Solutions to the exercises of ESL, Second Edition

Lorenzo Battarra

January 19, 2020

Contents

Notation

- Data vectors/matrices are denoted with boldface letters, e.g. $\mathbf{X}$, $\boldsymbol{\epsilon}$, while single variable values or abstract random variables are denoted with standard letters, e.g. X, ϵ.
- In case of vector or tensor-valued random variables, the sample dimension is always assumed to come first. For example, the matrix $\mathbf{X}$ containing N samples of a vector variable of dimension p has shape (N, p).
- e_p is the vector of ones of size p.

Chapter 2: Overview of supervised learning

Ex. 2.1

Since $||t_k||^2 = 1$, one has:

$$||t_k - \hat{y}||^2 = 1 - 2\hat{y}_k + ||\hat{y}||^2 \tag{1}$$

hence:

$$\text{argmin}_k ||t_k - \hat{y}|| = \text{argmin}_k ||t_k - \hat{y}||^2 = \text{argmax}_k \, \hat{y}_k \tag{2}$$

Ex. 2.2

This assignment is slightly ambiguous, since the we are told that the 100 examples are generated for each class but the *a priori* probabilities $P(Y = \pm 1)$ are not specified. We interpret the exercise as $P(Y = \pm 1) = 1/2$, and obtain the probability distribution:

$$\begin{aligned} P(Y = \pm 1) &= \frac{1}{2} && (3) \\ P(x|Y = \pm 1) &= \frac{1}{10} \sum_{k=1}^{10} \mathcal{N}(x; m_k^{\pm}, \mathbb{I}/5) && (4) \end{aligned}$$

The decision boundary is the set points x for which:

$$P(Y = +1|X = x) = P(Y = -1|X = x) \tag{5}$$

From Bayes theorem and the fact that $P(Y = +1) = P(Y = -1)$, this is equivalent to:

$$p(x|Y = +1) = p(x|Y = -1) \tag{6}$$

which upon simplification reads:

$$\sum_{k=1}^{10} \exp\left(5\, m_k^{+T} x - \frac{5}{2} m_k^{+T} m_k^{+}\right) = \sum_{k=1}^{10} \exp\left(5\, m_k^{-T} x - \frac{5}{2} m_k^{-T} m_k^{-}\right) \tag{7}$$

Note how with one gaussian per class instead of 10 the decision boundary becomes linear (LDA).

Ex. 2.3

The variables $\rho_i \equiv ||x_i||$ are i.i.d. with c.d.f.:

$$P(\rho_i \leq \bar{\rho}) = \bar{\rho}^p \tag{8}$$

So letting $\rho_m \equiv \min(\{\rho_i\}_{i=1,\ldots,N})$, one has:

$$P\left(\rho_m \geq \bar{\rho}\right) = \prod_{i=1}^{N} P\left(\rho_i \geq \bar{\rho}\right) = \left(1 - \rho^p\right)^N \tag{9}$$

The median of ρ_m is the value $\rho^\star$ s.t. $P\left(\rho_m \geq \rho^\star\right) = 1/2$. Using (9) it is easy to see that:

$$\rho^\star = \left(1 - \frac{1}{2^{1/N}}\right)^{1/p} \tag{10}$$

Ex. 2.4

The components $(X_j)_{j=1,\ldots,p}$ of X are i.i.d. centered and unit variance gaussians. Hence, the linear combination $Z \equiv a^T X$ is still gaussian and centered, and its variance is also one:

$$\text{Var}\left(\sum_{j=1}^{p} a_j X_j\right) = \sum_{j=1}^{p} a_j^2 = 1 \tag{11}$$

Ex. 2.5

The random variables y_0 and $\hat{y}_0$ are independent, hence:

$$\begin{aligned}
\mathrm{EPE}(x_0) &\equiv \mathbb{E}_{\mathcal{T},y_0|x_0}\left[(y_0-\hat{y}_0)^2\right] = \left(\mathbb{E}_{\mathcal{T},y_0|x_0}\left[y_0-\hat{y}_0\right]\right)^2 + \mathrm{Var}_{\mathcal{T},y_0|x_0}\left(y_0-\hat{y}_0\right)\\
&= \left(\mathbb{E}_{y_0|x_0}\left[y_0\right]-\mathbb{E}_{\mathcal{T}}\left[\hat{y}_0\right]\right)^2 + \mathrm{Var}_{y_0|x_0}(y_0) + \mathrm{Var}_{\mathcal{T}}(\hat{y}_0)
\end{aligned}$$

We obtain the decomposition of expected prediction error as a sum of irreducible variance, squared bias and estimation variance:

$$\begin{aligned}
\mathrm{EPE}(x_0) &= \mathrm{Var}_{y_0|x_0}(y_0) + \mathrm{Bias}^2_{\mathcal{T},y_0|x_0}(y_0) + \mathrm{Var}_{\mathcal{T}}(\hat{y}_0) && (12)\\
\mathrm{Var}_{y_0|x_0}(y_0) &\equiv \mathbb{E}_{y_0|x_0}\left[\left(y_0-\mathbb{E}_{y_0|x_0}\left[y_0\right]\right)^2\right] && (13)\\
\mathrm{Bias}_{\mathcal{T},y_0|x_0}(y_0) &\equiv \mathbb{E}_{y_0|x_0}\left[y_0\right]-\mathbb{E}_{\mathcal{T}}\left[\hat{y}_0\right] && (14)\\
\mathrm{Var}_{\mathcal{T}}(\hat{y}_0) &\equiv \mathbb{E}_{\mathcal{T}}\left[\left(\mathbb{E}_{\mathcal{T}}\left[\hat{y}_0\right]-\hat{y}_0\right)^2\right] && (15)
\end{aligned}$$

Equation (2.27) was obtained under two assumptions:

- the underlying distribution for Y conditional on X is:

  $$y = \beta^T X + \epsilon, \quad \epsilon \sim \mathcal{N}(0,\sigma^2)$$

  from which it follows that:

  $$\begin{aligned}
  \mathbb{E}_{y_0|x_0}\left[y_0\right] &= \beta^T x_0\\
  \mathrm{Var}_{y_0|x_0}(y_0) &= \sigma^2
  \end{aligned}$$

