Figure 3.3 in [2].

(2) To generate a confidence band for f using C_β, for each point x_0, we include in our band the points

$$\{\mathbf{x}_0^\top\beta : \beta \in C_\beta\}.$$

So for a fixed point $\mathbf{x}_0$, we investigate local minima and maxima of the map $g : \beta \mapsto \mathbf{x}_0^\top\beta$. For nonzero $\mathbf{x}_0$, we see by investigating ∇g that no local maxima or minima occur on the interior. Hence, the maximum and minimum for each point will occur on the boundary of C_β. Note that $(\chi^2_4)^{.975} = 11.14$. For convenience, denote by $11.14\sigma^2 := \epsilon$. Define an innerproduct that depends on $\mathbf{X}^\top\mathbf{X}$ in the following way:

$$\langle v, w\rangle_{\mathbf{X}^\top\mathbf{X}} = \langle(\mathbf{X}^\top\mathbf{X})^{1/2}v, (\mathbf{X}^\top\mathbf{X})^{1/2}w\rangle,$$

where $\langle\cdot,\cdot\rangle$ denotes the standard euclidean innerproduct. Given a fixed data point $\mathbf{x}_0$ and estimate $\hat{\beta}$, notice that $\mathbf{x}_0^\top\hat{\beta}$ is a constant. Hence,

$$\max_{\|\beta-\hat{\beta}\|^2_{\mathbf{X}^\top\mathbf{X}}\leq\epsilon} \mathbf{x}_0^\top\beta = \mathbf{x}_0^\top\hat{\beta} + \max_{\|\beta-\hat{\beta}\|^2_{\mathbf{X}^\top\mathbf{X}}\leq\epsilon} \mathbf{x}_0^\top(\beta-\hat{\beta}).$$

Similarly for the minimum. Hence, changing variables, it is equivalent to consider the problem

$$\max_{\|\mathbf{w}\|^2_{\mathbf{X}^\top\mathbf{X}}=\epsilon} \mathbf{x}_0^\top\mathbf{w}.$$

Notice that if $\mathbf{w}_{\max}$ is the argmax of the above, then $-\mathbf{w}_{\max}$ will be the argmin of the corresponding minimization problem. Hence, to find the solution to both the maximization and minimization problems simultaneously, it suffices to find the argmax of the square. I.e., we study

$$\max_{\|\mathbf{w}\|^2_{\mathbf{X}^\top\mathbf{X}}=\epsilon} \langle\mathbf{x}_0^\top\mathbf{w}, \mathbf{x}_0^\top\mathbf{w}\rangle = \max_{\|\mathbf{w}\|^2_{\mathbf{X}^\top\mathbf{X}}=\epsilon} \langle\mathbf{x}_0\mathbf{x}_0^\top\mathbf{w}, \mathbf{w}\rangle.$$

Writing in terms of $\widetilde{\mathbf{w}} := \frac{1}{\sqrt{\epsilon}}\mathbf{w}$, we have

$$\max_{\|\mathbf{w}\|^2_{\mathbf{X}^\top\mathbf{X}}=\epsilon} \langle\mathbf{x}_0\mathbf{x}_0^\top\mathbf{w}, \mathbf{w}\rangle = \epsilon \max_{\|\widetilde{\mathbf{w}}\|^2_{\mathbf{X}^\top\mathbf{X}}=1} \langle\mathbf{x}_0\mathbf{x}_0^\top\widetilde{\mathbf{w}}, \widetilde{\mathbf{w}}\rangle.$$

Using that $(\mathbf{X}^\top\mathbf{X})^{-1/2}(\mathbf{X}^\top\mathbf{X})^{1/2} = I$, and changing variables to $\mathbf{v} = (\mathbf{X}^\top\mathbf{X})^{1/2}\widetilde{\mathbf{w}}$, we can rewrite the above as

$$\epsilon \max_{\|\widetilde{\mathbf{w}}\|^2_{\mathbf{X}^\top\mathbf{X}}=1} \langle\mathbf{x}_0\mathbf{x}_0^\top\widetilde{\mathbf{w}}, \widetilde{\mathbf{w}}\rangle = \epsilon \max_{\|\mathbf{v}\|^2=1} \langle\mathbf{x}_0\mathbf{x}_0^\top(\mathbf{X}^\top\mathbf{X})^{-1/2}\mathbf{v}, (\mathbf{X}^\top\mathbf{X})^{-1/2}\mathbf{v}\rangle$$

$$= \epsilon \max_{\|\mathbf{v}\|^2=1} \langle(\mathbf{X}^\top\mathbf{X})^{-1/2}\mathbf{x}_0\mathbf{x}_0^\top(\mathbf{X}^\top\mathbf{X})^{-1/2}\mathbf{v}, \mathbf{v}\rangle$$

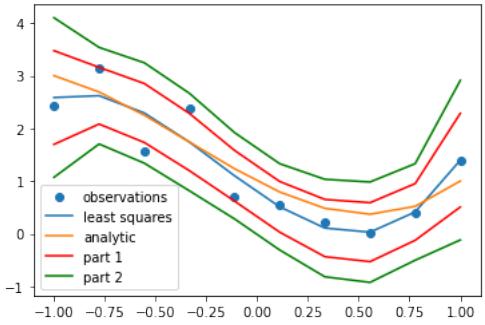

FIGURE 1. 95% Confidence intervals for fitting a cubic polynomial from 10 noisey observations.

where on the final line we used that $(\mathbf{X}^\top\mathbf{X})^{-1/2}$ is self-adjoint. Letting

$$A = (\mathbf{X}^\top\mathbf{X})^{-1/2}\mathbf{x}_0\mathbf{x}_0^\top(\mathbf{X}^\top\mathbf{X})^{-1/2},$$

this can be rewritten as an eigenvalue problem:

$$\epsilon \max_{\|\mathbf{v}\|^2=1} \langle A\mathbf{v}, \mathbf{v}\rangle = \epsilon\lambda_{\max}(A).$$

This uses the fact that A is self-adjoint (easy to check). Tracing through definitions, we can find the original maximum and minimum in terms of $\epsilon\lambda_{\max}(A)$. In particular, the 95% confidence band for $\mathbf{x}_0^\top\beta$ is given by

$$\mathbf{x}_0^\top\hat{\beta} \pm \sqrt{\epsilon\lambda_{\max}(A)} = \mathbf{x}_0^\top\hat{\beta} \pm \hat{\sigma}\sqrt{11.14\lambda_{\max}(A)}.$$

This completes the discussion of part (2).

The bands in part (2) will certainly be wider. Intuitively, this follows from the fact that generating bands in terms of β is done in 4 dimensions, and hence the confidence bands depend on $(\chi^2_{4+1})^{.975}$, while generating bands in the standard way occurs in 1 dimension, and the bands depend on 1.96, the 97.5-th percentile of the standard normal distribution. More formally, notice that the term in part (1) is given by

$$\sqrt{\mathbf{x}_0^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_0} = \sqrt{\mathrm{tr}\left(\mathbf{x}_0^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_0\right)} = \sqrt{\mathrm{tr}\left((\mathbf{X}^\top\mathbf{X})^{-1/2}\mathbf{x}_0\mathbf{x}_0^\top(\mathbf{X}^\top\mathbf{X})^{-1/2}\right)} =$$

$$= \sqrt{\mathrm{tr}\,(A)}.$$

where we used the cyclic property of trace. The trace of a matrix is the sum of its eigenvalues. However, since $\mathbf{x}_0\mathbf{x}_0^\top$ is rank 1, its clear that A also has rank 1, since $(\mathbf{X}^\top\mathbf{X})^{-1/2}$ is invertible. Hence, the sum of the eigenvalues of A is equal to its maximum eigenvalue, $\lambda_{\max}(A)$. Therefore, provided that $\hat{\sigma} \approx \sigma$, we have that $\hat{\sigma}\sqrt{11.14} > 1.96\sigma$. Hence, the bands in part 2 will be larger.

We display the results for the numerical experiment performed as follows. We generated a dataset X of 10 points in the interval $[-1, 1]$. For each $x_i \in X$, we saved noisey observations from the polynomial

$$f(x) = x^3 + x^2 - 2x + 1.$$

Namely, our observations $Y = \{y_1, \ldots, y_{10}\}$ were given by

$$y_i = f(x_i) + \epsilon_i,$$

where $\epsilon_i \sim \mathcal{N}(0, .5^2)$. We then performed least squares fit, and computed the upper and lower confidence intervals as described above. The results are shown in Figure 1. Our analysis is confirmed, and the confidence bands using the method described in part (2) are indeed wider. ■

Exercise 3.3. *Gauss-Markov theorem:*

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

(b) The matrix inequality $\mathbf{B} \preceq \mathbf{A}$ 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 $\widetilde{\mathbf{V}}$ is the variance-covariance matrix of any other linear unbiased estimate, then $\hat{\mathbf{V}} \preceq \widetilde{\mathbf{V}}$.

Solution. (a) Let $a^\top\beta$ denote the true quantity, and suppose we are given fixed N noisey observations $\mathbf{y}$ where the noise has mean 0 and variance σ^2. Let $\mathbf{X}$ be the design matrix so that $\mathbf{y} = \mathbf{X}\beta + \epsilon$. Denote by $b^\top\mathbf{y} = a^\top(\mathbf{X}^\top\mathbf{X})^{-1}X^\top\mathbf{y}$ the least squares estimate of the parameter, and let $\mathbf{c}^\top\mathbf{y}$ be any other unbiased, linear estimate. Note that we can write $\mathbf{c}^\top\mathbf{y} = a^\top(\mathbf{X}^\top\mathbf{X})^{-1}X^\top\mathbf{y} + \gamma^\top\mathbf{y}$, for some nonzero γ. Since $\mathbf{c}^\top\mathbf{y}$ is unbiased,

$$a^\top\beta = \mathbb{E}[\mathbf{c}^\top\mathbf{y}] = a^\top\beta + \gamma^\top\mathbb{E}\left[\mathbf{y}\right].$$

Hence,

$$\gamma^\top\mathbf{X}\beta = 0. \tag{3.3}$$

Let us use this to calculate the variance of $\mathbf{c}^\top\mathbf{y}$ in terms of the variance of the least squares estimate. Namely,

$$\begin{aligned}\mathrm{Var}(\mathbf{c}^\top\mathbf{y}) &= \mathbb{E}\left[\left(a^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} + \gamma^\top\mathbf{y} - a^\top\beta\right)^2\right] \\ &= \mathbb{E}\left[(v - \gamma^\top\mathbf{y})^2\right],\end{aligned}$$

where $v = a^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} - a^\top\beta$. Expanding the above yields

$$\mathbb{E}\left[v^2\right] - 2\mathbb{E}\left[v\mathbf{y}^\top\right]\gamma + \gamma^\top\mathbb{E}\left[\mathbf{y}\mathbf{y}^\top\right]\gamma.$$

The first term in the equation above is precisely the variance of the least squares estimate. It is easy to see that the second term vanishes completely, using that $\mathbb{E}\left[\mathbf{y}^\top\right] = \beta^\top\mathbf{X}^\top$, $\mathbb{E}\left[\mathbf{y}\mathbf{y}^\top\right] = \mathbf{X}\beta\beta^\top\mathbf{X} + \sigma^2\mathbf{I}$, and (3.3). Similarly, the final term simplifies to $\sigma^2\gamma^\top\gamma$. Putting it all together, we have that

$$\mathrm{Var}(\mathbf{c}^\top\mathbf{y}) = \mathrm{Var}(b^\top\mathbf{y}) + \sigma^2\|\gamma\|^2.$$

This shows that $\mathrm{Var}(b^\top\mathbf{y}) \leq \mathrm{Var}(\mathbf{c}^\top\mathbf{y})$, and hence completes part (a).

(b) The outline of this problem is similar. Let $\hat{\beta} := (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ be the least squares estimate, and $\widetilde{\beta} := \mathbf{A}\mathbf{y}$ be any other linear unbiased estimate. Write

$$\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + (\mathbf{A} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top) := (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{B}.$$

Using that $\mathbb{E}\tilde{\beta} = \mathbb{E}\hat{\beta} = \beta$, the above equation shows, after taking expectations and multiplying on the right by $\mathbf{y}$, that

$$0 = \mathbb{E}\left[\mathbf{B}\mathbf{y}\right] = \mathbf{B}\mathbf{X}\beta. \tag{3.4}$$

We keep this in mind, and expand out the variance-covariance expression for $\mathbf{Ay}$:

$$\begin{aligned}\widetilde{\mathbf{V}} &= \mathbb{E}\left[\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}+\mathbf{By}-\beta\right)\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}+\mathbf{By}-\beta\right)^\top\right] \\ &= \mathbb{E}\left[(\mathbf{v}+\mathbf{By})\left(\mathbf{v}^\top-\mathbf{y}^\top\mathbf{B}^\top\right)\right]\end{aligned}$$

where $\mathbf{v}=\hat{\beta}-\beta$. Notice that $\mathbb{E}\left[\mathbf{vv}^\top\right]=\mathrm{Var}(\hat{\beta})$. Hence, expanding the above yields

$$\begin{aligned}\hat{\mathbf{V}} &- \mathbb{E}\left[\mathbf{Byy}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}-\mathbf{By}\beta^\top\right] \\ &- \mathbb{E}\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{yy}^\top\mathbf{B}^\top-\beta\mathbf{y}^\top\mathbf{B}^\top\right)\right]+\mathbb{E}\left[\mathbf{Byy}^\top\mathbf{B}^\top\right].\end{aligned}$$

Using that $\mathbb{E}\left[\mathbf{yy}^\top\right]=\mathbf{X}\beta\beta^\top\mathbf{X}^\top+\sigma^2\mathbf{I}$, $\mathbb{E}\left[\mathbf{y}\right]=\mathbf{X}\beta$, and (3.4), we see that both cross terms vanish. Moreover, the last term simplifies to $\sigma^2\mathbf{BB}^\top$. We are left with

$$\widetilde{\mathbf{V}}=\hat{\mathbf{V}}+\sigma^2\mathbf{BB}^\top.$$

$\mathbf{BB}^\top$ is clearly positive semi-definite, since for any vector $\mathbf{w}\in\mathbb{R}^{p+1}$,

$$\langle\mathbf{BB}^\top\mathbf{w},\mathbf{w}\rangle=\langle\mathbf{B}^\top\mathbf{w},\mathbf{B}^\top\mathbf{w}\rangle\geq 0.$$

Hence, $\hat{\mathbf{V}}\preceq\widetilde{\mathbf{V}}$, which was to be demonstrated.

■

Exercise 3.4. *Show how the vector of least squares coefficients can be obtained from a single path of the Gram-Schmidt procedure (Algorithm 3.1). Represent your solution in terms of the QR decomposition of* $\mathbf{X}$.