- $\hat{y}_0$ is an OLS estimate of Y at $X = x_0$:

  $$\hat{y}_0 = x_0^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{X}\beta+\boldsymbol{\epsilon}\right) = x_0^T\beta + x_0^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\epsilon}$$

By assumption ϵ and X are independent random variables. Hence, denoting $p_X \equiv \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}x_0$ such that $\hat{y}_0 = x_0^T\beta + p_X^T\boldsymbol{\epsilon}$, one has:

$$\begin{aligned}
\mathbb{E}_{\mathcal{T}}\left[p_X^T\,\epsilon\right] &= \mathbb{E}_{\mathbf{X}}\left[p_X^T\right]\mathbb{E}_{\epsilon}\left[\boldsymbol{\epsilon}\right]\\
\mathbb{E}_{\mathcal{T}}\left[\hat{y}_0\right] &= x_0^T\beta\\
\mathrm{Var}_{\mathcal{T}}(\hat{y}_0) &= \mathbb{E}_{\mathcal{T}}\left[(p_X^T\boldsymbol{\epsilon})^2\right] = \mathbb{E}_{\mathcal{T}}\left[p_X^T\boldsymbol{\epsilon}\,\boldsymbol{\epsilon}^T p_X\right] = \mathbb{E}_{\mathcal{T}}\left[\mathrm{Tr}\left(p_X^T\boldsymbol{\epsilon}\,\boldsymbol{\epsilon}^T p_X\right)\right]\\
&= \mathbb{E}_{\mathcal{T}}\left[\mathrm{Tr}\left(p_X\,p_X^T\boldsymbol{\epsilon}\,\boldsymbol{\epsilon}^T\right)\right]\\
&= \mathrm{Tr}\left(\mathbb{E}_{\mathbf{X}}\left[p_X p_X^T\right]\mathbb{E}_{\boldsymbol{\epsilon}}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right]\right) = \sigma^2\mathbb{E}_{\mathbf{X}}\left[x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0\right]
\end{aligned}$$

The last equality follows from the cyclicity and linearity of trace. Using $\mathbb{E}_{\boldsymbol{\epsilon}}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right] = \sigma^2$ one has:

$$\begin{aligned}
\mathrm{Var}_{\mathcal{T}}(\hat{y}_0) &= \sigma^2\,\mathrm{Tr}\left(\mathbb{E}_{\mathbf{X}}\left[p_X p_X^T\right]\right)\\
&= \sigma^2\,\mathbb{E}_{\mathbf{X}}\left[\mathrm{Tr}\left(p_X p_X^T\right)\right] = \sigma^2\,\mathbb{E}_{\mathbf{X}}\left[||p_X||^2\right]\\
&= \sigma^2\mathbb{E}_{\mathbf{X}}\left[x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0\right]
\end{aligned}$$

Putting everything together:

$$\begin{aligned}
\mathrm{Var}_{y_0|x_0}(y_0) &= \sigma^2 \\
\mathrm{Bias}_{\mathcal{T},y_0|x_0}(y_0) &= 0 \\
\mathrm{Var}_{\mathcal{T}}(\hat{y}_0) &= \sigma^2\, \mathbb{E}_{\mathbf{X}}\left[x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0\right] \\
\mathrm{EPE}(x_0) &= \sigma^2\, \left(1 + \mathbb{E}_{\mathbf{X}}\left[x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0\right]\right)
\end{aligned}$$

which proves Eq. (2.27).

To prove Eq. (2.28) under the specified assumption:

$$(\mathbf{X}^T\mathbf{X})^{-1} \to N^{-1}\mathrm{Cov}^{-1}(X) \quad \text{as } N \to \infty$$

requires a simple manipulation involving the cyclicity and linearity of the trace operation, as well as the independence of $\mathbf{X}$ and x_0:

$$\begin{aligned}
\mathbb{E}_{x_0}\left[EPE(x_0)\right] &= \sigma^2\left(1 + \mathbb{E}_{x_0,\mathbf{X}}\left[x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0\right]\right) \\
&= \sigma^2\left(1 + \mathbb{E}_{x_0,\mathbf{X}}\left[\mathrm{Tr}\left(x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0\right)\right]\right) \\
&= \sigma^2\left(1 + \mathbb{E}_{x_0,\mathbf{X}}\left[\mathrm{Tr}\left(x_0x_0^T(\mathbf{X}^T\mathbf{X})^{-1}\right)\right]\right) \\
&= \sigma^2\left(1 + \mathrm{Tr}\left(\mathbb{E}_{x_0,\mathbf{X}}\left[x_0x_0^T(\mathbf{X}^T\mathbf{X})^{-1}\right]\right)\right) \\
&= \sigma^2\left(1 + \mathrm{Tr}\left(\mathbb{E}_{x_0}\left[x_0x_0^T\right]\mathbb{E}_{\mathbf{X}}\left[(\mathbf{X}^T\mathbf{X})^{-1}\right]\right)\right) \\
&= \sigma^2\left(1 + \mathrm{Tr}\left(\mathrm{Cov}(x_0)\,\mathbb{E}_{\mathbf{X}}\left[(\mathbf{X}^T\mathbf{X})^{-1}\right]\right)\right) \\
&\to \sigma^2\left(1 + \frac{1}{N}\mathrm{Tr}\left(\mathrm{Cov}(x_0)\,\mathrm{Cov}^{-1}(X)\right)\right) \\
&= \sigma^2\left(1 + \frac{p}{N}\right)
\end{aligned}$$

The last equality follows from the fact that x_0 is drawn from the same distribution as X.

Ex. 2.6

One has:

$$\sum_{i:\, x_i = x} \left(y_i - f_\theta(x_i)\right)^2 = n_x \left(\bar{y}_x - f_\theta(x)\right)^2 + \sum_{i:\, x_i = x} \left(y_i - \bar{y}_x\right)^2$$

where $n_x \equiv \sum_{i:\, x_i = x} 1$ and $\bar{y}_x \equiv n_x^{-1} \sum_{i:\, x_i = x} y_i$. The second term does not contribute to the regression since it does not depend on f_θ, so we can just fit using the average $\bar{y}_x$ as response and weights proportional to n_x.

Ex. 2.7

Point (a)

One has:

$$\begin{array}{rl}
\text{linear regression:} & \hat{f}(x_0) = x_0^T\hat{\beta} = x_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TY \\
& l_i(x_0, \mathcal{X}) = \left(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x_0\right)_i \\
\text{k-nn:} & l_i(x_0, \mathcal{X}) = \dfrac{1}{k}\,\mathbf{I}\,(x_i \text{ is among the } k \text{ closest neighbors of } x_0)
\end{array}$$

Point (b)

Notice that:

$$\begin{aligned}
\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\hat{f}(x_0)\right] &= \boldsymbol{l}^T(x_0;\mathcal{X})\,\boldsymbol{f}(X), \\
\mathrm{Var}_{\mathcal{Y}|\mathcal{X}}\left(\hat{f}(x_0)\right) &= \boldsymbol{l}^T(x_0;\mathcal{X})\,\mathrm{Cov}(\boldsymbol{\epsilon})\,\boldsymbol{l}(x_0;\mathcal{X}) = \sigma^2\,||\boldsymbol{l}(x_0;\mathcal{X})||^2
\end{aligned}$$

where we denoted $(\boldsymbol{l}(x_0;\mathcal{X}))_i \equiv l_i(x_0;\mathcal{X})$ and $(\boldsymbol{f}(X))_i \equiv f(x_i)$. Hence:

$$\begin{aligned}
\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\left(f(x_0) - \hat{f}(x_0)\right)^2\right] &= \mathrm{Bias}^2_{\mathcal{Y}|\mathcal{X}}(y_0) + \mathrm{Var}_{\mathcal{Y}|\mathcal{X}}(y_0) \\
\mathrm{Bias}^2_{\mathcal{Y}|\mathcal{X}}(y_0) &\equiv \left(f(x_0) - \boldsymbol{l}^T(x_0;\mathcal{X})\,\boldsymbol{f}(X)\right)^2 \\
\mathrm{Var}_{\mathcal{Y}|\mathcal{X}}(y_0) &\equiv \sigma^2||\boldsymbol{l}(x_0;\mathcal{X})||^2
\end{aligned}$$

The first term represents the bias, while the second represents the variance of the estimator as the training responses vary for fixed $\mathcal{X}$.

Point (c)

For this part can go down a similar road, using now the independence of X and ϵ:

$$\begin{aligned}
\mathbb{E}_{\mathcal{Y},\mathcal{X}}\left[\hat{f}(x_0)\right] &= \mathbb{E}_{\mathcal{X}}\left[\boldsymbol{l}^T(x_0;\mathcal{X})\,\boldsymbol{f}(X)\right], \\
\mathrm{Var}_{\mathcal{Y},\mathcal{X}}\left(\hat{f}(x_0)\right) &= \mathrm{Var}_{\mathcal{X}}\left(\boldsymbol{l}^T(x_0;\mathcal{X})\,\boldsymbol{f}(X)\right) + \sigma^2\mathbb{E}_{\mathcal{X}}\left[||\boldsymbol{l}(x_0;\mathcal{X})||^2\right],
\end{aligned}$$

Hence:

$$\begin{aligned}
\mathbb{E}_{\mathcal{Y},\mathcal{X}}\left[\left(f(x_0) - \hat{f}(x_0)\right)^2\right] &= \mathrm{Bias}^2_{\mathcal{Y},\mathcal{X}}(y_0) + \mathrm{Var}_{\mathcal{Y},\mathcal{X}}(y_0) \\
\mathrm{Bias}^2_{\mathcal{Y},\mathcal{X}}(y_0) &\equiv \left(f(x_0) - \mathbb{E}_{\mathcal{X}}\left[\boldsymbol{l}^T(x_0;\mathcal{X})\,\boldsymbol{f}(X)\right]\right)^2 \\
\mathrm{Var}_{\mathcal{Y},\mathcal{X}}(y_0) &\equiv \mathrm{Var}_{\mathcal{X}}\left(\boldsymbol{l}^T(x_0;\mathcal{X})\,\boldsymbol{f}(X)\right) + \sigma^2\mathbb{E}_{\mathcal{X}}\left[||\boldsymbol{l}(x_0;\mathcal{X})||^2\right]
\end{aligned}$$

Point (d)

Combining the equations above one can see that:

$$\mathrm{Bias}^2_{\mathcal{Y},\mathcal{X}}(y_0) + \mathrm{Var}_{\mathcal{Y},\mathcal{X}}(y_0) = \mathbb{E}_{\mathcal{X}}\left(\mathrm{Bias}^2_{\mathcal{Y}|\mathcal{X}}(y_0) + \mathrm{Var}_{\mathcal{Y}|\mathcal{X}}(y_0)\right)$$

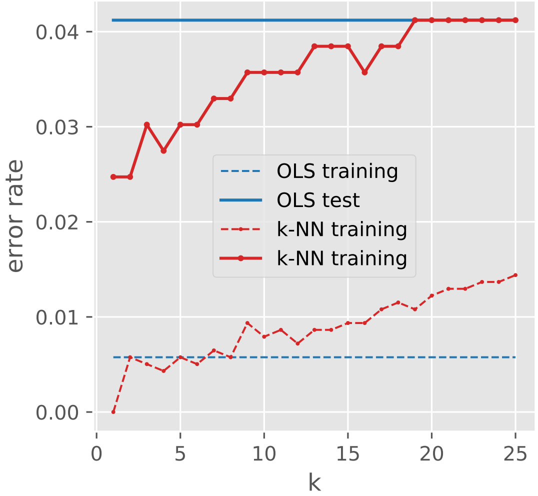
which is a simple consequence of the conditional expectations identity:

$$\mathbb{E}_{\mathcal{Y},\mathcal{X}}\left[\left(f(x_0) - \hat{f}(x_0)\right)^2\right] = \mathbb{E}_{\mathcal{X}}\left[\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\left(f(x_0) - \hat{f}(x_0)\right)^2\right]\right]$$

This might be the relationship the authors wanted us to find.