Solution. By definition of least squares, we have that

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

and

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

A single pass of the Gram-Schmidt algorithm (i.e. algorithm 3.1) yields an orthogonal matrix $\mathbf{Q}$ (i.e. $\mathbf{Q}^\top\mathbf{Q}=\mathbf{I}$), and an upper-triangular matrix $\mathbf{R}$ such that

$$\mathbf{X}=\mathbf{QR}.$$

Plugging in the QR-decomposition for $\hat{\beta}$, we have that

$$\hat{\beta}=(\mathbf{R}^\top\mathbf{Q}^\top\mathbf{QR})^{-1}\mathbf{R}^\top\mathbf{Q}^\top\mathbf{y}.$$

Since $\mathbf{R}$ is invertible and $\mathbf{Q}$ is orthogonal, this simplifies to

$$\hat{\beta}=\mathbf{R}^{-1}(\mathbf{R}^{-1})^\top\mathbf{R}^\top\mathbf{Q}^\top\mathbf{y}=\mathbf{R}^{-1}\mathbf{Q}^\top\mathbf{y}.$$

Hence,

$$\hat{\mathbf{y}}=\mathbf{QR}\hat{\beta}=\mathbf{QRR}^{-1}\mathbf{Q}^\top\mathbf{y}=\mathbf{QQ}^\top\mathbf{y}.$$

This shows how a single pass of Gram-Schmidt yields the vector of least-squares coefficients, and represents that vector, as well $\hat{\mathbf{y}}$, in terms of the QR-decomposition. ■

Exercise 3.5. *Consider the ridge regression problem (3.41). Show that this problem is equivalent to the problem*

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

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. The minimization problem in (3.41) is given by

$$\text{argmin}_{\beta} \left\{ \sum_{i=1}^{N} [y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j]^2 + \lambda \sum_{j=1}^{p} (\beta_i)^2 \right\}.$$

Simply adding and subtracting $\sum_{j=1}^{p} \overline{x}_j$, we obtain that the above is given by

$$\text{argmin}_{\beta} \left\{ \sum_{i=1}^{N} [y_i - (\beta_0 - \sum_{j=1}^{p} \overline{x}_j) - \sum_{j=1}^{p} (x_{ij} - \overline{x}_j)\beta_j]^2 + \lambda \sum_{j=1}^{p} (\beta_i)^2 \right\}.$$

Defining $\beta_0^c := (\beta_0 - \sum_{j=1}^{p} \overline{x}_j)$, and $\beta_j^c = \beta_j$ for $j = 1, \ldots, p$, we have that the above is equivalent to

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

This establishes that the two minimizations are equivalent, and gives a clear correspondence between $\hat{\beta}$ and $\hat{\beta}^c$. While β_j^c for $j \geq 1$ are found in the same way, we can use this to find a closed form expression for β_0^c. Namely, setting the derivative of the above quantity w.r.t. β_0^c to 0 yields the equation

$$\sum_{i=1}^{N} [y_i - \beta_0^c - \sum_{j=1}^{p} (x_{ij} - \overline{x}_j)\beta_j^c] = 0.$$

Distributing the sum, we see that the second term vanishes: $\sum_{i=1}^{N} \sum_{j=1}^{p} (N\frac{1}{N} x_{ij} - \overline{x}_j) = \sum_{j=1}^{p} (N\frac{1}{N} \sum_{i=1}^{N} x_{ij} - N\overline{x}_j) = \sum_{j=1}^{p} (N\overline{x}_j - N\overline{x}_j) = 0$. Hence, we obtain that

$$\beta_0^c = \frac{1}{N} \sum_{i=1}^{N} y_i$$

is the solution for β_0^c. Therefore, replacing $\mathbf{y}$ with the centered data $\mathbf{y}_c$, as well as $\mathbf{X}$ with the $N \times p$ centered matrix $\mathbf{X}_c$ (note: *not* $N \times (p+1)$), we obtain that the term in the curly brackets above can be rewritten as

$$(\mathbf{y}_c - \mathbf{X}_c\beta^c)^\top (\mathbf{y}_c - \mathbf{X}_c\beta^c) + \lambda \|\beta^c\|^2. \tag{3.5}$$

Setting the derivative w.r.t. β equal to 0 yields the equation

$$-2\mathbf{X}_c^\top (\mathbf{y}_c - \mathbf{X}_c\beta) + 2\lambda\beta = 0.$$

Solving gives the minimizer:

$$\hat{\beta}^c = (\mathbf{X}_c^\top \mathbf{X}_c + \lambda \mathbf{I})^{-1} \mathbf{X}_c^\top \mathbf{y}.$$

For the lasso, only the penalty term is changed. Hence, the exact analysis as above can be used to recenter the data, and we can rewrite the lasso problem as

$$\text{argmin}_{\beta^c}\{(\mathbf{y}_c - \mathbf{X}_c\beta^c)^\top(\mathbf{y}_c - \mathbf{X}_c\beta^c) + \lambda\|\beta^c\|_{L^1}.\},$$

where $\|\beta^c\|_{L^1} = \sum_{j=1}^p |\beta_j^c|$. We note, as the book does, that this no longer has a closed-form solution in general, and becomes a quadratic programming problem. ■

Exercise 3.6. *Show that the ride regression estimate is the mean (and mode) of the posterior distribution, under a Gaussian prior $\beta \sim \mathcal{N}(0, \tau\mathbf{I})$, and Gaussian sampling model $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbf{I})$. Find the relationship between the regularization parameter λ in the ridge formula, and the variances τ and σ.*

Solution. The probability density function associated to β is given by

$$C\exp\left(\frac{-\|\beta\|^2}{2\tau}\right),$$

while the probability density function associated to the sampling model (i.e., probability of $(\mathbf{X}, \mathbf{y})$ given β) is given by

$$D\exp\left(\frac{-\|\mathbf{y} - \mathbf{X}\beta\|^2}{2\sigma^2}\right).$$

where C and D are constants irrelevant in this setting. Bayes' theorem states that posterior, probability of β given $(\mathbf{X}, \mathbf{y})$ is proportional to the product of the above probabilities:

$$p(\beta|(\mathbf{X}, \mathbf{y})) = \text{Const}\exp\left(\frac{-\|\mathbf{y} - \mathbf{X}\beta\|^2}{2\sigma^2} - \frac{\|\beta\|^2}{2\tau}\right).$$

for some proper normalization making the above into a probability density function. The above is again Gaussian, and hence the mean of this distribution is given by $\hat{\beta}$ which maximizes the above quantity. This is equivalent to maximizing the following:

$$\hat{\beta} = \text{argmax}_\beta \log\left(p(\beta|(\mathbf{X}, \mathbf{y}))\right) = \text{argmax}_\beta \log(\text{Const}) - \frac{\|\mathbf{y} - \mathbf{X}\beta\|^2}{2\sigma^2} - \frac{\|\beta\|^2}{2\tau}.$$

The constant in front does not depend on β, so it can be removed without affecting the argmax. Hence, it's equivalent to minimizing the negative of the above. I.e.,

$$\hat{\beta} = \text{argmin}_\beta \frac{\|\mathbf{y} - \mathbf{X}\beta\|^2}{2\sigma^2} + \frac{\|\beta\|^2}{2\tau}.$$

Multipling by a constant does not change the argmin, so

$$\hat{\beta} = \text{argmin}_\beta \|\mathbf{y} - \mathbf{X}\beta\|^2 + \frac{\sigma^2\|\beta\|^2}{\tau}.$$

Referring back to (3.5), we see that this is equivalent to the ridge regression estimate with $\lambda = \frac{\sigma^2}{\tau}$. ■

Exercise 3.7. *Assume $y_i \sim \mathcal{N}(\beta_0 + x_i^\top\beta, \sigma^2)$, $i = 1, 2, \ldots, N$, and the parameters $\beta_j, j = 1, \ldots, p$ 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 prior, show that the (minus) log-posterior density of β is proportional to $\sum_{i=1}^N (y_i - \beta_0 - \sum_j x_{ij}\beta_j)^2 + \lambda\sum_{j=1^p}\beta_j^2$.*

Solution. The solution to this exercise is essentially the same as the one above. Note that in this situation, since β_0 is not governed by a prior, it appears in the probability density function of $(\mathbf{X}.\mathbf{y})$ given β, as it is in the formula for the mean, but β_0 does not appear in the other term. I.e, by the same reasoning above, we have that

$$p(\beta|((\mathbf{X},\mathbf{y})) = \text{Const}\exp\left(\frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i-\beta_0-\sum_j x_{ij}\beta_j)^2\right)\exp\left(\frac{1}{\tau^2}\sum_{j=1}^{p}\beta_j^2\right).$$

Taking the log and carrying out the same reasoning as before and multiplying by σ^2, we see that the minus log-posterior of β is proportional to the ridge problem, which was to be demonstrated. ■

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 $\mathbf{Q}_2$ and $\mathbf{U}$ span the same subspace, where $\mathbf{Q}_2$ is the sub-matrix of $\mathbf{Q}$ with the first column removed. Under what circumstances will they be the same, up to sign flips?*

Solution. One can easily check that $\mathbf{U}\mathbf{U}^\top$ is the projection onto the column space of $\tilde{\mathbf{X}}$. Hence, since the columns of $\mathbf{U}$ are orthogonal, it follows that they span the column space of $\tilde{\mathbf{X}}$:

$$\text{Col}(\tilde{\mathbf{X}}) = \text{Col}(\mathbf{U}).$$

So, it suffices to show that $\text{Col}(\mathbf{Q}_2) = \text{Col}(\tilde{\mathbf{X}})$. Notice that

$$\text{Col}(\mathbf{X}) = \text{Span}\{(1,\dots,1)^\top\} \oplus \text{Col}(\tilde{\mathbf{X}}),$$

and since the columns of $\tilde{\mathbf{X}}$ are centered, it follows that any vector in $\text{Span}\{(1,\dots,1)^\top\}$ is orthogonal to any vector in $\text{Col}(\tilde{\mathbf{X}})$. Hence, $\text{Col}(\tilde{\mathbf{X}})$ is the unique orhtogonal complement to $\text{Span}\{(1,\dots,1)^\top\}$.
Similarly to $\mathbf{U}$, the columns of $\mathbf{Q}$ form an orthonormal basis for the column space of $\mathbf{X}$ (see Exercise 3.4). Hence

$$\text{Col}(\mathbf{X}) = \text{Col}(\mathbf{Q}) = \text{Span}\{\mathbf{q}_1\} \oplus \text{Col}(\mathbf{Q}_2),$$

where $\mathbf{q}_1$ denotes the first column of $\mathbf{Q}$, and $\mathbf{Q}_2$ is the unique orthogonal complement to $\text{Span}\{\mathbf{q}_1\}$. The first column of $\mathbf{Q}$ is just the normalized vector of all ones (indeed, see algorithm 3.1.), we have

$$\text{Col}(\mathbf{X}) = \text{Span}\{(1,\dots,1)^\top\} \oplus \text{Col}(\mathbf{Q}_2),$$

where again every vector in $\text{Col}(\mathbf{Q}_2)$ is orthogonal to any vector in $\text{Span}\{(1,\dots,1)^\top\}$. By uniqueness of the orthogonal complement, it follows that

$$\text{Col}(\mathbf{Q}_2) = \text{Col}(\tilde{\mathbf{X}}) = \text{Col}(\mathbf{U}),$$

which was to be demonstrated.
To see when the matrices $\mathbf{U}$ and $\mathbf{Q}_2$ are the same, consider the following. By definition, the columns of $\mathbf{U}$ are the normalized vectors corresponding to the p nonzero eigenvalues of $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top$:

$$\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top\mathbf{u}_i = \sigma_i\mathbf{u}_i,$$

where $\mathbf{u}_i$ is the i-th column of $\mathbf{U}$. Since the columns of $\mathbf{Q}_2$ are normalized, $\mathbf{Q}_2 = \mathbf{U}$ is equivalent to

$$\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top \mathbf{q}_{i+1} = \sigma_i \mathbf{q}_{i+1},$$

for $i = 1, \dots p$. ■

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. Performing QR decomposition on $\mathbf{X}_1$, we obtain q *orthonormal* vectors $\{\mathbf{q}_i\}$. For $i = 1, \dots, p - q$, denote by $\mathbf{q}_{q+1}^{(i)}$ the i-th column of X_2, orthogonalized with respect to $\{\mathbf{q}_j\}_{j=1}^q$. For each $i = 1, \dots, p - q$, extend $\{\mathbf{q}_j\}_{j=1}^q \cup \{\mathbf{q}_{q+1}^{(i)}\}$ to a basis for $\mathbb{R}^N$, with vectors $\{\mathbf{q}_j^{(i)}\}_{j=q+2}^p$. Notice that the current residual is given by

$$\mathbf{r} := \mathbf{y} - \hat{\mathbf{y}} \quad = \quad \sum_{j=q+1}^{N} \langle \mathbf{y}, \mathbf{q}_j^{(i)} \rangle \mathbf{q}_j^{(i)}.$$

Note that the above has norm independent of i, and for each i is given by

$$\|\mathbf{r}\|^2 = |\langle \mathbf{y}, \mathbf{q}_{q+1}^{(i)} \rangle|^2 + \sum_{j=q+2}^{N} |\langle \mathbf{y}, \mathbf{q}_j^{(i)} \rangle|^2.$$

It follows that the residual $\mathbf{r}_i$ after adding the i-th column of $\mathbf{X}_2$ into model has norm

$$\|\mathbf{r}_i\|^2 = \sum_{j=q+2}^{N} |\langle \mathbf{y}, \mathbf{q}_j^{(i)} \rangle|^2.$$

Hence, choosing

$$\text{argmax}_i |\langle \mathbf{y}, \mathbf{q}_{q+1}^{(i)} \rangle|^2$$

results in a model that decreases the residual sum of squares the most. This completes the problem. ■

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. Let RSS_1 denote the residual sum of squares for the model with p-parameters, and $\text{RSS}_{0,j}$ a model with $p - 1$ parameters, after dropping the j-th coefficient. We wish to minimize the increase

$$w_j := \text{RSS}_{0,j} - \text{RSS}_1.$$

I.e., we wish to find

$$j^* := \text{argmin}_j w_j.$$

Multiplying by a constant does not change the argmin. Hence,

$$j^* := \text{argmin}_j \frac{w_j/(p-(p-1))}{\text{RSS}_1/(N-p-1)}.$$

Notice that the above is precisely the F statistic (3.13) for dropping a single coefficient. Exercise 3.1 showed this is precisely the corresponding z-score. Hence

$$j^* := \text{argmin}_j z_j,$$

where z_j denotes the j-th Z-score. This completes the problem. ■

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 $\mathbf{\Sigma}_i$ are different for each observation?*

Comment. The multivariate linear regression corresponds to the situation where each observation y_i is a vector in $\mathbb{R}^K$, and each observation has some random noise ϵ, which is again a vector in $\mathbb{R}^K$. Denoting $\text{Cov}(\epsilon) = \mathbf{\Sigma}$ the covariance matrix for the noise associated with each observation, we have the minimization criterion

$$\text{RSS}(\mathbf{B}; \mathbf{\Sigma}) = \sum_{i=1}^{N} (y_i - f(x_i))^\top \mathbf{\Sigma}^{-1} (y_i - f(x_i)). \tag{3.40}$$

where $\mathbf{B}$ is the $(p+1) \times K$ matrix of parameters. The solution in the text mentioned above is given by

$$\hat{\mathbf{B}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}, \tag{3.39}$$

where $\mathbf{Y}$ is the $N \times K$ response matrix. ■

Solution. We begin by rewriting (3.40) more suggestively:

$$\text{RSS}(\mathbf{B}; \mathbf{\Sigma}) = \text{tr}\left((\mathbf{Y} - \mathbf{XB}) \mathbf{\Sigma}^{-1} (\mathbf{Y} - \mathbf{XB})^\top \right),$$

which is easy to check. Since $\mathbf{\Sigma} = \mathbb{E}\epsilon\epsilon^\top$ is a $K \times K$ symmetric and positive-definite matrix, it has a symmetric, positive definite square root $\mathbf{\Sigma}^{-1/2}$. Writing $\tilde{\mathbf{B}} = \mathbf{B}\mathbf{\Sigma}^{-1/2}$, $\tilde{\mathbf{Y}} = \mathbf{Y}\mathbf{\Sigma}^{-1/2}$, we can rewrite (3.40) as

$$\text{RSS}(\mathbf{B}; \mathbf{\Sigma}) = \text{tr}\left((\tilde{\mathbf{Y}} - \mathbf{X}\tilde{\mathbf{B}})(\tilde{\mathbf{Y}} - \mathbf{X}\tilde{\mathbf{B}})^\top \right) = \text{tr}\left((\tilde{\mathbf{Y}} - \mathbf{X}\tilde{\mathbf{B}})^\top (\tilde{\mathbf{Y}} - \mathbf{X}\tilde{\mathbf{B}}) \right).$$

We minimize the above over all matrices $\mathbf{B}\mathbf{\Sigma}^{-1/2}$. According to the standard solution from (3.39), we have that

$$\hat{\tilde{\mathbf{B}}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \tilde{\mathbf{Y}}.$$

Since the minimization was taken over matrices of the form $\mathbf{B}\mathbf{\Sigma}^{-1/2}$, the matrix for which the minimization is achieved has the form $\hat{\tilde{\mathbf{B}}} = \hat{\mathbf{B}}\mathbf{\Sigma}^{-1/2}$. Multiplying on the right by $\mathbf{\Sigma}^{1/2}$ shows that the solution is given by

$$\hat{\mathbf{B}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y},$$

which was to be demonstrated.
If the correlations $\mathbf{\Sigma}_i$ vary with each observation, we have the equation

$$\mathrm{RSS}(\mathbf{B}; \mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_N) = \sum_{i=1}^{N} (y_i - f(x_i))^\top \mathbf{\Sigma}_i^{-1} (y_i - f(x_i)).$$