Ex. 2.8

We decide to restrict the data prior to training to the examples corresponding to the digits 2 and 3. The results are summarized below:



Model	Training error	Test error
OLS	0.58 %	4.12 %
1-NN	0	2.47 %
3-NN	0.50 %	3.02 %
5-NN	0.58 %	3.02 %
7-NN	0.65 %	3.30 %
15-NN	0.94 %	3.85 %

Ex. 2.9

Let's denote $R(Z_{tr}, Z_{te})$ the average of squared residuals for Z_{te} using the OLS coefficients from Z_{tr}. Using the notation from the exercise:

$$R_{tr}(\hat{\beta}) \equiv R(Z_{tr}, Z_{tr}), \qquad R_{te}(\hat{\beta}) \equiv R(Z_{tr}, Z_{te})$$

It is easy to verify that the expected value $\mathbb{E}_{Z_{te}}\left[R(Z_{tr}, Z_{te})\right]$ does not depend on $M \equiv |Z_{te}|$ (assuming the test examples are i.i.d.). This allows us to take $M = N$. Now, by definition of OLS estimates we have:

$$R(Z_{te}, Z_{te}) \leq R(Z_{tr}, Z_{te})$$

Now that $M = N$, the lhs has the same distribution as $R(Z_{tr}, Z_{tr})$, so when taking the expectation value:

$$\mathbb{E}_{Z_{te},Z_{tr}}\left[R(Z_{tr}, Z_{tr})\right] = \mathbb{E}_{Z_{te},Z_{tr}}\left[R(Z_{te}, Z_{te})\right] \leq \mathbb{E}_{Z_{te},Z_{tr}}\left[R(Z_{tr}, Z_{te})\right]$$

which proves the assertion.

Chapter 3: Linear methods for regression

Ex. 3.1

We can use the results from Section 3.2.3 to prove the statement for the last predictor $j = p$. This generalizes for all other predictors, since neither the z-score nor the F-statistic depend on the order of predictors.

Letting $\mathbf{z}_j$ be the columns of Z as in Eq. (3.30) in the text and assuming additive gaussian errors, we have the following:

$$\begin{aligned} \mathrm{Var}(\hat{\beta}_p) &= \frac{\sigma^2}{||\mathbf{z}_p||^2} \qquad \text{(Eq. (3.29) in the text)} \\ \mathbf{z_j} \cdot \mathbf{z_k} &= 0 \quad \text{for } j \neq k \end{aligned}$$

The z-score for predictor j had been defined (see Eq. (3.12) in the text) as the ratio between the OLS value $\hat{\beta}_j$ and the square root of the estimate of $\mathrm{Var}(\hat{\beta}_p)$ obtained by replacing σ with $\hat{\sigma}$, hence:

$$z_p^2 \equiv \frac{\hat{\beta}_p^2}{\hat{\sigma}^2/||\mathbf{z}_p||^2}$$

Adopting the notation used to defined the F-score in Eq. (3.13) in the text:

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}_1}{N-p-1}$$

where RSS_1 is the sum of squared residuals when the p-th predictor is included. We have then:

$$z_p^2 = \frac{\hat{\beta}_p^2 \, ||\mathbf{z}_p||^2}{\mathrm{RSS}_1/(N-p-1)}$$

Since $\mathbf{z}_p$ is orthogonal to all other predictors, it is easy to check that:

$$\mathrm{RSS}_1 = \mathrm{RSS}_0 - \hat{\beta}_p^2 \, ||\mathbf{z}_p||^2$$

Hence, according to Eq. (3.13) in the text:

$$F_p \equiv \frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\mathrm{RSS}_1/(N-p-1)} = \frac{\hat{\beta}_p^2 \, ||\mathbf{z}_p||^2}{\mathrm{RSS}_1/(N-p-1)} = z_p^2$$

Ex. 3.2

(see Jupyter Notebook)

The premise of the two confidence band estimates is that the posterior distribution of the true value β given the data is that of a gaussian vector, with center $\hat{\beta}$ and covariance $\hat{\sigma}^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$. In the case described by the exercise $\mathbf{X}$ is the matrix whose columns contains increasing powers of the samples of X. The two confidence bands are then obtained as follows:

1. Since $\beta^T x_0$ is also gaussian with mean $\hat{y}_0 \equiv \hat{\beta}^T x_0$ and variance $\hat{\sigma}^2\, x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0$, we can define a confidence band for its value via:

$$\mathcal{C}_1 = \left\{ y \;:\; (y-\hat{y}_0)^2 \leq \hat{\sigma}^2\, x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0\, {\chi_1^2}^{(1-\alpha)} \right\}$$

where ${\chi_1^2}^{(1-\alpha)}$ is the $1-\alpha$ percentile of a chi-squared distribution with one degree of freedom, i.e. the distribution of the squared of a normal random variable:

$$P(\beta^T x_0 \in \mathcal{C}_1 \,|\, Y) = 1-\alpha$$

2. One has (cf. Eq. (3.15) in the text):

$$(\hat{\beta}-\beta)^T\,\mathbf{X}^T\mathbf{X}\,(\hat{\beta}-\beta) \sim \hat{\sigma}^2\chi^2_{p+1}$$

We can thus define a confidence interval for the whole vector β as:

$$\mathcal{C}_{2,\beta} = \left\{ \beta \;:\; \frac{1}{\hat{\sigma}^2}(\hat{\beta}-\beta)^T\,\mathbf{X}^T\mathbf{X}\,(\hat{\beta}-\beta) \leq {\chi^2_{p+1}}^{(1-\alpha)} \right\}$$
$$P(\beta \in \mathcal{C}_{2,\beta} \,|\, Y) = 1-\alpha$$

This in turns generate a confidence interval for $\beta^T x_0$ as:

$$\mathcal{C}_2 = \left\{ y = \beta^T x_0 :\; \beta \in \mathcal{C}_{2,\beta} \right\}$$

The two confidence bands are very much related, as we now show. First, one can easily show that both $\mathcal{C}_1$ and $\mathcal{C}_2$ do not change if we start with a different set of predictors, related to the original ones via a non-singular linear transformation. Hence, we can assume that the predictors are orthogonal and normalised, $\mathbf{X}^T\mathbf{X} = \mathbb{I}_p$. We then have:

$$\mathcal{C}_1 = \left\{ y \;:\; (y-\hat{y}_0)^2 \leq \hat{\sigma}^2\, ||x_0||^2\, {\chi_1^2}^{(1-\alpha)} \right\} \tag{16}$$

and:

$$\mathcal{C}_{2,\beta} = \left\{ \beta \;:\; ||\hat{\beta}-\beta||^2 \leq \hat{\sigma}^2{\chi^2_{p+1}}^{(1-\alpha)} \right\}$$

or equivalently:

$$\begin{aligned} \mathcal{C}_{2,\delta\beta} &\equiv \left\{ \delta\beta \;:\; ||\delta\beta||^2 \leq \hat{\sigma}^2\, {\chi^2_{p+1}}^{(1-\alpha)} \right\} \\ \mathcal{C}_2 &= \left\{ y = \hat{y}_0 + \delta\beta^T x_0 :\; \delta\beta \in \mathcal{C}_{2,\delta\beta} \right\} \end{aligned}$$

One can easily prove that $\mathcal{C}_2$ is, like $\mathcal{C}_1$, an interval centered around $\hat{y_0}$. Finding its upper limit corresponds to solving:

$$\max \delta\beta^T x_0 \quad \text{on} \quad ||\delta\beta||^2 \leq \hat{\sigma}^2 \, {\chi^2_{p+1}}^{(1-\alpha)}$$

The solution is:

$$\max \delta\beta^T x_0 = \left(\hat{\sigma}^2 ||x_0||^2 \, {\chi^2_{p+1}}^{(1-\alpha)}\right)^{1/2}$$

Hence:

$$\mathcal{C}_2 = \left\{ y \; : \; (y - \hat{y_0})^2 \leq \hat{\sigma}^2 ||x_0||^2 \, {\chi^2_{p+1}}^{(1-\alpha)} \right\} \tag{17}$$

Comparing (17) with (16) we see that:

$$\frac{\mathrm{diam}^2(\mathcal{C}_2)}{\mathrm{diam}^2(\mathcal{C}_1)} = \frac{{\chi^2_{p+1}}^{(1-\alpha)}}{{\chi^2_1}^{(1-\alpha)}} \geq 1 \tag{18}$$

The equality holds only for $p = 0$, i.e. when we only fit a constant, or more generally with only a single predictor. So, the point-wise confidence bands are narrower.

We can also provide a simple graphical interpretation for this result. First, notice that the size of both confidence intervals scales linearly with $||x_0||$, hence to get an idea of interval sizes we can set $||x_0|| = 1$. The quantity $\beta^T x_0$ can now be interpreted as the euclidean projection of the random β vector onto the unit vector x_0. Using orthonormal predictors, the confidence band size is independent on the direction of x_0 (see (17) and (16)) and we can set $x_0 = e_1$, hence $\beta^T x_0 = \beta_1$. With this choice, we see that the first confidence band estimate corresponds to a band $\{|\beta_1 - \hat{\beta}_1| \leq c\}$ with probability $1 - \alpha$. On the other hand, the second choice consists in finding a ball in the β space with the same probability, then to project this set onto the e_1 axis. It is then obvious that the ball diameter needs to be wider than previous band-sized set in order for their probabilities to be the same (see Figures 1 and 2).

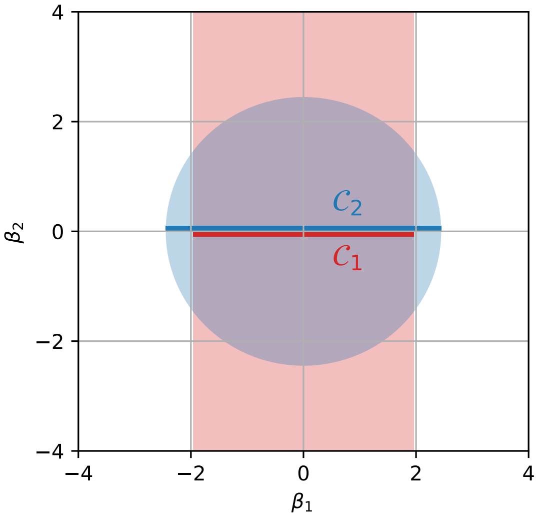


Figure 1: Confidence bands: single projection versus full vector. The two-dimensional vector β is taken to be centered and to have unit covariance. Both highlighted areas have 95% probability, but the band-shaped one has smaller projection on the β_1 axis.

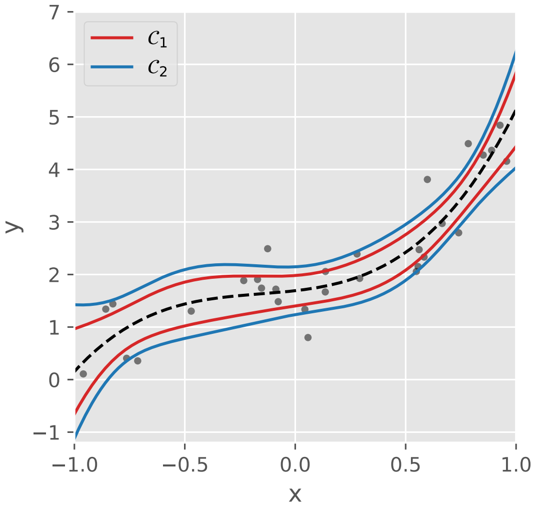


Figure 2: Confidence bands for a cubic univariate model. Values of x have been drawn from a uniform distribution in $[-1, 1]$, and y has been generated as $\beta^T x^{(3)} + \epsilon$, where β is a centered, unit-covariance gaussian vector, $x^{(3)} \equiv (1, x, x^2, x^3)$ and ϵ is a centered gaussian with variance equal to 0.25.