Let $\mathbf{y}$ denote the column vector of size $N \cdot K$, obtained by stacking each observation, with observation y_i sitting as a column vector below observation y_{i-1}. Perform the same operation for $f(x_i)$ to obtain $\mathbf{f}$. Let $\mathbf{\Sigma}_{NK \times NK}$ denote the block diagonal matrix $\mathbf{\Sigma}_{NK\times NK} = \mathrm{diag}(\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_N)$. It is easy to check that

$$\mathrm{RSS}(\mathbf{B}; \mathbf{\Sigma}_1 \ldots, \mathbf{\Sigma}_N) = (\mathbf{y} - \mathbf{f})^\top \mathbf{\Sigma}_{NK\times NK}^{-1} (\mathbf{y} - \mathbf{f}).$$

Moreover, notice that $\mathbf{f} = \mathrm{diag}(\mathbf{B}^\top, \ldots, \mathbf{B}^\top)\mathbf{x}$, where $\mathbf{x}$ is again the N vectors of size $p+1$, $\mathbf{x}_i \in \mathbb{R}^{p+1}$, stacked ontop of each other to obtain a vector in $\mathbb{R}^{(p+1)N}$. Using this, it is easy to see that the map

$$\psi(\mathbf{b}_1, \mathbf{b}_2) = \mathrm{RSS}(\mathbf{B}_1 + \mathbf{B}_2; \mathbf{\Sigma}_1 \ldots, \mathbf{\Sigma}_N) - \mathrm{RSS}(\mathbf{B}_1; \mathbf{\Sigma}_1 \ldots, \mathbf{\Sigma}_N) - \mathrm{RSS}(\mathbf{B}_2; \mathbf{\Sigma}_1 \ldots, \mathbf{\Sigma}_N),$$

where $\mathbf{b}_i$ is the $K(p+1)$ vector corresponding to matrix $\mathbf{B}_i$ with the columns stacked ontop of each other, is bi-linear. Moreover, we see for any scalar $a \in \mathbb{R}$, we have

$$\mathrm{RSS}(a\mathbf{B}; \mathbf{\Sigma}_1 \ldots, \mathbf{\Sigma}_N) = a^2 \mathrm{RSS}(\mathbf{B}; \mathbf{\Sigma}_1 \ldots, \mathbf{\Sigma}_N).$$

By definition, the above defines a quadratic form. Define a symmetric matrix

$$\mathbf{A} \in \mathbb{R}^{(p+1)K \times (p+1)K}$$

with entries

$$(\mathbf{A})_{ij} = \frac{1}{2}\psi(\mathbf{e}_i, \mathbf{e}_j),$$

where $\mathbf{e}_i$ denotes the $K(p+1)$ vector with a 1 in the i-th position, and 0's elsewhere. It follows from the theory of quadratic forms that

$$\langle \mathbf{A}\mathbf{b}, \mathbf{b} \rangle = \mathrm{RSS}(\mathbf{B}; \mathbf{\Sigma}_1 \ldots, \mathbf{\Sigma}_N).$$

Hence, to minimize the above is to find the eigenvector $\hat{\mathbf{b}}$ of $\mathbf{A}$ corresponding to the smallest eigenvalue of $\mathbf{A}$. Using the identification of vectors in $\mathbb{R}^{K(p+1)}$ with $(p+1) \times K$ matrices mentioned above by stacking columns, we obtain the parameter set $\hat{\mathbf{B}}$ that minimizes $\mathrm{RSS}(\mathbf{B}; \mathbf{\Sigma}_1 \ldots, \mathbf{\Sigma}_N)$, which was to be demonstrated. ■

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 artifical data having response 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. Recall that with centered data, the ridge regression problem is given by

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

Notice that if we denote the augmented data set by $(\tilde{\mathbf{X}}, \tilde{\mathbf{y}})$, we have that

$$\begin{aligned}(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\beta)^\top(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\beta) &= \left(\begin{bmatrix}\mathbf{y}\\ \mathbf{0}\end{bmatrix} - \begin{bmatrix}\mathbf{X}\\ \sqrt{\lambda}\mathbf{I}\end{bmatrix}\beta\right)^\top \left(\begin{bmatrix}\mathbf{y}\\ \mathbf{0}\end{bmatrix} - \begin{bmatrix}\mathbf{X}\\ \sqrt{\lambda}\mathbf{I}\end{bmatrix}\beta\right) \\ &= \left(\begin{bmatrix}\mathbf{y}^\top & \mathbf{0}\end{bmatrix} - \beta^\top\begin{bmatrix}\mathbf{X}^\top & \sqrt{\lambda}\mathbf{I}\end{bmatrix}\right)\left(\begin{bmatrix}\mathbf{y}\\ \mathbf{0}\end{bmatrix} - \begin{bmatrix}\mathbf{X}\\ \sqrt{\lambda}\mathbf{I}\end{bmatrix}\beta\right) \\ &= \|\mathbf{y}\|^2 - \mathbf{y}^\top\mathbf{X}\beta - \beta^\top\mathbf{X}^\top\mathbf{y} + \beta^\top + \mathbf{X}^\top\mathbf{X}\beta + \lambda\|\beta\|^2 \\ &= (\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta) + \lambda\|\beta\|^2. \end{aligned} \tag{3.6}$$

This shows that the ridge regression is ordinary least squares on the augmented data, and hence completes the problem. ■

Exercise 3.13. *Derive the expression (3.62) and show that* $\hat{\beta}^{\text{pcr}}(p) = \hat{\beta}^{\text{ls}}$.

Comment. ■

Solution. Let $\mathbf{z}_0, \mathbf{z}_1 \ldots, \mathbf{z}_M$ denote the first M principal component directions of $\mathbf{X}$, with $\mathbf{z}_0 = (1, \ldots, 1)$. By definition, Principal component regression calls to regress $\mathbf{y}$ on the first M principal components. I.e.,

$$\hat{\mathbf{y}}^{\text{pcr}} = \overline{y}\mathbf{1} + \sum_{i=1}^{M} \frac{\langle \mathbf{z}_i, \mathbf{y}\rangle}{\langle \mathbf{z}_i, \mathbf{z}_i\rangle}\mathbf{z}_i := \overline{y}\mathbf{1} + \sum_{i=1}^{M} \hat{\theta}_i \mathbf{z}_i,$$

where we used that $\langle \mathbf{z}_0, \mathbf{z}_0\rangle = N$, and $\langle \mathbf{y}, \mathbf{z}_0\rangle = \sum_{i=1}^{N} y_i$. However, by definition the principal components are given by

$$\mathbf{z}_i = \mathbf{X}v_i,$$

where

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$$

and v_i denotes the i-th column of $\mathbf{V}$. Hence,

$$\overline{y}\mathbf{1} + \mathbf{X}\hat{\beta}^{\text{pcr}} = \hat{\mathbf{y}}^{\text{pcr}} = \overline{y}\mathbf{1} + \mathbf{X}\left(\sum_{i=1}^{M} \hat{\theta}_i v_i\right).$$

Since $\mathbf{X}$ has full rank, for any linear combination of its columns, we can uniquely determine the coefficients. Hence,

$$\hat{\beta}^{\text{pcr}}(M) = \sum_{i=1}^{M} \hat{\theta}_i v_i,$$

which was to be demonstrated.

When $M = p$, the solution is given by

$$\hat{\mathbf{y}}^{\text{pcr}}(p) = \overline{y}\mathbf{1} + \sum_{i=1}^{p} \frac{\langle \mathbf{z}_i, \mathbf{y}\rangle}{\sqrt{\langle \mathbf{z}_i, \mathbf{z}_i\rangle}} \frac{\mathbf{z}_i}{\sqrt{\langle \mathbf{z}_i, \mathbf{z}_i\rangle}} = \overline{y}\mathbf{1} + \sum_{i=1}^{p} \langle \mathbf{u}_i, \mathbf{y}\rangle \mathbf{u}_i = \overline{y}\mathbf{1} + \mathbf{U}\mathbf{U}^\top\mathbf{y},$$

where we have used that the i-th principal component direction points in the direction of $\mathbf{u}_i$, the i-th column of $\mathbf{U}$. Hence, normalizing $\mathbf{z}_i$, we obtain $\mathbf{u}_i$. Using Exercise 3.8., we see that the above can be written

$$\hat{\mathbf{y}}^{\text{pcr}}(p) = \overline{y}\mathbf{1} + \mathbf{Q}_2\mathbf{Q}_2\mathbf{y},$$

since both $\mathbf{Q}_2\mathbf{Q}_2$ and $\mathbf{UU}$ are orthogonal projections onto the same subspace. This can again be rewritten

$$\hat{\mathbf{y}}^{\text{pcr}}(p) = \mathbf{Q}\mathbf{Q}^\top\mathbf{y},$$

where $\mathbf{Q}$ comes from the QR decomposition of the full matrix $[\mathbf{1}, \mathbf{X}]$. It follows that $\hat{\mathbf{y}}^{\text{pcr}}(p) = \hat{\mathbf{y}}^{\text{ls}}$, the least squares solution. Again, since the columns of $\mathbf{X}$ are linearly independent, it follows that the linear combination of the columns of $\mathbf{X}$ uniquely determines the coefficients. Hence,

$$\hat{\beta}^{\text{pcr}} = \hat{\beta}^{\text{ls}},$$

which was to be demonstrated. ■

Exercise 3.14. *Show that in the orthogonal case, PLS stops after $m = 1$ steps, because subsequent $\hat{\varphi}_{mj}$ in step 2 in Algorithm 3.3 are zero.*

Solution. Notice that $\mathbf{z}_1$ is a linear combination of the columns, $\mathbf{x}_j$. Hence, we can compute in step $2d$ that

$$\mathbf{x}_j^{(1)} = \mathbf{x}_j - \frac{\hat{\varphi}_{0j}}{\sum_{k=1}^p \hat{\varphi}_{0k}^2} \sum_{i=1}^p \hat{\varphi}_{0i}\mathbf{x}_i.$$

Hence,

$$\hat{\varphi}_{1j} = \langle \mathbf{x}_j^{(1)}, \mathbf{y} \rangle = \hat{\varphi}_{0j} - \frac{\hat{\varphi}_{0j}}{\sum_{k=1}^p \hat{\varphi}_{0k}^2} \sum_{i=1}^p \hat{\varphi}_{0i}^2 = \hat{\varphi}_{0j} - \hat{\varphi}_{0j} = 0.$$

This shows that the coefficients in step 2 are 0, and hence that the algorithm terminates after the first step. ■

Comment. Note that since the matrix $\mathbf{X}$ is orthogonal, $\mathbf{z}_1$ is already the projection of $\mathbf{y}$ onto the subspace spanned by the columns of $\mathbf{X}$. Hence, the least squares solution is obtained in this case. ■

Exercise 3.15. *Verify expression* (3.64)*, and hence show that the partial least squares directions are a compromise between the ordinary regression coefficients and the principal component directions.*

Comment. The equation referenced above in the book is given by

$$\max_\alpha \text{Corr}^2(\mathbf{y}, \mathbf{X}\alpha)\text{Var}(\mathbf{X}\alpha), \quad \text{subject to } \|\alpha\| = 1, \alpha^\top \mathbf{S}\hat{\varphi}_\ell = 0, \ell = 0, 1, \ldots, m-1. \quad (3.64)$$

In the above, $\mathbf{S}$ is the sample variance-covariance matrix. We note that here the correlation and variance refer to the sample correlation and variance (i.e., these are not matrices). We emphasize that for this method, the columns of $\mathbf{X}$ are assume to be centered. Similarly, we will center the output $\mathbf{y}$ as well. The claim is that $\hat{\varphi}_m$, output by algorithm 3.3, solves the above. We emphasize that technically, $\hat{\varphi}_m$ will only be proportional to the above problem, since these vectors are not guaranteed to have norm 1. ■

Solution. We begin by simplifying the above expression to maximize. Since the data is centered, we have

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

Since the denominator is constant, we only need to maximize the numerator over α.

First, we check for $m = 1$, where there is no orthogonality constraint. To maximize the inner product, we simply have

$$\langle \mathbf{y}, \mathbf{X}\alpha\rangle = \langle \mathbf{X}^\top \mathbf{y}, \alpha\rangle,$$

from which it follows that $\alpha = (1/\|\mathbf{X}^\top \mathbf{y}\|)\mathbf{X}^\top \mathbf{y}$. Notice that indeed,

$$\hat{\varphi}_{1j} = \mathbf{x}_j^\top \mathbf{y},$$

by definition. This shows that (3.64) in the case $m = 1$.

Now, suppose (3.64) for $\ell = 1, 2, \ldots, m-1$. We emphasize that

$$\langle \hat{\varphi}_i, \hat{\varphi}_j\rangle_{\mathbf{S}} = 0$$

whenever $i \neq j$, where

$$\langle v, w\rangle_{\mathbf{S}} := \langle \mathbf{S}^{1/2} v, \mathbf{S}^{1/2} w\rangle$$

denotes the inner-product with respect to the symmetric positive definite matrix $\mathbf{S}$. Due to the orthogonality constraints, the quantity to maximize can be altered to the following:

$$\alpha^\top \left(\mathbf{X}^\top \mathbf{y} - c_1 \mathbf{S}\hat{\varphi}_1 - \cdots - c_{m-1}\mathbf{S}\hat{\varphi}_{m-1}\right).$$

Indeed, all terms except the first vanish due to the orthogonality constraints. Moreover, this is true for any choice of constants $c_1, \ldots, c_{m-1}$. Hence, if there exist a choice of constants $c_1, \ldots, c_{m-1}$ such that

$$\mathbf{X}^\top \mathbf{y} - c_1 \mathbf{S}\hat{\varphi}_1 - \cdots - c_{m-1}\mathbf{S}\hat{\varphi}_{m-1}$$

satisfies the orthogonality conditions, then the solution α will be the normalized vector pointing in the direction of

$$\mathbf{X}^\top \mathbf{y} - c_1 \mathbf{S}\hat{\varphi}_1 - \cdots - c_{m-1}\mathbf{S}\hat{\varphi}_{m-1},$$

since the inner-product of two vectors of fixed norm will always be maximized whenever they point in the same direction. Let us use the Gram-Schmidt procedure to orthogonalize (in $\langle\cdot,\cdot\rangle_{\mathbf{S}}$ inner-product) $\mathbf{X}^\top \mathbf{y}$ with respect to mutually orthogonal vectors $\{\hat{\varphi}_i\}_{i=1}^{m-1}$. I.e., consider the vector

$$\hat{\alpha} = \mathbf{X}^\top \mathbf{y} - \sum_{i=1}^{m-1} c_i \hat{\varphi}_i,$$

where $c_i = \langle \mathbf{X}^\top \mathbf{y}, \hat{\varphi}_i\rangle_{\mathbf{S}} / \langle \hat{\varphi}_i, \hat{\varphi}_i\rangle_{\mathbf{S}}$. By the Gram-Schmidt procedure, for any $i = 1, \ldots, m-1$, we have that

$$\langle \hat{\alpha}, \hat{\varphi}_i\rangle_{\mathbf{S}} = \hat{\alpha}^\top \mathbf{S}\hat{\varphi}_i = 0.$$

It follows that $\hat{\alpha}/\|\hat{\alpha}\|$ solves the maximization problem. It remains use iterative algebra and the fact that

$$\hat{\varphi}_{\ell j} = \langle \mathbf{x}_j^{(\ell-1)}, \mathbf{y}\rangle, \qquad \mathbf{x}_j^{(\ell)} = \mathbf{x}_j^{(\ell-1)} - [\langle \mathbf{z}_\ell, \mathbf{x}_j^{\ell-1}\rangle / \langle \mathbf{z}_\ell, \mathbf{z}_\ell\rangle]\mathbf{z}_\ell$$

where $\mathbf{z}_\ell = \sum_{j=1}^p \hat{\varphi}_{\ell j} \mathbf{x}_j^{(\ell-1)}$, to conclude that the result $\hat{\alpha}$ is indeed a multiple of $\hat{\varphi}_m$. This completes the proof of (3.64).