Ex. 3.3

Any linear, unbiased estimate $c^T Y$ of $a^T\beta$ can be written as the OLS estimator plus an additional linear estimate with zero expectation:

$$\begin{aligned} &c^T Y \equiv a^T\hat{\beta} + b^T Y = \left(a^T \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T + b^T\right) Y \\ &\mathbb{E}\left[b^T Y\right] = b^T\mathbf{X}\beta = 0 \end{aligned}$$

Since the second equation must hold for any β, we conclude $b^T\mathbf{X} = 0$. This implies that the two terms in:

$$c = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} a + b$$

are orthogonal vectors. Hence, the variance:

$$\text{Var}\left(c^T Y\right) \quad = \quad \sigma^2 ||c||^2$$

is minimized for $b = 0$, i.e. when $c^T Y = a^T\hat{\beta}$.

Similarly, any linear unbiased estimate $C^T Y$ of the β vector can be written as:

$$\begin{aligned} C^T Y &= \left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T + B^T\right) Y \equiv \left(\hat{B} + B\right)^T Y \\ B^T\mathbf{X} &= 0 \quad \Longrightarrow \quad B^T\hat{B} = 0 \end{aligned} \tag{19}$$

Its variance-covariance matrix is:

$$\text{Cov}(C^T Y) = C^T C = \hat{B}^T\hat{B} + B^T B$$

where the last equality follows from (19). Since $B^T B$ is positive semi-definite, this proves that:

$$\text{Cov}(\hat{\beta}) = \hat{B}^T\hat{B} \lesssim \text{Cov}(C^T Y)$$

Ex. 3.4

Using the QR decomposition $\mathbf{X} = \mathbf{Q}_{Np} R_{pp}$, $\mathbf{Q}^T\mathbf{Q} = \mathbb{I}$, we see that:

$$\begin{aligned}\hat{\beta} &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = R^{-1}\mathbf{Q}^T\mathbf{y} \equiv R^{-1}\hat{\beta}^q \\ \hat{\beta}^q &\equiv \mathbf{Q}^T\mathbf{y}\end{aligned}$$

This identity can be given a procedural interpretation via the Gram-Schmidt procedure.

1. Start with $\hat{\beta} = 0$.

2. For $j = 1, \ldots, p$:

   - Compute the j-th column $\mathbf{q}_j$ of Q, and the j-th column of R as per the Gram-Schmit procedure. This corresponds to:

     $$\mathbf{x}_j = R_{1j}\mathbf{q}_1 + \ldots + R_{jj}\mathbf{q}_j$$

     where $\mathbf{q}_j$ satisfies $\langle \mathbf{q}_i, \mathbf{q}_j \rangle = \delta_{ij}$.

   - Compute the regression coefficient $\hat{\beta}^q_j$ of $\mathbf{y}$ against $\mathbf{q}_j$ as $\langle \mathbf{q}_j, \mathbf{y} \rangle$. Notice that, the $\mathbf{q}_j$'s being orthogonal, $\hat{\beta}^q_j$ is the coefficient of $\mathbf{q}_j$ in the regression of $\mathbf{y}$ against all of the $\mathbf{q}_j$'s:

     $$\mathbf{y} \sim \hat{\beta}^q_1\,\mathbf{q}_1 + \ldots + \hat{\beta}^q_p\,\mathbf{q}_p$$

   - Compute the j-th column of R^{-1} via backward substitution. Notice that this only requires the first j columns of R, which are available at this stage. We have then:

     $$\mathbf{q}_j = (R^{-1})_{1j}\,\mathbf{x}_1 + \ldots + (R^{-1})_{jj}\,\mathbf{x}_j \tag{20}$$

   - Update the coefficients of the original predictors according to the new term in the regression:

     $$\begin{aligned}\hat{\beta}^q_j\,\mathbf{q}_j &= \hat{\beta}^q_j\left((R^{-1})_{1j}\,\mathbf{x}_1 + \ldots + (R^{-1})_{jj}\,\mathbf{x}_j\right) \\ \hat{\beta}_i &\to \hat{\beta}_i + (R^{-1})_{ij}\,\hat{\beta}^q_j \quad i = 1, \ldots, j\end{aligned}$$

Ex. 3.5

One has:

$$\begin{aligned}
\beta_0 + \sum_j x_{ij}\beta_j &= \beta_0^c + \sum_j (x_{ij} - \bar{x}_j)\beta_j \\
\beta_0^c &\equiv \beta_0 + \sum_j \bar{x}_j \beta_j
\end{aligned}$$

Since β_0 is a dummy variable in the OLS minimization, we can use β_0^c instead, which proves the equivalence and provides the relationship between β and β^c:

$$\begin{aligned}
\beta_0^c &= \beta_0 + \sum_j \bar{x}_j \beta_j \\
\beta_j^c &= \beta_j, \quad j = 1, \ldots, p
\end{aligned}$$

As a side note, this shows that as long as we include the constant in a Ridge regression, we should not worry about the remaining predictors being centered: their average can be absorbed in the coefficient of the constant which is not penalized.

Ex. 3.6

From Bayes theorem:

$$P_{posterior}(\beta|\mathbf{y}) \propto P(\mathbf{y}|\beta)\, P_{prior}(\beta)$$

The proportionality factor only depends on $\mathbf{y}$ and is fixed by the normalization of the posterior distribution. Assuming[1]:

$$\begin{aligned}
l_{prior}(\beta) &\equiv \log P_{prior}(\beta) = c_1 - \frac{1}{2\tau^2}\beta^T\beta \\
\log P(\mathbf{y}|\beta) &= c_2 - \frac{1}{2\sigma^2}\left(\mathbf{y} - \mathbf{X}\beta\right)^T\left(\mathbf{y} - \mathbf{X}\beta\right)
\end{aligned}$$

we get:

$$\begin{aligned}
l_{posterior}(\beta|\mathbf{y}) &= c_3 - \frac{1}{2\tau^2}\beta^T\beta - \frac{1}{2\sigma^2}\left(\mathbf{y} - \mathbf{X}\beta\right)^T\left(\mathbf{y} - \mathbf{X}\beta\right) \qquad (21) \\
&= c(\mathbf{y}) - \frac{1}{2}\left(\beta - \hat{\beta}\right)^T\left(\frac{\mathbf{X}^T\mathbf{X}}{\sigma^2} + \frac{1}{\tau^2}\right)\left(\beta - \hat{\beta}\right) \\
\hat{\beta} &= \left(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\tau^2}\right)^{-1}\mathbf{X}^T\mathbf{y}
\end{aligned}$$

This shows that the posterior distribution of β is gaussian, with center given by the Ridge estimate with parameter $\lambda = \sigma^2/\tau^2$.

[1] The text exercise uses τ instead of τ^2. This seems inconsistent with the next exercise and with the convention for variances, so we use τ^2 instead.

Ex. 3.7

This is almost equivalent to (21), which was derived in the previous exercise. The only difference is the presence of the constant β_0. In the absence of a prior for β_0, the expression in the exercise:

$$\sum_{i=1}^{N}\left(y_i - \beta_0 \sum_j x_{ij}\beta_j\right)^2 + \lambda \sum_j \beta_j^2$$

is not a valid log-likelihood for $\beta \equiv (\beta_j)_{j=1,\ldots,p}$, since β_0 remains unspecified. Two possibilities come to mind: first, that β_0 is deterministic ($\tau_0 = 0$) and known in advance, in which case (21) yields the results in the exercise when we replace $\mathbf{y}$ by $\mathbf{y} - \beta_0$.

The second, more likely possibility is that we have no prior on β_0. This can be modelled as a gaussian prior with $\tau_0 \to \infty$, in which case we can repeat the steps of the previous exercise to get:

$$l_{posterior}(\beta, \beta_0|\mathbf{y}) = c - \frac{1}{2\tau_0^2}\beta_0^2 - \frac{1}{2\tau^2}\beta^T\beta - \frac{1}{2\sigma^2}\left(\mathbf{y} - \beta_0 - \mathbf{X}\beta\right)^T\left(\mathbf{y} - \beta_0 - \mathbf{X}\beta\right)$$

Since we are interested in the posterior distribution for β, we should integrate over β_0 the exponential of $l_{posterior}(\beta, \beta_0|\mathbf{y})$. As $\tau_0 \to \infty$, one can show that upon integration:

$$\begin{aligned} l_{posterior}(\beta|\mathbf{y}) &= c(\mathbf{y}) - \frac{1}{2\tau^2}\beta^T\beta - \frac{1}{2\sigma^2}\left(\mathbf{y} - \mathbf{X}\beta\right)^T\left(\mathbb{I} - \frac{ee^T}{N}\right)\left(\mathbf{y} - \mathbf{X}\beta\right) \\ &= c(\mathbf{y}) - \frac{1}{2\tau^2}\beta^T\beta - \frac{1}{2\sigma^2}\left(\mathbf{y}_c - \mathbf{X}_c\beta\right)^T\left(\mathbf{y}_c - \mathbf{X}_c\beta\right) \end{aligned}$$

where $\mathbf{y}_c$ and $\mathbf{X}_c$ are the centered versions of $\mathbf{y}$ and $\mathbf{X}$.

Ex. 3.8

We begin by proving that $\mathbf{Q}_2$ is the 'Q' matrix in the QR decomposition of $\tilde{\mathbf{X}}$. Denote $(\mathbf{q}_0 = \mathbf{e}, \mathbf{q}_1, \ldots, \mathbf{q}_p)$ the columns of $\mathbf{Q}$, the matrix appearing in the QR decomposition $\mathbf{X}$. The matrix $\mathbf{Q}_2$ has columns $(\mathbf{q}_1, \ldots, \mathbf{q}_p)$. The centered matrix $\tilde{\mathbf{X}}$ can be written as:

$$\tilde{\mathbf{X}} = \left(\mathbf{X} - \frac{1}{N}\mathbf{e}\,\mathbf{e}^T\mathbf{X}\right)E_{p+1,p} \tag{22}$$

where $E_{p+1,p}$ is the $(p+1) \times p$ matrix containing an identity block starting at position $(1,0)$, so as to remove the first column of the matrix on its left. Replacing the QR decomposition of $\mathbf{X}$ we get:

$$\tilde{\mathbf{X}} = \left(\mathbb{I}_{p+1,p+1} - \frac{1}{N}\mathbf{e}\,\mathbf{e}^T\right)\mathbf{Q}\,R\,E_{p+1,p} \tag{23}$$

Since $\mathbf{q}_0 = \mathbf{e}$ and the remaining columns of $\mathbf{q}$ are orthogonal to $\mathbf{e}$ as per the Gram-Schmidt procedure, the first term on the r.h.s. has the effect of multiplying by zero the first column of $\mathbf{Q}$ while leaving the remaining columns $\mathbf{Q}_2$ untouched:

$$\tilde{\mathbf{X}} = \mathbf{Q}_2\, E_{p,p+1}\, R\, E_{p+1,p} \tag{24}$$

where $E_{p,p+1}$ is the $p \times (p+1)$ matrix containing an identity block starting at index $(0,1)$. The matrix $E_{p,p+1}\, R\, E_{p+1,p}$ is nothing but the 1 to $p+1$ submatrix of R, which is still upper diagonal. This proves that $\mathbf{Q}_2$ is the 'Q' matrix in the QR decomposition of $\tilde{\mathbf{X}}$.

Let now $\tilde{\mathbf{X}} = UDV^T$. Since DV^T is a non-singular $p \times p$ matrix, the columns of U must span the same subspace as the columns of both $\tilde{\mathbf{X}}$ and $\mathbf{Q}_2$. From now on we can drop both the tildes and the subscript, since both the QR and SVD decompositions are performed on the centered matrix.