Since ordinary least squares maximizes the correlation between $\mathbf{y}$ and $\mathbf{X}\beta$, while principal component regression maximizes the variance $\mathrm{Var}(\mathbf{X}\beta)$, subject to orthogonality conditions, one can see that partial least squares, which maximizes the product subject to orthogonality conditions, is in this sense a compromise of the two notions. ■

Exercise 3.16. *Derive the entries in Table 3.4, the explicit forms for estimators in the orthogonal case.*

Solution. We begin by showing the formula for the best subset of size M. Note that best subset selection corresponds to keeping the M columns of $\mathbf{X}$, which are obtained iteratively by Exercise 3.9. I.e., we check

$$\mathrm{argmax}_i |\langle \mathbf{y}, \mathbf{q}_i \rangle|,$$

where this is the i-th column of Q in the QR decomposition of $\mathbf{X}$. Since $\mathbf{X}$ has orthonormal columns already, $\mathbf{q}_i = \mathbf{x}_i$. Let $\tilde{\mathbf{X}}$ denote the $N \times M$ matrix obtained by keeping only the columns for which the value $\mathbf{x}_i^\top \mathbf{y}$ was among the M largest values. The solution is then given by

$$\hat{\tilde{\beta}} = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \mathbf{y} = \tilde{\mathbf{X}}^\top \mathbf{y},$$

since the columns of $\tilde{\mathbf{X}}$ are still orthonormal. However, note that the ordinary least squares solution is given by

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

Moreover, recall that the absolute value of the $\hat{\beta}_i$, $|\mathbf{x}_i^\top \mathbf{y}|$, is the quantity being maximized above. Hence, the solution is given by the set of $\hat{\beta}_j$ such that $\hat{\beta}_j \geq \hat{\beta}_{(M)}$, where $\hat{\beta}_{(M)}$ denotes the M-th largest component in the vector $\hat{\beta}$. This can be written

$$\hat{\beta}_j \cdot \mathbf{I}\left(|\hat{\beta}_j| \geq |\hat{\beta}_{(M)}| \right),$$

which was to be demonstrated.

For ridge regression, the solution is given by

$$\hat{\beta}^{\mathrm{ridge}} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y} = (1+\lambda)^{-1} \mathbf{X}^\top \mathbf{y} = (1+\lambda)^{-1} \hat{\beta}.$$

This shows immediately that the coefficients are given by

$$\frac{\hat{\beta}_j}{1+\lambda}.$$

For the lasso (assuming centered data), we have to minimize the quantity

$$\frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) + \lambda \|\beta\|_{L^1}$$

over all β. Expanding the above and simplifying yields

$$\frac{1}{2}\mathbf{y}^\top \mathbf{y} - \langle \hat{\beta}, \beta \rangle + \frac{1}{2}\|\beta\|^2 + \lambda \|\beta\|_{L^1},$$

where in the above we used that $\mathbf{X}^\top\mathbf{X} = \mathbf{I}$, and $\mathbf{X}^\top\mathbf{y} = \hat{\beta}$. The term $\frac{1}{2}\mathbf{y}^\top\mathbf{y}$ is constant, and hence it suffices to minimize

$$\frac{1}{2}\|\beta\|^2 + \lambda\|\beta\|_{L^1} - \langle\hat{\beta}, \beta\rangle.$$

To take the derivative w.r.t. β_j, we need to consider the case when $\beta_j > 0$, and $\beta_j < 0$, as $|\beta_j|$ is not differentiable at 0.

(1) ($\beta_j > 0$) The above expression is

$$\frac{1}{2}\sum_{i=1}^{p}\beta_i^2 + \lambda\beta_j + \lambda\sum_{i\neq j}|\beta_i| - \sum_{i=1}^{p}\beta_i\hat{\beta}_i.$$

Setting the derivative of the above equation w.r.t. β_j equal to 0 yields

$$\beta_j + \lambda - \hat{\beta}_j = 0.$$

In this case we have

$$\beta_j = \hat{\beta}_j - \lambda.$$

For this to be a minimum, we require $\hat{\beta}_j - \lambda$ to be positive. This is easily seen by looking at the above expression as a quadratic polynomial of β_j.

(2) ($\beta_j < 0$) Now the expression becomes

$$\frac{1}{2}\sum_{i=1}^{p}\beta_i^2 - \lambda\beta_j + \lambda\sum_{i\neq j}|\beta_i| - \sum_{i=1}^{p}\beta_i\hat{\beta}_i.$$

Setting the derivative of the above equation w.r.t. β_j equal to 0 yields

$$\beta_j = \hat{\beta}_j + \lambda.$$

Similarly, for the quantity to be a minimum, and not a maximum, we require the above to be positive.

Let β_j denote the minimizer for the lasso obtained above. We can rewrite both cases as

$$\beta_j = (\hat{\beta}_j - \text{sign}(\beta_j)\lambda)_+.$$

Since $\hat{\beta}_j = \text{sign}(\hat{\beta}_j)|\hat{\beta}_j|$, it only remains to show that $\text{sign}(\beta_j) = \text{sign}(\hat{\beta}_j)$. Since β_j minimizes the polynomial

$$\frac{1}{2}x^2 \pm \lambda x - \hat{\beta}_j x,$$

we see that their signs must agree. Indeed, if $\hat{\beta}_j > 0$ and $\beta_j < 0$, we obtain that

$$\frac{1}{2}\beta_j^2 - (\lambda + \hat{\beta}_j)\beta_j,$$

is a minimum, where $(\lambda + \hat{\beta}_j) > 0$. However, choosing β_j to be larger (closer to 0 but still negative), we achieve a smaller value. This contradicts that β_j minimizes the above quantity. The same reasoning results in a contradiction if $\hat{\beta}_j < 0$ and $\beta_j > 0$. This shows

$$\beta_j = \text{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda)_+,$$

which was to be demonstrated. This completes the problem. ■

Exercise 3.17. *Repeat the analysis of Table 3.3 on the spam data discussed in Chapter* 1.

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. Conjugate gradient algorithms provide a method of solution to the problem $\mathbf{X}\beta = \mathbf{y}$ by iteratively finding "better guesses" $\beta^{(1)}, \ldots, \beta^{(m)}$, while PLS iteratively finds solutions $\hat{\mathbf{y}}^{(1)}, \ldots, \hat{\mathbf{y}}^{(m)}$. They are related in the sense that

$$\hat{\mathbf{y}}^{(m)} = \mathbf{X}\beta^{(m)}.$$

■

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. Equation (3.48) is [2] is given by

$$\mathbf{X}\hat{\beta}^{\text{ridge}} = \sum_{j=1}^{p} \mathbf{u}_j \frac{d_j^2}{d_j^2 + \lambda} \mathbf{u}_j^\top \mathbf{y}, \tag{3.47}$$

where $\mathbf{X} = \mathbf{UDV}$ is the singular value decomposition of $\mathbf{X}$, and $\mathbf{D}$ is diagonal with entries d_i. From this we see

$$\|\hat{\beta}^{\text{ridge}}\|^2 \geq \frac{1}{d_{\max}^2} \|\mathbf{X}\hat{\beta}^{\text{ridge}}\|^2 = \frac{1}{d_{\max}^2} \sum_{j=1}^{p} \frac{d_j^2 \langle \mathbf{u}_j, \mathbf{y} \rangle}{d_j^2 + \lambda}.$$

This shows clearly that $\|\hat{\beta}^{\text{ridge}}\|^2$ increases as $\lambda \to 0$.

For the lasso, notice that (3.84) in [2] for estimating the lasso coefficients is given by

$$\tilde{\beta}_j(\lambda) \leftarrow S\left(\sum_{i=1}^{N} x_{ij}(y_i - \tilde{y}_i^{(j)}), \lambda \right) \tag{3.84}$$

where $S(t, \lambda) = \text{sign}(t)(|t| - \lambda)_+$. This does have a norm increasing as $\lambda \to 0$. This suggests $\|\hat{\beta}^{\text{lasso}}\|$ does get larger as $\lambda \to 0$.

From studying the method of solution for PLS in Exercise 3.15, we see that

$$\|\mathbf{X}^\top \mathbf{y}\|^2 = \|c_1 \mathbf{S}\hat{\varphi}_1 - \cdots + c_{m-1}\mathbf{S}\hat{\varphi}_{m-1} + \hat{\varphi}_m\|^2.$$

Using that $\hat{\varphi}_m$ is orthogonal to all the other terms, we can compute

$$\|\mathbf{X}^\top \mathbf{y}\|^2 = \|c_1 \mathbf{S}\hat{\varphi}_1 + \cdots + c_{m-1}\mathbf{S}\hat{\varphi}_{m-1}\|^2 + \|\hat{\varphi}_m\|^2.$$

Since the left-hand-side is constant, one would suspect that the norms do not blow up with successive steps in the algorithm. ■

Exercise 3.20. *Consider the canonical correlation problem (3.67). Show that the leading pair of canonical variates u_1, v_1 solve the problem*

$$\max_{u^\top(\mathbf{Y}^\top\mathbf{Y})u=1, v^\top(\mathbf{X}^\top\mathbf{X})v=1} u^\top(\mathbf{Y}^\top\mathbf{X})v, \tag{3.86}$$

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

$$(\mathbf{Y}^\top\mathbf{Y})^{-1/2}(\mathbf{Y}^\top\mathbf{X})(\mathbf{X}^\top\mathbf{X})^{-1/2} = \mathbf{U}^*\mathbf{D}^*(\mathbf{V}^*)^\top \tag{3.87}$$

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

Solution. Assuming centered data, the canonical correlation problem (3.67) can be rewritten as the maximization of

$$\text{Corr}^2(\mathbf{Y}u, \mathbf{X}v) = \frac{\langle \mathbf{Y}u, \mathbf{X}v\rangle}{\|\mathbf{Y}u\|\|\mathbf{X}v_i\|} = \frac{u^\top \mathbf{Y}^\top \mathbf{X}v}{\|\mathbf{Y}u\|\|\mathbf{X}v\|}.$$

Notice that the above is invariant under scaling u or v by positive constants. Hence, we can replace u and v by multiples of themselves ensuring that both terms in the denominator are 1. The maximization problem then becomes

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

which was to be demonstrated. Rewriting in terms of $\tilde{u} = (\mathbf{Y}^\top\mathbf{Y})^{-1/2}u$, $\tilde{v} = (\mathbf{X}^\top\mathbf{X})^{-1/2}u$, the problem becomes

$$\max_{\|\tilde{u}\|,\|\tilde{v}\|=1} \langle A\tilde{v}, \tilde{u}\rangle,$$

where

$$A = (\mathbf{Y}^\top\mathbf{Y})^{-1/2}(\mathbf{Y}^\top\mathbf{X})(\mathbf{X}^\top\mathbf{X})^{-1/2}.$$

By Cauchy-Schwarz, we have that

$$\langle A\tilde{v}, \tilde{u}\rangle \leq \|A\tilde{v}\|\|\tilde{u}\| = \|A\tilde{v}\|.$$

Recall that

$$\max_{\|\tilde{v}\|=1}\sqrt{\langle A\tilde{v}, A\tilde{v}\rangle} = \max_{\|\tilde{v}\|=1}\sqrt{\langle A^\top A\tilde{v}, \tilde{v}\rangle} = \sigma_1,$$

the leading singular value of A. Hence,

$$\max_{\|\tilde{u}\|,\|\tilde{v}\|=1} \langle A\tilde{v}, \tilde{u}\rangle \leq \sigma_1.$$

It suffices to find a pair of vectors on which this maxima is obtained. Writing $A = \sum_{i=1}^{\min(k,p)} \sigma_i u_i^*(v_i^*)^\top$, we see that choosing $\tilde{v} = v_1^*$, $\tilde{u} = u_1^*$ yields the maximum. Tracing back the original definition of $\tilde{u}$ and $\tilde{v}$ gives the result.

For the successive maxima, note that the requirement that the linear combination $\mathbf{X}v_2$ be uncorrelated to $\mathbf{X}v_1$ can be rewritten as

$$0 = \langle \mathbf{X}v_1, \mathbf{X}v_2\rangle = \langle \mathbf{X}^\top\mathbf{X}v_1, v_2\rangle = \langle (\mathbf{X}^\top\mathbf{X})^{1/2}\tilde{v}_1, v_2\rangle = \langle \tilde{v}_1, \tilde{v}_2\rangle.$$

Similarly for u_1 and u_2. Hence, the maximization at step 2 can be rewritten as

$$\max_{\|\tilde{u}\|,\|\tilde{v}\|=1,\tilde{u}\perp\tilde{u}_1,\tilde{v}\perp\tilde{v}_1} \langle A\tilde{v}, \tilde{u}\rangle.$$

Note that, similar to before, using Cauchy-Schwarz we have

$$\max_{\|\tilde{u}\|,\|\tilde{v}\|=1,\tilde{u}\perp\tilde{u}_1,\tilde{v}\perp\tilde{v}_1} \langle A\tilde{v}, \tilde{u}\rangle \leq \max_{\|\tilde{v}\|=1,\tilde{v}\perp\tilde{v}_1} \sqrt{\langle A^\top A\tilde{v}, \tilde{v}\rangle}.$$

Again, from Rayleigh quotients, it is known that the above problem has a maximum of σ_2. Hence

$$\max_{\|\tilde{u}\|,\|\tilde{v}\|=1,\tilde{u}\perp\tilde{u}_1,\tilde{v}\perp\tilde{v}_1} \langle A\tilde{v}, \tilde{u}\rangle \leq \sigma_2.$$

Using the same decomposition of A as before, we see this maximum is obtained when $\tilde{v} = v_2^*$, and $\tilde{u} = u_2^*$. The exact same reasoning repeated $\min(K, p)$ times, each time picking up extra orthogonality conditions, yields the result. ■

Exercise 3.21. *Show that the solution to the reduced-rank regression problem (3.68), with* $\mathbf{\Sigma}$ *estimated by* $\mathbf{Y}^\top\mathbf{Y}/N$*, is given by (3.69). Hint: Transform* $\mathbf{Y}$ *to* $\mathbf{Y}^* := \mathbf{Y}\mathbf{\Sigma}^{-1/2}$*, and solve in terms of the canonical vectors* u_m^**. Show that* $\mathbf{U}_m = \Sigma^{-1/2}\mathbf{U}_m^*$*, and a generalized inverse is* $(\mathbf{U}_m^-)^\top\mathbf{\Sigma}^{1/2}$*.*

Solution. This problem will be completed at a later date. ■

Exercise 3.22. *Show that the solution in Exercise 3.21 does not change if* $\mathbf{\Sigma}$ *is estimated by the more natural quantity* $(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^\top(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})/(N - pK)$*.*

Solution. This problem will be completed at a later date. ■

Exercise 3.23. *Consider a regression problem with all variables and response having mean zero and standard deviation one. Suppose a;so that each variable has identical absolute correlation with the response:*

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

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.*

(a) Show that

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

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

(b) Show that these correlations are all equal to

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

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 monotonically decreasing, as claimed in (3.55)*.*

Comment. The notation around equation (3.55) is given as follows. LAR is an iterative method, and so $\mathcal{A}_k$ denotes the set of variables in the model at the beginning of step k, while $\beta_{\mathcal{A}_k}$ denotes the coefficient vector for these variables at this step. $\beta_{\mathcal{A}_k}$ has length k, and $k-1$ of these values are nonzero, at the beginning of step k. Defining $\mathbf{r}_k := \mathbf{y} - \mathbf{X}_{\mathcal{A}_k}\beta_{\mathcal{A}_k}$, equation (3.55) is given by

$$\delta_k = (\mathbf{X}_{\mathcal{A}_k}^\top\mathbf{X}_{\mathcal{A}_k})^{-1}\mathbf{X}_{\mathcal{A}_k}^\top\mathbf{r}_k \tag{3.55}$$

. As [2] claims, the coefficient profile is then adjusted in this direction by $\beta_{\mathcal{A}_k}(\alpha) = \beta_{\mathcal{A}_k} + \alpha \cdot \delta_k$. ■

Solution. (a) Recall that $\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$. From this it follows that

$$\mathbf{x}_j^\top\mathbf{X}\hat{\beta} \;=\; \mathbf{x}_j^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \tag{3.7}$$

Notice that $\mathbf{x}_j^\top \mathbf{X}$ is the j-th row of $\mathbf{X}^\top\mathbf{X}$. Hence, $\mathbf{x}_j^\top \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}$ is a $1 \times p$ row vector with a 1 in the jth position, and zeros elsewhere. It follows that $\mathbf{x}_j^\top \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is again $\mathbf{x}_j^\top$. Hence,

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

which was to be demonstrated.

(b) The correlation differs by factors depending on the norms of $\mathbf{x}_j$ and $\mathbf{y} - \mathbf{u}(\alpha)$. We have

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

since the $\mathbf{x}_j$ have standard deviation 1. Simplifying the denominator further, we have

$$\langle \mathbf{y} - \mathbf{u}(\alpha), \mathbf{y} - \mathbf{u}(\alpha)\rangle = \langle \mathbf{y}, \mathbf{y}\rangle - 2\alpha\langle \mathbf{y}, \mathbf{X}\hat{\beta}\rangle + \alpha^2\langle \mathbf{X}\hat{\beta}, \mathbf{X}\hat{\beta}\rangle.$$

But $\langle \mathbf{X}\hat{\beta}, \mathbf{X}\hat{\beta}\rangle = \langle \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}, \mathbf{X}\hat{\beta}\rangle = \langle \mathbf{y}, \mathbf{X}\hat{\beta}\rangle$. Hence,

$$\|\mathbf{y} - \mathbf{u}(\alpha)\|^2 = \langle \mathbf{y}, \mathbf{y}\rangle + \alpha(\alpha-2)\langle \mathbf{y}, \mathbf{X}\hat{\beta}\rangle.$$

Notice from the above that that RSS $= \langle \mathbf{y}, \mathbf{y}\rangle - \langle \mathbf{y}, \mathbf{X}\hat{\beta}\rangle$.. We add and subtract $\alpha(2-\alpha)$RSS. This yields

$$\begin{aligned}\|\mathbf{y} - \mathbf{u}(\alpha)\|^2 &= \langle \mathbf{y}, \mathbf{y}\rangle - \alpha(2-\alpha)\langle \mathbf{y}, \mathbf{y}\rangle + \alpha(2-\alpha)\mathrm{RSS} \\ &= N - \alpha(2-\alpha)N + \alpha(2-\alpha)\mathrm{RSS} \\ &= N(1-\alpha)^2 + \alpha(2-\alpha)\mathrm{RSS}..\end{aligned}$$

Dividing by N and taking the square root yields the result.

(c) Part (b) shows that when $\alpha = 0$, the correlations are λ, and when $\alpha = 1$, we have $\lambda(\alpha) = 0$. Moreover, one can inspect $\frac{d\lambda}{d\alpha}$ to see the derivative is negative for all $\alpha \in [0, 1]$. It follows that the correlations are monotonically decreasing, as claimed in (3.55).

■

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. For simplicity, let $\mathbf{X}_{\mathcal{A}_k}$ be denoted by $\mathbf{X}$ an $N \times k$ matrix, with column $\mathbf{x}_k$ just added. Let $\mathbf{x}_j$ be any one of its columns. Similarly let $\mathbf{u} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{r}$, where $\mathbf{r}$ is the residual at the current kth step. Let θ_j denote the angle between $\mathbf{x}_j$ and $\mathbf{u}$, and let ϕ_j be the angle between $\mathbf{r}$.

$$\cos\theta_j = \frac{\langle \mathbf{x}_j, \mathbf{u}\rangle}{\|\mathbf{u}\|} = \frac{\mathbf{x}_j^\top \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{r}}{\|\mathbf{u}\|} = \frac{\mathbf{x}_j^\top\mathbf{r}}{\|\mathbf{u}\|} = \frac{\|\mathbf{r}\|}{\|\mathbf{u}\|} = \cos\phi_j,$$

where in the above we used the same reasoning as part (a) in Exercise 3.23 to simplify $\mathbf{x}_j^\top \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. By definition, newest predictor $\mathbf{x}_k$ has equal absolute correlation with the current residual as all the other predictors, and hence $\phi_1 = \phi_2 = \cdots = \phi_k$, up to a possible phase shift of π radians. From this it follows that $\theta_1 = \theta_2 = \cdots = \theta_j$, up to possible phase shifts of π radians. I.e., negative correlation vs. positive correlation. Hence $\mathbf{u}$ makes an equal angle with all predictors (again, up to a possible phase shift of π radians). ■

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. For each $\mathbf{x}_j$, let $c_{j,1} = \langle \mathbf{x}_j, \mathbf{r}_k \rangle$, and $c_{j,2} = \langle \mathbf{x}_j, \mathbf{u}_k \rangle$. For any $i, j \in \mathcal{A}_k$, we have by the previous problem that

$$c_{i,1} = \pm c_{j,1}, \quad c_{i,2} = \pm c_{j,2}.$$

Moreover, $c_{j,1}$ and $c_{j,2}$ have the same sign. Denote the positive versions of these constants by c_1 and c_2 respectively. For any j, the covariance between $\mathbf{x}_j$ and the residual dependent on α is given by

$$\langle \mathbf{x}_j, \mathbf{r}_k(\alpha) \rangle = \langle \mathbf{x}_j, \mathbf{y} - \hat{\mathbf{y}}_k - \alpha \mathbf{u}_k \rangle = \langle \mathbf{x}_j, \mathbf{r}_k \rangle - \alpha \langle \mathbf{x}_j, \mathbf{u}_k \rangle = c_{1,j} - \alpha c_{2,j}.$$

To find the α for which $\mathbf{x}_j$ for $j \notin \mathcal{A}_k$, has correlation equal to the predictors in the current set, we simply solve the equation

$$\left| \frac{c_{1,j} - \alpha c_{2,j}}{\|\mathbf{x}_j\| \|\mathbf{r}_k(\alpha)\|} \right| = \left| \frac{c_1 - \alpha c_2}{\|\mathbf{x}_{i^*}\| \|\mathbf{r}_k(\alpha)\|} \right|.$$

for any $i^* \in \mathcal{A}_k$. By the equal variance assumption, we can cancel the denominators. If $c_{1,j} - \alpha c_{2,j}$ has the same sign as $c_1 - \alpha c_2$, we solve for α and obtain

$$\alpha_j = \frac{c_1 - c_{1,j}}{c_2 - c_{2,j}}.$$

If the signs are flipped, then we have

$$\alpha'_j = \frac{c_1 + c_{1,j}}{c_{2,j} + c_2}.$$

We see that the required stepsize is given by

$$\min_j \mathbf{1}_{S_j} \alpha_j + \mathbf{1}_{S_j^c} \alpha'_j$$

where $\mathbf{1}_{S_j}$ denotes the characteristic function of the set $S_j = \{x \in [0,1] : \text{sign}(c_1 - x c_2) = \text{sign}(c_{1,j} - x c_{2,j})$. Moreover, the variable added will be the argument for which the above minimum is achieved. ■

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 $x_{j.\mathcal{A}}$ 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 the largest in magnitude.]*

Solution. Note first that if we can show the hint is true, then we are finished. Indeed, by the previous problem, LAR adds the index j for which a line of the form $c_{j,1}(1 - c_{j,2}\alpha)$ intersects the line $C_1(1 - C_2\alpha)$ for the smallest value of α, while Forward stepwise regression simply computes which $\mathbf{x}_j$ has the largest correlation with the residual (which is equivalent to simply choosing the j for which $c_{j,1}$ above is the largest).

To see that the hint is true, consider the forward-stepwise regression model. Since $x_{j.\mathcal{A}}$ is "linearly adjusted for all other variables in the model", this is to say that $\mathbf{x}_j$ is orthogonalized with respect to other variables in the model. Hence,

$$\text{argmax}_j \text{Corr}(x_{j.\mathcal{A}}, \mathbf{r}) = \text{argmax}_j \frac{\langle \mathbf{r}, x_{j.\mathcal{A}} \rangle}{\|\mathbf{r}\| x_{j.\mathcal{A}}\|} = \text{argmax}_j \langle \mathbf{r}, \frac{x_{j.\mathcal{A}}}{\|x_{j.\mathcal{A}}\|} \rangle.$$

This is precisely the algorithm derived in Exercise 3.9. This completes the problem. ■

Exercise 3.27. *Lasso and LAR: Consider the lasso problem with Lagrange multiplier form: with* $L(\beta) = \frac{1}{2}\sum_i (y_i - \sum_j x_{ij}\beta_j)^2$, *we minimize*

$$L(\beta) + \lambda \sum_j |\beta_j| \tag{3.88}$$

for fixed $\lambda > 0$.

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

$$L(\beta) + \lambda \sum_j (\beta_j^+ + \beta_j^-) - \sum_j \lambda_j^+ \beta_j^+ - \sum_j \lambda_j^- \beta_j^- \tag{3.89}$$

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:*

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

Hence show that for any "active" predictor having $\beta_j \neq 0$, *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* j*th 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*

$$\hat{\beta}(\lambda) = \hat{\beta}(\lambda_0) - (\lambda - \lambda_0)\gamma_0 \tag{3.90}$$

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

Solution. (a) Note that if $\beta_j \leq 0$, then $\beta_j = -\beta_j^-$, while if $\beta_j \geq 0$, we have $\beta = \beta_j^+$. Hence, in either case $|\beta_j| = \beta_j^+ + \beta_j^-$. It follows that (3.88) becomes

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

Minimizing this subect to $\beta_j^+, \beta_j^- \geq 0$ by definition gives the lagrange dual function

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

which was to be demonstrated. The first two KKT conditions are obtained by simply setting the derivative of the above expression taking with respect to β_j^+ and β_j^- equal to 0 , while the last two come from the slackness conditions. The partial w.r.t. β_j^+ yields

$$\sum_i (y_i - \sum_j x_{ij}\beta_j) \cdot x_{ij} + \lambda - \lambda_j^+ = 0,$$

which is the first KKT condition since $\sum_i (y_i - \sum_j x_{ij}\beta_j) \cdot x_{ij}$ is indeed the jth component of $\nabla L(\beta)$. Similarly, since $\beta_j = \beta_j^+ - \beta_j^-$, the derivative of (3.89) w.r.t. β_j^- yields

$$\sum_i (y_i - \sum_j x_{ij}\beta_j) \cdot (-x_{ij}) + \lambda - \lambda_j^- = 0,$$

which is the second KKT conditions. The last two equations are precisely the slackness constraints, since the terms $-\sum_j \lambda_j^+ \beta_j^+$ and $-\sum_j \lambda_j^- \beta_j^-$ in (3.89) come from the non-negativity constrains on β_j^+, β_j^- respectively. This completes part (a).

(b) Adding the first two KKT conditions and solving for λ yields

$$\lambda = \frac{\lambda_j^+ + \lambda_j^-}{2}.$$

Now subtracting the second equation from the first yields

$$\nabla L(\beta)_j = \frac{\lambda_j^+ - \lambda_j^-}{2}.$$

Hence

$$|\nabla L(\beta)_j| \leq \frac{\lambda_j^+ + \lambda_j^-}{2} = \lambda,$$

by non-negativity of λ_j^-, λ_j^+. If $\lambda = 0$, then the first equation shows $\nabla L(\beta)_j$ is non-negative, while the second shows it is non-positive. Hence, $\nabla L(\beta)_j = 0$ for all j in this case. If $\lambda > 0$ and $\beta_j^+ > 0$, then the third equation shows $\lambda_j^+ = 0$, in which case the first equation shows $\nabla L(\beta)_j = -\lambda$. This shows that $\lambda_j^- = 2\lambda > 0$, and hence that $\beta_j^- = 0$ from the fourth KKT condition. The exact same reasoning shows the third implication holds true. The above 3 cases exhaust all situations where $\beta_J \neq 0$. This shows $\nabla L(\beta)_j = \pm\lambda$, depending on β_j. (i.e., $\beta_j < 0$ when $\beta_j^- > 0$, and $\beta_j > 0$ when $\beta_j^+ > 0$.) Writing

$$L(\beta) = \frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta),$$

we have that

$$\nabla L(\beta) = \frac{1}{2}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\beta).$$

I.e,

$$\nabla L(\beta)_j = \frac{1}{2}\mathbf{x}_j^\top(\mathbf{y} - \mathbf{X}\beta) = \frac{1}{2}\langle \mathbf{x}_j, \mathbf{y} - \mathbf{X}\beta\rangle.$$

Denoting $\mathbf{r} = \mathbf{y} - \mathbf{X}\beta$ as the current residual, we have

$$\text{Corr}(\mathbf{x}_j, \mathbf{r}) = \frac{\langle \mathbf{x}_j, \mathbf{r}\rangle}{\|\mathbf{x}_j\|\|\mathbf{r}\|} = \frac{-2\text{sign}(\beta_j)\lambda}{\|\mathbf{r}\|},$$

where we used that the predictors each have variance 1.

(c) Let $\text{sign}(\beta)$ denote the $p \times 1$ vector with entries ± 1 or 0, depending on the whether β_j is positive, negative, or 0. Then the equation obtained in the previous part can be rewritten as

$$\mathbf{X}^\top\mathbf{y} - \mathbf{X}^\top\mathbf{X}\beta = -2\lambda\mathbf{v} + \mathbf{v}',$$

where $\mathbf{v}, \mathbf{v}'$ are $p \times 1$ vectors with entries $\mathbf{v}_j = \text{sign}(\beta_j)$ for nonzero β_j, and 0 otherwise, while $\mathbf{v}'_j = 0$ for nonzero β_j, and $\mathbf{x}_j^\top\mathbf{y}$ otherwise. Re-arranging and solving for β yields

$$\beta(\lambda) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} + 2\lambda(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{v} - \mathbf{v}'.$$

Since the active set of predictors does not change from λ_0 to λ_1, it follows that $\mathbf{v}', \mathbf{v}$ are constant vectors for $\lambda \in [\lambda_1, \lambda_0]$. It follows that

$$\beta(\lambda) = \beta(\lambda_0) - (\lambda - \lambda_0)(\mathbf{X}^\top\mathbf{X})^{-1}(-2\mathbf{v}),$$

which was to be demonstrated.

■

Exercise 3.28. *Suppose for a given t in (3.51), the fitted lasso coefficient for variable X_j is $\hat{\beta}_j = a$. Suppose we augment our set of variables with an identical copy $X_j^* = X_j$. Characterize the effect of this exact collinearity by describing the set of solutions for $\hat{\beta}_j$ and $\hat{\beta}_j^*$, using the same value of t.*

Solution. The lasso problem attempts to minimize

$$\frac{1}{2}\sum_i (y_i - x_{ij}\beta_j - \sum_{k\neq j} x_{ik}\beta_k)^2 + t|\beta_j| + t\sum_{i\neq j}|\beta_i|.$$

Denote the above by $L(\beta)$, where $\beta = (\beta_1, \ldots, \beta_j, \ldots, \beta_p)^\top$. This has solution $\hat{\beta}$, for which we are told $(\hat{\beta})_j = a$. The corresponding problem with the above collinearity is given by

$$\begin{aligned}
&= \frac{1}{2}\sum_i (y_i - x_{ij}(\beta_j^* + \beta_j) - \sum_{k\neq j} x_{ik}\beta_k)^2 + t(|\beta_j^*| + |\beta_j|) + t\sum_{i\neq j}|\beta_i| \\
&= \frac{1}{2}\sum_i (y_i - x_{ij}(\beta_j^* + \beta_j) - \sum_{k\neq j} x_{ik}\beta_k)^2 + t(|\beta_j^* + \beta_j|) + t\sum_{i\neq j}|\beta_i| + t(|\beta_j^*| + |\beta_j| - |\beta_j^* + \beta_j|) \\
&= L(\beta') + t(|\beta_j^*| + |\beta_j| - |\beta_j^* + \beta_j|),
\end{aligned}$$

where $\beta' = (\beta_1, \ldots, \beta_j + \beta_j^*, \ldots, \beta_p)^\top$. Notice that if we choose a solution $\hat{\beta}'$ for the above which is optimal for $L(\beta')$, and satisfies that $t(|\beta_j^*| + |\beta_j| - |\beta_j^* + \beta_j|) = 0$, then we know this would be an optimal solution. Indeed,

$$\min_{\beta'} \left(L(\beta') + t(|\beta_j^*| + |\beta_j| - |\beta_j^* + \beta_j| \right) \geq \min_{\beta'} L(\beta').$$

Hence, a solution that minimizes $L(\beta')$ and keeps the extra penalty 0 achieves equality. Choosing $\beta_j = \beta_j^* = a/2$, and all other terms the same, certainly minimizes $L(\beta')$, since $a/2 + a/2 = a$. Moreover, with this choice the extra penalty term above becomes 0. By the above reasoning, we see that this is an optimal solution. Hence, the solution for β_j, and β_j^* using the same value of t is given by $\beta_j = \beta_j^* = a/2$. This completes the problem. ■

Exercise 3.29. *Suppose we run a ridge regression with parameter λ on a single variable X and get coefficient a. We now include an exact copy $X^* = X$, and refit our ridge regression. Show that both coefficients are identical, and derive their value. Show in general that if m copies of a variable X_j are included in a ridge regression, their coefficients are all the same.*

Solution. Ridge regression has a closed form solution given by

$$\hat{\beta}^{\text{ridge}} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}.$$

If $\mathbf{X}_1$ is simply a $N \times 1$ matrix with response $\mathbf{y}$, this reduces to

$$a = \hat{\beta}^{\text{ridge}} = \frac{\langle \mathbf{X}_1, \mathbf{y} \rangle}{\lambda + \|\mathbf{X}_1\|^2}.$$

Now letting $\mathbf{X}$ denote the $N \times 2$ matrix with repeated columns $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_1]$, we have that

$$\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I} = \begin{bmatrix} \|\mathbf{X}_1\|^2 + \lambda & \|\mathbf{X}_1\|^2 \\ \|\mathbf{X}_1\|^2 & \|\mathbf{X}_1\|^2 + \lambda \end{bmatrix},$$

and hence

$$(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} = \frac{1}{2\lambda \|\mathbf{X}_1\|^2 + \lambda^2} \begin{bmatrix} \|\mathbf{X}_1\|^2 + \lambda & -\|\mathbf{X}_1\|^2 \\ -\|\mathbf{X}_1\|^2 & \|\mathbf{X}_1\|^2 + \lambda \end{bmatrix},$$

so we can compute

$$\hat{\beta}^{\text{ridge}} = \frac{1}{2\lambda \|\mathbf{X}_1\|^2 + \lambda^2} \begin{bmatrix} \|\mathbf{X}_1\|^2 + \lambda & -\|\mathbf{X}_1\|^2 \\ -\|\mathbf{X}_1\|^2 & \|\mathbf{X}_1\|^2 + \lambda \end{bmatrix} \begin{bmatrix} \langle \mathbf{X}_1, \mathbf{y} \rangle \\ \langle \mathbf{X}_1, \mathbf{y} \rangle \end{bmatrix} = \begin{bmatrix} \langle \mathbf{X}_1, \mathbf{y} \rangle / (\lambda + 2\|\mathbf{X}_1\|^2) \\ \langle \mathbf{X}_1, \mathbf{y} \rangle / (\lambda + 2\|\mathbf{X}_1\|^2). \end{bmatrix}$$

The general result for when $\mathbf{X}$ is a $N \times m$ matrix with m identical columns can be solved easily. One can simply check explicitly that

$$(\hat{\beta}^{\text{ridge}})_i = \frac{1}{m\|\mathbf{X}_1\|^2 + \lambda}$$

minimizes

$$L(\beta) := \|\mathbf{y} - \mathbf{X}_1 \left(\sum_{i=1}^m \beta_i \right) \|^2 + \lambda \|\beta\|^2$$

by checking the first derivative and Hessian. This completes the problem. ■

Exercise 3.30. *Consider the elastic-net optimization problem:*

$$\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda[\alpha \|\beta\|_2^2 + (1 - \alpha)\|\beta\|_1]. \tag{3.91}$$

Show how one can turn this into a lasso problem, using an augmented version of $\mathbf{X}$ and $\mathbf{y}$.

Solution. Replace $\mathbf{y}$ with $\tilde{\mathbf{y}}$, by appending p zeros to the end of $\mathbf{y}$. Additionally, replace $\mathbf{X}$ with $\tilde{\mathbf{X}}$, by appending p additional rows of $\sqrt{(\lambda\alpha)}\mathbf{I}$. By Exercise 3.12, we can rewrite the above as

$$\min_\beta \|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\beta\|^2 + \lambda(1-\alpha)\|\beta\|_1.$$

This is by definition a lasso problem on $(\tilde{\mathbf{X}}, \tilde{\mathbf{y}})$ with coefficient $\lambda(1-\alpha)$. This completes the problem. ■

4. Solutions to Chapter 4

Exercise 4.1. *Show how to solve the generalized eigenvalue problem* $\max a^\top \mathbf{B} a$ *subject to* $a^\top \mathbf{W} a = 1$ *by transforming to a standard eigenvalue problem.*

Solution. Define an innerproduct by

$$\langle a, b\rangle_{\mathbf{W}} := (\mathbf{W}^{1/2}a)^\top(\mathbf{W}^{1/2}b).$$

Notice that $\langle a, a\rangle_{\mathbf{W}} = a^\top \mathbf{W} a$, and

$$\langle \mathbf{W}^{-1}\mathbf{B}a, a\rangle_{\mathbf{W}} = a^\top \mathbf{W}^{1/2}\mathbf{W}^{-1/2}\mathbf{B}a = a^\top \mathbf{B}a.$$

Hence, we can rewrite the problem as

$$\max_{\langle a,a\rangle_{\mathbf{W}}=1} \langle \mathbf{W}^{-1}\mathbf{B}a, a\rangle_{\mathbf{W}}.$$

It is easy to check that $\mathbf{W}^{-1}\mathbf{B}$ is indeed self-adjoint w.r.t. the $\langle\cdot,\cdot\rangle_{\mathbf{W}}$ innerproduct. Hence, the above is indeed a standard eigenvalue problem, and the proof is complete. ■

Exercise 4.2. *Suppose we have features* $x \in \mathbb{R}^p$*, a two-class response, with class sizes* N_1*,* N_2*, and the target coded as* $-N/N_1$*,* N/N_2*.*

(a) Show that the LDA rule classifies to class 2 *if*

$$x^\top \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2}(\hat{\mu}_2 + \hat{\mu}_1)^\top \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) - \log(N_2/N_1),$$

and class 1 *otherwise.*

(b) Consider the minimization of the least squares criterion

$$\sum_{i=1}^N (y_i - \beta_0 - x_i^\top \beta)^2. \tag{4.55}$$

Show that the solution $\hat{\beta}$ *satisfies*

$$\left[(N-2)\hat{\mathbf{\Sigma}} + \hat{\mathbf{\Sigma}}_B\right]\beta = N(\hat{\mu}_2 - \hat{\mu}_1) \tag{4.56}$$

(after simplification), where $\mathbf{\Sigma}_B = \frac{N_1 N_2}{N^2}(\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)^\top$*.*

(c) Hence show that $\mathbf{\Sigma}_B\beta$ *is in the direction* $(\hat{\mu}_2 - \hat{\mu}_1)$ *and thus*

$$\hat{\beta} \propto \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1). \tag{4.57}$$

Therefore the least-squares regression coefficient is identical to the LDA coefficient, up to a scalar multiple.

(d) Show that this result holds for any (distinct) coding of the two classes.

(e) Find the solution $\hat{\beta}_0$ (up to the same scalar multiple as in (c), and hence the predicted value $\hat{f}(x) = \hat{\beta}_0 + x^\top \hat{\beta}$). Consider the following rule: classify to class 2 if $\hat{f}(x) > 0$ and class 1 otherwise. Show that this is not the same as the LDA rule unless the classes have equal numbers of observations.

Solution. (a) The LDA rule classifies to class 2 if and only if

$$\log \frac{\Pr(G = 1|X = x)}{\Pr(G = 2| = x)} < 0.$$

The above can be simplified, using estimates for the mean, covariance, and priors, to

$$\log \frac{N_1}{N_2} - \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)^\top \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_1 - \hat{\mu}_2) + x^\top \hat{\mathbf{\Sigma}^{-1}}(\hat{\mu}_1 - \hat{\mu_2}) < 0.$$

Adding $x^\top \hat{\mathbf{\Sigma}^{-1}}(\hat{\mu}_2 - \hat{\mu_1})$ to both sides, and using that $\log(N_1/N_2) = -\log(N_2/N_1)$ yields the result.

(b) We first simplify the inside $(N-2)\hat{\mathbf{\Sigma}} + N\hat{\sigma}_B$. Denote by x_i the $p \times 1$ vector associated with the i-th trial. We have

$$\begin{aligned}
(N-2)\hat{\mathbf{\Sigma}} + N\hat{\sigma}_B &= \sum_{x_i \in G_1} x_i x_i^\top + N_1 \hat{\mu}_1 \hat{\mu}_1^\top - \sum_{i \in G_1} (\hat{\mu}_1 x_i^\top + x_i \hat{\mu}_1^\top) \\
&+ \sum_{x_i \in G_2} x_i x_i^\top + N_2 \hat{\mu}_2 \hat{\mu}_2^\top - \sum_{i \in G_2} (\hat{\mu}_2 x_i^\top + x_i \hat{\mu}_2^\top) \\
&+ \frac{N_1 N_2}{N} \left[\hat{\mu}_1 \hat{\mu}_1^\top + \hat{\mu}_2 \hat{\mu}_2^\top - (\hat{\mu}_1 \hat{\mu}_2^\top + \hat{\mu}_2 \hat{\mu}_1^\top) \right] \\
&= \sum_{i=1}^{N} x_i x_i^\top - N_1 \hat{\mu}_1 \hat{\mu}_1^\top - N_2 \hat{\mu}_2 \hat{\mu}_2^\top \\
&+ \frac{N_1 N_2}{N} \left[\hat{\mu}_1 \hat{\mu}_1^\top + \hat{\mu}_2 \hat{\mu}_2^\top - (\hat{\mu}_1 \hat{\mu}_2^\top + \hat{\mu}_2 \hat{\mu}_1^\top) \right].
\end{aligned}$$

Using that $-N_i + \frac{N_1 N_2}{N} = -N_i^2/N$, we obtain that the above simplifies to

$$\begin{aligned}
&= \sum_{i=1}^{N} x_i x_i^\top - \frac{1}{N}(N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2)(N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2)^\top \\
&= \sum_{i=1}^{N} x_i x_i^\top - \frac{1}{N} \left(\sum_{i=1}^{N} x_i \right) \left(\sum_{i=1}^{N} x_i^\top \right).
\end{aligned}$$

Computing $\mathbf{X}^\top \mathbf{X}$, the $(p+1) \times (p+1)$ matrix, in terms of the $p \times 1$ vectors x_i, and using that $\mathbf{X}^\top \mathbf{X} \beta = \mathbf{X}^\top \mathbf{y}$, it is easy to check that

$$\left[\sum_{i=1}^{N} x_i x_i^\top - \frac{1}{N} \left(\sum_{i=1}^{N} x_i \right) \left(\sum_{i=1}^{N} x_i^\top \right) \right] \hat{\beta} = \sum_{i=1}^{N} y_i x_i - \frac{1}{N} \left(\sum_{i=1}^{N} x_i \right) \left(\sum_{i=1}^{N} y_i \right),$$

where $\hat{\beta}$ denotes the last p components of the least squares solution vector. But $\sum_{i=1}^{N} y_i = N_1 N/N_1 - N_2 N/N_2 = 0$, so the second term vanishes. Hence,

$$\begin{aligned}\left[(N-2)\hat{\mathbf{\Sigma}} + N\hat{\sigma}_B\right]\hat{\beta} &= \sum_{i=1}^{N} y_i x_i = N\sum_{x_i \in G_2} x_i/N_2 + N\sum_{x_i \in G_1} x_i/N_1 \\ &= N(\hat{\mu}_2 - \hat{\mu}_1).\end{aligned}$$

This completes the problem.

(c) Notice that $(\hat{\mu}_2 - \hat{\mu}_1)^\top \hat{\beta}$ is a scalar, so

$$\hat{\mathbf{\Sigma}}_B \hat{\beta} = \frac{N_1 N_2 \langle (\hat{\mu}_2 - \hat{\mu}_1), \hat{\beta}\rangle}{N^2}(\hat{\mu}_2 - \hat{\mu}_1) := C(\hat{\mu}_2 - \hat{\mu}_1).$$

It follows from (4.56) that

$$\hat{\mathbf{\Sigma}}\hat{\beta} = \frac{(N-C)}{N-2}(\hat{\mu}_2 - \hat{\mu}_1),$$

and hence

$$\hat{\beta} \propto \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1),$$

which was to be demonstrated.

(d) It suffices to show that an equation like (4.55) holds, regardless of the encoding. Recall that we had the left-hand-side simplified to

$$\sum_{i=1}^{N} y_i x_i - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)$$

without using any information about the target coding. Denote by c_1 the target coded value for the first class, and c_2 the target coded value for the second class. The above can be rewritten

$$\begin{aligned}\sum_{i=1}^{N} y_i x_i - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right) &= c_1 \sum_{x_i \in G_1} x_i + c_2 \sum_{x_i \in G_2} x_i - \frac{c_1 N_1 + c_2 N_2}{N}\left(\sum_{i=1}^{N} x_i\right) \\ &= c_1 N_1 \hat{\mu}_1 + c_2 N_2 \hat{\mu}_2 - \frac{c_1 N_1 + c_2 N_2}{N}(N_1\hat{\mu}_1 + N_2\hat{\mu}_2).\end{aligned}$$

Simplifying the above yields

$$\frac{N_1 N_2 (c_2 - c_1)}{N}(\hat{\mu}_2 - \hat{\mu}_1).$$

Replacing N in (4.55) with $\frac{N_1 N_2 (c_2 - c_1)}{N}$ holds true in general. Hence, so too does the proportionality result. This completes the problem.

(e) Notice that

$$\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} N & \sum_{i=1}^{N} x_i^\top \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i x_i^\top \end{bmatrix}$$

since the first column of $\mathbf{X}$ is all ones. Since

$$\mathbf{X}^\top \mathbf{X} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta} \end{bmatrix} = \mathbf{X}^\top \mathbf{y},$$

as this is the least-squares solution, we have from the first row of the above equation that

$$N\hat{\beta}_0 + \sum_{i=1}^{N} x_i^\top \hat{\beta} = \sum_{i=1}^{N} y_i.$$

Notice that the right-hand-side of the above is 0. Hence,

$$\hat{\beta}_0 = \frac{-1}{N} \sum_{i=1}^{N} x_i^\top \hat{\beta}.$$

Hence, we have that

$$\hat{f}(x) = \left(\frac{-1}{N} \sum_{i=1}^{N} x_i^\top + x^\top \right) \hat{\beta}.$$

To investigate the difference of this expression from the LDA rule, we will rewrite the above as follows:

$$\left(\frac{-1}{N} \sum_{i=1}^{N} x_i^\top + x^\top \right) \hat{\beta} \;=\; \frac{-C}{N}(N_1\hat{\mu}_1 + N_2\hat{\mu}_2)^\top \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) + Cx^\top \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1),$$

where $C > 0$ is the proportionality constant from earlier parts. Since the constant above does not affect the sign, classifying to class 2 iff $\hat{f}(x) > 0$, or iff $\hat{f}(x)/C > 0$ is the same rule. Hence, this is equivalent to classifying to class 2 if and only if

$$x^\top \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{N}(N_1\hat{\mu}_1 + N_2\hat{\mu}_2)^\top \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1).$$

From part (a), notice that if $N_1 = N_2 = N/2$, then the log term in (4.55) vanishes, and $N/2$ can be factored out of the right hand term in the equation above. In this case, one can see that the classification rules are the same. Otherwise, there is an additional log term in LDA, and the term $\frac{1}{N}(N_1\hat{\mu}_1 + N_2\hat{\mu}_2)^\top \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)$ does not simplify to $\frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)^\top \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)$. Hence, when $N_1 \neq N_2$, the classification rules are not necessarily the same.

■

Exercise 4.3. *Suppose we transform the original predictors* $\mathbf{X}$ *to* $\hat{\mathbf{Y}}$ *via linear regression. In detail, let* $\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} = \mathbf{X}\hat{\mathbf{B}}$*, where* $\mathbf{Y}$ *is the indicator response matrix. Similarly for any input* $x \in \mathbb{R}^p$*, we can a transformed vector* $\hat{y} = \hat{\mathbf{B}}^\top x \in \mathbb{R}^K$*. Show that LDA using* $\hat{\mathbf{Y}}$ *is equivalent to LDA in the original space.*

Solution. Notice that under the augmented data, the following tranformations occur:

$$\begin{aligned} x^\top &\mapsto x^\top\hat{\mathbf{B}} \\ \hat{\mu}_k &\mapsto \hat{\mathbf{B}}^\top\hat{\mu}_k \\ \hat{\mathbf{\Sigma}} &\mapsto \hat{\mathbf{B}}^\top\hat{\mathbf{\Sigma}}\hat{\mathbf{B}} \end{aligned}$$

We remark that for the transformed version of $\mathbf{\Sigma}$ to be invertible, we require that $K \leq p$, since $\hat{\mathbf{B}}$ is $p \times K$. We also note that $\hat{\mathbf{B}}$ has rank K. Under these transformations, the formula

for δ_k becomes

$$x^\top \hat{\mathbf{B}}(\hat{\mathbf{B}}^\top \hat{\mathbf{\Sigma}} \hat{\mathbf{B}})^{-1} \hat{\mathbf{B}}^\top \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^\top \hat{\mathbf{B}}(\hat{\mathbf{B}}^\top \hat{\mathbf{\Sigma}} \hat{\mathbf{B}})^{-1} \hat{\mathbf{B}}^\top \hat{\mu}_k + \log \pi_k.$$

Simplifying into inner-product form yields

$$\langle \hat{\mathbf{B}}(\hat{\mathbf{B}}^\top \hat{\mathbf{\Sigma}} \hat{\mathbf{B}})^{-1} \hat{\mathbf{B}}^\top \hat{\mu}_k, x \rangle - 1/2 \langle \hat{\mathbf{B}}(\hat{\mathbf{B}}^\top \hat{\mathbf{\Sigma}} \hat{\mathbf{B}})^{-1} \hat{\mathbf{B}}^\top \hat{\mu}_k, \hat{\mu}_k \rangle + \log \pi_k$$

A very lengthy exercise in matrix algebra shows that this is equal to the standard discriminant for LDA. Some key steps in the derivation include writing $\hat{\mathbf{\Sigma}}$ as a product of matrices:

$$\hat{\mathbf{\Sigma}} = \frac{1}{N-K} \mathbf{X}^\top \left(\mathbf{I} - \mathbf{Y} \mathbf{N}_K^{-1} \mathbf{Y}^\top \right) \mathbf{X},$$

where $\mathbf{N}_K = \operatorname{diag}(N_1, \ldots, N_K)$. Deriving the above uses the fact that

$$\hat{\mu}_k = \mathbf{X}^\top y_k,$$

where y_k is the kth column of $\mathbf{Y}$, as well as the fact that $\mathbf{Y}^\top \mathbf{Y} = \mathbf{N}_K$. At this point, $(\hat{\mathbf{B}}^\top \hat{\mathbf{\Sigma}} \hat{\mathbf{B}})^{-1}$ can be computed in a slightly more explicit way. In particular, using the definition of $\hat{\mathbf{B}}$, some $\mathbf{X}$ cancellation occurs, and we see that

$$(\hat{\mathbf{B}}^\top \hat{\mathbf{\Sigma}} \hat{\mathbf{B}})^{-1} = (N-K) \left(\mathbf{I} - \mathbf{N}_K^{-1} \mathbf{A} \right)^{-1} \mathbf{A}^{-1},$$

where $\mathbf{A}$ is the self-adjoint matrix $\mathbf{A} = \hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{Y}$. Another simple but lengthy computation then reveals that

$$\hat{\mathbf{B}}(\hat{\mathbf{B}}^\top \hat{\mathbf{\Sigma}} \hat{\mathbf{B}})^{-1} \hat{\mathbf{B}} \mathbf{X}^\top \mathbf{Y} = \hat{\mathbf{\Sigma}}^{-1} \mathbf{X}^\top \mathbf{Y}.$$

The i-th column of the above matrix equality shows that each of the first two terms above are identical to the standard LDA terms, since $\hat{\mu}_k = \mathbf{X}^\top y_k$. Since the priors π_k are unaffected by the transformations involving $\hat{\mathbf{B}}$, this completes the proof. ■

Exercise 4.4. *Consider the multilogit model with K classes (4.17). Let β be the $(p+1)(K-1)$ vector consisting of all the coefficients. Define a suitably enlarged version of the input vector x to accommodate this vectorized coefficient matrix. Derive the Newton-Ralphson algorithm for maximizing the multinomial log-likelihood, and describe how you would implement this algorithm.*

Solution. The $(p+1)(K-1)$ coefficient vector β is given by

$$\beta = \begin{pmatrix} \beta_{10} \\ \beta_1 \\ \beta_{20} \\ \beta_2 \\ \vdots \\ \beta_{(K-1)0} \\ \beta_{K-1} \end{pmatrix}$$

where each β_i is itself a vector of length p. Of course, enlarging x by adding a constant to account for the intercept, then stacking $(K-1)$ of these ontop of each other allows for this enlarged x to be multiplied by β. Defining Y_k to be a $(p+1)(K-1) \times (p+1)(K-1)$ diagonal

matrix with 1's along the diagonal entries $k(p+1)-p, k(p+1)-p+1, \ldots, k(p+1)$ and zeros elsewhere. With these definitions, defining X to be the enlarged version of x, we have

$$\beta^\top Y_k X = (\beta_{k0}\ \beta_k^\top) \begin{pmatrix} 1 \\ x \end{pmatrix}.$$

We can use this to write the log likelihood in a convenient way, and then derive the Newton-Ralphson algorithm in this setting. In particular, recall that the log likelihood for K classes is given by

$$\ell(\beta) = \sum_{i=1}^N \left[\sum_{j=1}^{K-1} \chi_{ij} \log p_j(x_i;\beta) + \chi_{iK} \log(p_K(x_i;\beta)) \right],$$

where χ_{ij} here is 1 if y_i is class j and zero otherwise. Recall that

$$\begin{aligned} p_j(x_i;\beta) &= \frac{\exp(\beta_j^\top x_i)}{1+\sum_{\ell=1}^{K-1}\exp(\beta_\ell^\top x_i)}, \qquad j=1,\ldots,K-1 \\ p_K(x_i;\beta) &= \frac{1}{1+\sum_{\ell=1}^{K-1}\exp(\beta_\ell^\top x_i)}. \end{aligned}$$

Here, a x_i is understood to have a 1 appended to the beginning, and β_j is a vector of coefficients corresponding to class j. Using this, $\ell(\beta)$ can be simplified to

$$\begin{aligned} \ell(\beta) &= \sum_{i=1}^N \left[\sum_{j=1}^{K-1} \left[\chi_{ij}\beta_j^\top x_i - \chi_{ij} \log\left(1+\sum_{\ell=1}^{K-1}\exp(\beta_\ell^\top x_i)\right)\right] - \chi_{iK}\log(1+\sum_{\ell=1}^{K-1}\exp(\beta_\ell^\top x_i))\right] \\ &= \sum_{i=1}^N \left[\sum_{j=1}^{K-1} \chi_{ij}\beta_j^\top x_i - \log(1+\sum_{\ell=1}^{K-1}\exp(\beta_\ell^\top x_i))\right] \end{aligned}$$

since $\sum_j \chi_{ij} = 1$. Let $Y_j^{(i)}$ be equal to Y_j if the class of y_i is j, and 0 otherwise. We can then rewrite the above in terms of $Y_j, Y_j^{(i)}, X_i$, and β:

$$\ell(\beta) = \sum_{i=1}^N \left[\sum_{j=1}^{K-1} \beta^\top Y_j^{(i)} X_i - \log\left(1+\sum_{\ell=1}^{K-1}\exp\left(\beta^\top Y_\ell X_i\right)\right)\right]$$

From this we can see that

$$\begin{aligned} \frac{\partial \ell}{\partial \beta} &= \sum_{i=1}^N \left[\sum_{j=1}^{K-1} Y_j^{(i)} X_i - \frac{\sum_{\ell=1}^{K-1}\exp\left(\beta^\top Y_\ell X_i\right) Y_\ell X_i}{\left(1+\sum_{\ell=1}^{K-1}\exp\left(\beta^\top Y_\ell X_i\right)\right)}\right] \\ &= \sum_{i=1}^N \sum_{j=1}^{K-1} \left[Y_j^{(i)} X_i - \frac{\exp\left(\beta^\top Y_j X_i\right) Y_j X_i}{\left(1+\sum_{\ell=1}^{K-1}\exp\left(\beta^\top Y_\ell X_i\right)\right)}\right] \end{aligned}$$

The Hessian is then given by

$$\frac{\partial^2 \ell}{\partial\beta\partial\beta^\top} = \sum_{i=1}^N \sum_{j=1}^{K-1} \left[\frac{\exp\left(\beta^\top Y_j X_i\right)\sum_{\ell=1}^{K-1}\exp\left(\beta^\top Y_\ell X_i\right) X_i^\top Y_\ell^\top Y_j X_i}{\left(1+\sum_{\ell=1}^{K-1}\exp\left(\beta^\top Y_\ell X_i\right)\right)^2} - \frac{\exp\left(\beta^\top Y_j X_i\right) X_i^\top Y_j^\top Y_j X_i}{\left(1+\sum_{\ell=1}^{K-1}\exp\left(\beta^\top Y_\ell X_i\right)\right)}\right].$$

Since Y_j and Y_ℓ are diagonal with 1's along the diagonal in distinct positions when $\ell \neq j$ and zeros otherwise, its easy to see that $Y_\ell Y_j = \delta_{\ell j} Y_j$. Hence,

$$\frac{\partial^2 \ell}{\partial \beta \partial \beta^\top} = \sum_{i=1}^{N} \sum_{j=1}^{K-1} \left[\frac{\exp\left(\beta^\top Y_j X_i\right) \exp\left(\beta^\top Y_j X_i\right) X_i^\top Y_j^\top Y_j X_i}{\left(1 + \sum_{\ell=1}^{K-1} \exp\left(\beta^\top Y_\ell X_i\right)\right)^2} - \frac{\exp\left(\beta^\top Y_j X_i\right) X_i^\top Y_j^\top Y_j X_i}{\left(1 + \sum_{\ell=1}^{K-1} \exp\left(\beta^\top Y_\ell X_i\right)\right)} \right].$$

From the above formulas, we can compute the Hessian given data X and a fixed β, and we can compute the gradient $\frac{\partial \ell}{\partial \beta}$ given data X, labels y, and fixed β. The Newton-Ralphson algorithm is given by the update rule

$$\beta^{(k+1)} = \beta^{(k)} - \left(\frac{\partial^2 \ell}{\partial \beta \partial \beta^\top} \right)^{-1} \frac{\partial \ell}{\partial \beta}.$$

Typically, one can choose $\beta^{(0)} = \vec{0}$. Here is some psuedo-code for implementing the algorithm.

```
initialize beta
initialize error
initialize tol
while error > tol:
   beta_old = beta
   H = return_Hessian(beta, X)
   g = return_gradient(beta, X, y)
   beta_new = beta_old - H^(-1)g
   error = norm(beta_new - beta_old)
   beta = beta_new
```

■

Exercise 4.5. *Consider a two class logistic regression problem with $x \in \mathbb{R}$. Characterize the maximum likelihood estimates of the slope and intercept parameter if the sample x_i for the two classes are separated by a point $x_0 \in \mathbb{R}$. Generalize this result to (a) $x \in \mathbb{R}^p$ and (b) more than two classes.*

Solution. Let class 1, encoded by $y_i = 1$, denote the class corresponding to $x_i > x_0$. The log likelihood for two classes with $x \in \mathbb{R}$ can be written

$$\begin{aligned} \ell(\beta) &= \sum_i \left[y_i(\beta_0 + \beta x_i) - \log(1 + \exp\left(\beta_0 + \beta x\right)) \right] \\ &= \sum_{x_i > x_0} \left[\beta_0 + \beta x_i - \log(1 + \exp\left(\beta_0 + \beta x_i\right)] + \sum_{x_i < x_0} \left[-\log(1 + \exp\left(\beta_0 + \beta x_i\right)) \right] \end{aligned}$$

The maximum is bounded below by a particular choice of $\beta_0 = -\beta x_0$. Hence,

$$\max_\beta \ell(\beta) \geq \sum_{x_i > x_0} \left[\beta(x_i - x_0) - \log(1 + \exp\left(\beta(x_i - x_0)\right)) \right] + \sum_{x_i < x_0} \left[-\log(1 + \exp\left(\beta(x_i - x_0)\right)) \right].$$

Notice that in the second term, the argument on the exponential is negative. Hence, there is sufficiently large β such that each term in the second summand is bounded below by

$-\log(2)$. Hence,

$$\max_{\beta} \ell(\beta) \;\geq\; \sum_{x_i > x_0} \left[\beta(x_i - x_0) - \log(1 + \exp\left(\beta(x_i - x_0)\right)\right] - N_2 \log(2).$$

where N_2 is the number of elements in class 2. Its clear that each term in the sum of the above function diverges to $+\infty$ as $\beta \to +\infty$, since

$$\frac{\partial}{\partial \beta}\left[C\beta - \log(1 + \exp\left(C\beta\right)\right] = C\left(1 - \frac{\exp\left(C\beta\right)}{1 + \exp\left(C\beta\right)}\right) > 0$$

for any $\beta > 0$, provided that $C > 0$ is some fixed constant independent of β. Hence, the maximum likelihood estimates in this case do not exist, as $\ell(\beta)$ has no global or local maximum.

(a) Now suppose $x_i \in \mathbb{R}^p$ (without a 1 appended), and suppose there is $x_0 \in \mathbb{R}^p$, and a unit vector $\mathbf{v} \in \mathbb{R}^p$ such that all elements x_i in class 1 satisfy $\mathbf{v}^\top x_i > \mathbf{v}^\top x_0$, and all elements x_i in class 2 satisfy $\mathbf{v}^\top x_i < \mathbf{v}^\top x_0$ (I believe this is the intended way to generalize the assumption). The log likelihood is given by

$$\begin{aligned}
\ell(\beta) &= \sum_i \left[y_i(\beta_0 + \beta^\top x_i) - \log(1 + \exp\left(\beta_0 + \beta^\top x_i\right))\right] \\
&= \sum_{\{x_i : \beta^\top(x_i - x_0) > 0\}} \left[\beta_0 + \beta^\top x_i - \log(1 + \exp\left(\beta_0 + \beta^\top x_i\right)\right] \\
&+ \sum_{\{x_i : \beta^\top(x_i - x_0) < 0\}} \left[-\log(1 + \exp\left(\beta_0 + \beta^\top x_i\right))\right]
\end{aligned}$$

Again, choosing $\beta_0 = -\beta^\top x_0$, the exact same argument as before shows the above diverges as $\|\beta\| \to \infty$. Of course, the slight difference is that here the limit is taken as $a \to \infty$, where $\beta = a\mathbf{v}$.

(b) For $K \geq 3$, suppose there is $\mathbf{v}$ and $x_0 \in \mathbb{R}^p$ such that $\mathbf{v}^\top x_i > \mathbf{v}^\top x_0$ for any x_i with label in class $1, 2, \ldots, K-1$, but $\mathbf{v}^\top x_i < \mathbf{v}^\top x_0$ for any x_i with label in class K. Then the log likelihood is

$$\ell(\beta) \;=\; \sum_i \left[\sum_{j=1}^{K-1} \chi_{ij}(\beta_{0j} + \beta_j^\top x_i) - \log(1 + \exp\left(\beta_{0K} + \beta_K^\top x_i\right))\right]$$

where β_{0j} is the intercept for the j-th class, β_j is the p vector of coefficients for the j-th class, and χ_{ij} is 1 if y_i is in class j and zero otherwise. Certainly $\max_\beta \ell(\beta)$ is bounded above by a particular choice of β for which all intercepts are equal: $\beta_0 := \beta_{01} = \beta_{02} = \cdots = \beta_{0K}$, and all feature coefficients are equal: $\beta := \beta_1 = \beta_2 = \cdots = \beta_K$. Hence,

$$\max_{\beta} \ell(\beta) \;\geq\; \sum_i \left[\sum_{j=1}^{K-1} \chi_{ij}(\beta_0 + \beta^\top x_i) - \log(1 + \exp\left(\beta_0 + \beta^\top x_i\right))\right]$$

If y_i is in class $1, 2, \ldots, K-1$, then $\sum_j \chi_{ij} = 1$. If y_i is in class j, then $\sum_j \chi_{ij} = 0$. Hence, the above becomes

$$\begin{aligned}\max_\beta \ell(\beta) \quad &\geq \sum_{\{x_i : y_i \text{in class } 1,\ldots,K-1\}} \left[(\beta_0 + \beta^\top x_i) - \log(1 + \exp\left(\beta_0 + \beta^\top x_i\right))\right] \\ &+ \sum_{\{x_i : y_i \text{ in class } K\}} \left[-\log(1 + \exp\left(\beta_0 + \beta^\top x_i\right))\right]\end{aligned}$$

Choosing $\beta_0 = -\beta^\top x_0$,

$$\begin{aligned}\max_\beta \ell(\beta) \quad &\geq \sum_{\{x_i : y_i \text{in class } 1,\ldots,K-1\}} \left[\beta^\top (x_i - x_0) - \log(1 + \exp\left(\beta^\top (x_i - x_0)\right))\right] \\ &+ \sum_{\{x_i : y_i \text{ in class } K\}} \left[-\log(1 + \exp\left(\beta^\top (x_i - x_0)\right))\right]\end{aligned}$$

It then becomes the same argument as before, choosing $\beta = a\mathbf{v}$ and taking $a \to \infty$. The second term is bounded above by $-N_K \log(2)$, and each term in the first sum is strictly increasing as β increases. This completes the problem.

■

Exercise 4.6. *Suppose we have N points X_i in $\mathbb{R}^p$ in general position, with class labels $y_i \in \{-1, 1\}$. Prove that the perceptron learning algorithm converges to a separating hyperplane in a finite number of steps:*

- *(a) Denote a hyperplane by $f(x) = \beta_1^\top x + \beta_0 = 0$, or in more compact notation $\beta^\top x^* = 0$ where $x^* = (x, 1)$ and $\beta = (\beta_1, \beta_0)$. Let $z_i = x_i^*/\|x_i^*\|$. Show that the separability implies the existence of a β_{sep} such that $y_i \beta_{sep}^\top z_i \geq 1$ for all i.*
- *(b) Given a current β_{old}, the perceptron algorithm identifies a point z_i that is misclassified, and produces the update $\beta_{\text{new}} \leftarrow \beta_{\text{old}} + y_i z_i$. Show that $\|\beta_{\text{new}} - \beta_{\text{sep}}\|^2 \leq \|\beta_{\text{old}} - \beta_{\text{sep}}\|^2 - 1$, and hence that the algorithm converges to a separating hyperplane in no more than $\|\beta_{\text{start}} - \beta_{\text{sep}}\|^2$ steps.*

Solution. (a) Suppose that the classes are separable. By definition, there exists a vector β such that for each x_i with label 1 we have $\beta^\top x_i^* > c_i \|x_i^*\| > 0$, and for each x_i with label -1 we have $\beta^\top x_i^* < \|x_i^*\| d_i < 0$. Let D denote the maximum of all d_i's, (i.e., negative number closest to 0) and C denote the minimum of all $c_i's$. Note that

$$\begin{aligned}\beta^\top x_i^* \geq C\|x_i^*\| > 0 \qquad \forall i : y_i = 1 \\ \beta^\top x_i^* \leq D\|x_i^*\| < 0 \qquad \forall i : y_i = -1.\end{aligned}$$

Let $A = \min\{C, |D|\}$. Then clearly

$$\begin{aligned}\beta^\top x_i^* \geq A\|x_i^*\| > 0 \qquad \forall i : y_i = 1 \\ \beta^\top x_i^* \leq -A\|x_i^*\| < 0 \qquad \forall i : y_i = -1.\end{aligned}$$

Let $\beta_{\text{sep}} = \frac{1}{A}\beta$. Then we have

$$\begin{aligned}\beta_{\text{sep}}^\top x_i^* \geq \|x_i^*\| > 0 \qquad \forall i : y_i = 1 \\ \beta_{\text{sep}}^\top x_i^* \leq -\|x_i^*\| < 0 \qquad \forall i : y_i = -1.\end{aligned}$$

This is of course equivalent to

$$\begin{aligned} y_i\beta_{\text{sep}}^\top z_i \geq 1 \qquad & \forall i : y_i = 1 \\ y_i\beta_{\text{sep}}^\top z_i \geq 1 \qquad & \forall i : y_i = -1, \end{aligned}$$

which was to be demonstrated.

(b) Notice that

$$\begin{aligned} \|\beta_{\text{new}} - \beta_{\text{sep}}\|^2 &= \|\beta_{\text{old}} + y_i z_i - \beta_{\text{sep}}\|^2 \\ &= \langle \beta_{\text{old}} - \beta_{\text{sep}}, \beta_{\text{old}} - \beta_{\text{sep}} \rangle + 2\langle \beta_{\text{old}} - \beta_{\text{sep}}, y_i z_i \rangle + \langle y_i z_i, y_i z_i \rangle. \\ &= \|\beta_{\text{old}} - \beta_{\text{sep}}\|^2 + 2\langle \beta_{\text{old}}, y_i z_i \rangle - 2\langle \beta_{\text{sep}}, y_i z_i \rangle + \frac{1}{\|x_i^*\|^2}\|x_i^*\|^2. \end{aligned}$$

By the previous part, $-2\langle \beta_{\text{sep}}, y_i z_i \rangle \leq -2$. Since β_{old} by definition misclassified z_i, $2y_i\langle \beta_{\text{old}}, z_i \rangle$ will always be negative, since y_i and $\langle \beta_{\text{old}}, z_i \rangle$ have opposite signs. Putting this together,

$$\|\beta_{\text{new}} - \beta_{\text{sep}}\|^2 \leq \|\beta_{\text{old}} - \beta_{\text{sep}}\|^2 - 0 - 2 + 1,$$

which was to be demonstrated. Hence, after $\|\beta_{\text{start}} - \beta_{\text{sep}}\|^2$, we have

$$\begin{aligned} \|\beta_{\text{newest}} - \beta_{\text{sep}}\|^2 &\leq \|\beta_{\text{newest}-1} - \beta_{\text{sep}}\|^2 - 1 \\ &\leq \|\beta_{\text{newest}-2} - \beta_{\text{sep}}\|^2 - 2 \\ &\leq \cdots \leq \\ &\leq \|\beta_{\text{start}} - \beta_{\text{sep}}\|^2 - \|\beta_{\text{start}} - \beta_{\text{sep}}\|^2. \end{aligned} \tag{4.1}$$

Of course, this assumes that at each step the guess is further 1 distance squared away from β_{sep}. Once $\|\beta_{\text{newest}} - \beta_{\text{sep}}\|^2 < 1$ (which, by the above, occurs in less than $\|\beta_{\text{start}} - \beta_{\text{sep}}\|^2$ steps), no points are misclassified. Indeed, if there were misclassified points, then one could apply the above to find $\beta_{\text{newest}+1}$ such that $\|\beta_{\text{newest}+1} - \beta_{\text{sep}}\|^2 \leq \|\beta_{\text{newest}} - \beta_{\text{sep}}\|^2 - 1 < 0$, a clear contradiction. This completes the problem.

■

Exercise 4.7. *Consider the criterion*

$$D^*(\beta, \beta_0) = -\sum_{i=1}^{N} y_i(x_i^\top \beta + \beta_0),$$

a generalization of (4.41) where we sum over all the observations. Consider minimizing D^* subject to $\|\beta\| = 1$. Describe the criterion in words. Does it solve the optimal separating hyperplane problem?

Solution. In words, $y_i(x_i^\top \beta + \beta_0)$ is related to the distance of x_i to the hyperplane. If x_i is correctly classified, then for either class $y_i(x_i^\top \beta + \beta_0)$ is positive. Hence, the above subtracts the distance to the hyperplane of correctly classified points, and adds the distance of misclassified points. The lagrangian associated to this problem is

$$\mathcal{L}(\beta, \lambda) = -\sum_{i=1}^{N} y_i(x_i^\top \beta + \beta_0) + \frac{\lambda}{2}(\|\beta\|^2 - 1)$$

Writing the sufficient condition $\frac{\partial \mathcal{L}}{\partial \beta} = 0$ and assuming that the covariates are linearly independent reveals that $\lambda \neq 0$ and

$$\beta = \frac{1}{\lambda} \sum_{i=1}^{N} y_i x_i.$$

According to the condition (4.53) for the solution to the optimal hyperplane problem, we see that if $x_j^\top \beta + \beta_0 \neq 1$ for some j, then x_j does not contribute to β. However, the solution to the problem of minimizing D^* is a linear combination of all $y_i x_i$, each with coefficient $\frac{1}{\lambda}$. This shows that minimizing D^* does not solve the optimal separating hyperplane problem. ■

Exercise 4.8. *Consider the multivariate Gaussian model* $X|G = k \sim \mathcal{N}(\mu_k, \mathbf{\Sigma})$*, with the additional restriction that* $\operatorname{rank}\{\mu_k\}_{k=1}^K = L < \max\{K-1, p\}$*. Derive the constrained MLEs for the* μ_k *and* $\mathbf{\Sigma}$*. Show that the Bayes classification rule is equivalent to classifying in the reduced subspce computed by LDA.*

Solution. This seems to be quite an involved problem, and is best solved by reading Appendix I of the paper [1], and working out all details. I hope to revisit this problem soon. ■

Exercise 4.9. *Write a computer program to perform a quadratic discriminant analysis by fitting a separate Gaussian model per class. Try it out on the vowel data, and compute the misclassification error for the test data. The data can be found in the book website www-stat.stanford.edu/ElemStatLearn.*

Solution. The following python code generates predictions using QDA and plots the confusion matrices for each class:

```
import pandas as pd
data_train = pd.read_csv('data/train.csv')
data_test = pd.read_csv('data/test.csv')

X_train = data_train.iloc[:,2:].copy()
y_train = data_train.iloc[:,1].copy()
X_test = data_test.iloc[:,2:].copy()
y_test = data_test.iloc[:,1].copy()

from sklearn.discriminant_analysis import QuadraticDiscriminantAnalysis
from sklearn.metrics import multilabel_confusion_matrix
qda = QuadraticDiscriminantAnalysis()
qda.fit(X_train, y_train)

y_preds_proba = qda.predict_proba(X_test)
y_preds = qda.predict(X_test)
conf = multilabel_confusion_matrix(y_test,y_preds)

import seaborn as sns
import matplotlib.pyplot as plt
for i in range(0,11):
    fig, ax = plt.subplots()
```

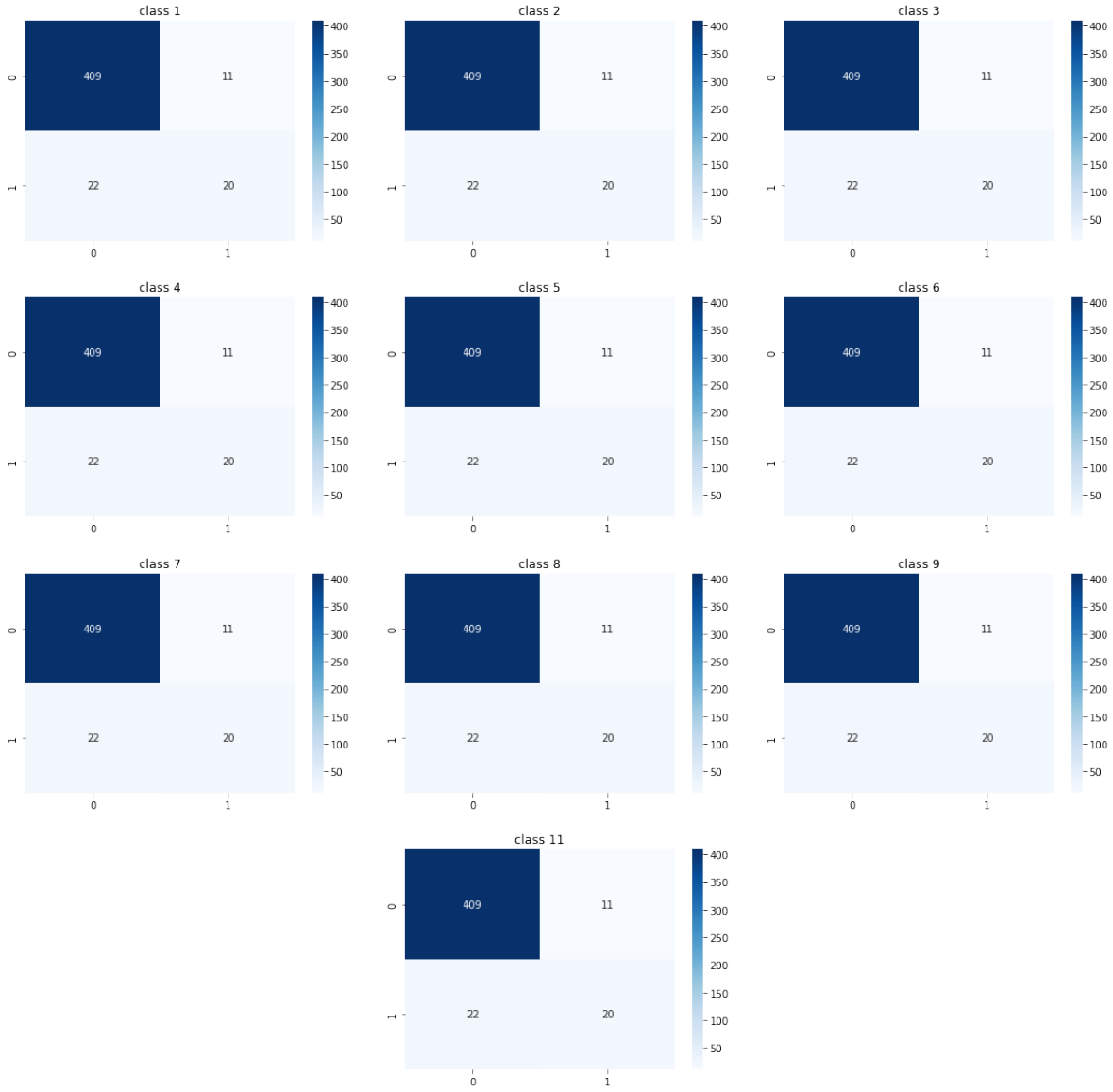

FIGURE 2. Displays confusion matrices for each class using QDA on the vowel data, as specified in Exercise 4.9.

```
ax = sns.heatmap(conf[10,:,:], annot = True,cmap='Blues', fmt='g')
ax.set_title(f"class {i+1}")
plt.show()
```

Figure 2 displays the confusion matrices for each class. We compute mean accuracy over each flass via the following code:

```
import numpy as np
accuracy = []
for i in range(0,11):
    accuracy.append((conf[i,:,:][0,0] + conf[i,:,:][1,1])/conf[i,:,:].sum())
```

```
np.asarray(accuracy).mean()
```

This reveals an average accuracy of about 93.4%. ■

5. Solutions to Chapter 5

6. Solutions to Chapter 6

7. Solutions to Chapter 7

References

[1] Trevor Hastie and Robert Tibshirani. Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):155–176, 1996.

[2] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Department of Mathematics, Pennsylvania State University, 107 McAllister Bld., University Park, State College, PA 16802, U.S.A.

Email address: jwp5828@psu.edu