For the second part of the question we note that, if the columns of $\mathbf{X}$ are mutually orthogonal, the QR decomposition is also an SVD decomposition, because the R matrix can be taken to be diagonal:

$$\begin{aligned} \mathbf{X} &= \mathbf{Q}D \\ D_{kk} &= ||\mathbf{x}_k|| \\ \mathbf{Q}^T\mathbf{Q} &= \mathbb{I}_p \end{aligned}$$

Since the SVD is unique[2] up to shuffling and sign flip of the columns of $\mathbf{U}$, the matrix $\mathbf{Q}$ can be obtained from any of the versions of $\mathbf{U}$ using such operations. One can in fact show that this is the only case in which the columns of $\mathbf{Q}$ coincide those of $\mathbf{U}$. Suppose that:

$$\mathbf{Q} = \mathbf{U}$$

By comparing the two decompositions one can easily show that:

$$V = D^{-1}R$$

This equality requires V to be upper-diagonal besides being orthogonal. This implies[3] that V is a diagonal, idempotent matrix, and:

$$\mathbf{X}^T\mathbf{X} = V^T D^2 V$$

is also diagonal, proving that the columns of $\mathbf{X}$ are mutually orthogonal.

[2] Here we suppose that the eigenvalues of the sample covariance matrix are all distinct.

[3] Any upper-diagonal orthogonal matrix V is necessarily diagonal. To show this, one can prove that V^{-1} is also upper-diagonal and, being equal to the transpose V^T, V must be diagonal.

Ex. 3.9

If we add a new predictor $\mathbf{z}$ to the fit, the residual-sum-of squares changes by:

$$RSS \longrightarrow RSS - \left(\frac{\langle \mathbf{y}, \mathbf{z}^{(r)} \rangle}{||\mathbf{z}^{(r)}||} \right)^2$$

where $\mathbf{z}^{(r)}$ is the OLS residual of $\mathbf{z}$ against the predictors already included in the fit, $\mathbf{X}_1$ in this case. Hence, it suffices to find the matrix $\mathbf{X}_2^{(r)}$ of residuals, which is easy to compute using the QR decomposition $\mathbf{X}_1 = \mathbf{Q}R$:

$$\begin{aligned} \mathbf{X}_2^{(r)} &= \mathbf{X}_2 - \mathbf{X}_1 \left(\mathbf{X}_1^T \mathbf{X}_1 \right)^{-1} \mathbf{X}_1^T \mathbf{X}_2 && (25) \\ &= \left(\mathbb{I} - \mathbf{Q}\mathbf{Q}^T \right) \mathbf{X}_2 && (26) \end{aligned}$$

Then, we choose the predictor $\mathbf{x}_{2,\bar{j}}$ such that:

$$\bar{j} = \operatorname{argmax} \frac{|\langle \mathbf{y}, \mathbf{x}_{2,j}^{(r)} \rangle|}{||\mathbf{x}_{2,j}^{(r)}||} = \operatorname{argmax} \frac{|\langle \mathbf{r}, \mathbf{x}_{2,j}^{(r)} \rangle|}{||\mathbf{x}_{2,j}^{(r)}||} = \operatorname{argmax} \frac{|\langle \mathbf{r}, \mathbf{x}_{2,j} \rangle|}{||\mathbf{x}_{2,j}^{(r)}||} \qquad (27)$$

The last two equalities follows from the fact that, by construction, $\mathbf{x}_{2,j}^{(r)}$ and $\mathbf{r}$ are orthogonal to the subspace generated by the columns of $\mathbf{X}_1$, hence to $\hat{\mathbf{y}} = \mathbf{y} - \mathbf{r}$ and $\mathbf{x}_{2,j} - \mathbf{x}_{2,j}^{(r)}$. The process can be continued by adding the column $\mathbf{x}_{2,\bar{j}}^{(r)}/||\mathbf{x}_{2,\bar{j}}^{(r)}||$ to Q, and updating R according to (25).

Note how one has generally $||\mathbf{x}_{2,j}^{(r)}|| \leq ||\mathbf{x}_{2,j}||$, the equality holding only when $\mathbf{x}_{2,j}$ is orthogonal to $\mathbf{X}_1$. Hence, the criterion (27) is not equivalent to:

$$\bar{j} = \operatorname{argmax} \frac{|\langle \mathbf{r}, \mathbf{x}_{2,j} \rangle|}{||\mathbf{x}_{2,j}||} \qquad (28)$$

Indeed, in forward stepwise regression we allow ourselves to change the coefficients of the predictors in the current active set $\mathbf{X}_1$, so as to only pick the ”new” part in $\mathbf{x}_{2,j}$. Conversely, criterion (28), which corresponds to forward stagewise regression, tends to penalize predictors which are strongly correlated with the predictors in the active set.

Ex. 3.10

In Exercise 3.1, we established that the F-score for dropping a single predictor equals the square of the corresponding z-score. Since the F score is proportional to the increase in residual-sum-of squares when the predictor is dropped, we see that the predictor with smallest absolute z-score is be the one to be dropped.

Ex. 3.11

Note that:

$$\mathrm{RSS}(B) = \mathrm{Tr}\left(\Sigma^{-1}\left(\mathbf{Y}-\mathbf{X}B\right)^T\left(\mathbf{Y}-\mathbf{X}B\right)\right)$$

When the matrix Σ is diagonal, this is a sum of K independent RSS's, and the solution for B is indeed:

$$B = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}\mathbf{Y} \tag{29}$$

More generally, since $\boldsymbol{\Sigma}$ is symmetric and positive definite, it admits a square root S, which allows us to fall back on the diagonal case:

$$\begin{aligned}
\mathrm{RSS}(B) &= \mathrm{Tr}\left(\left(\hat{\mathbf{Y}}-\mathbf{X}\hat{B}\right)^T\left(\hat{\mathbf{Y}}-\mathbf{X}\hat{B}\right)\right)\\
\hat{\mathbf{Y}} &\equiv \mathbf{Y}\,S\\
\hat{B} &\equiv B\,S
\end{aligned}$$

Hence the OLS solution is given by:

$$B = \hat{B}\,S^{-1} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}\hat{\mathbf{Y}}\,S^{-1} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}\mathbf{Y}$$

When the matrix Σ varies from instance to instance, the value of B which minimizes RSS is no longer given by a simple formula like (29). However, RSS is still a quadratic form in B, seen as a vector in $\mathbb{R}^{p\times K}$. So, the equation which determines B can be shown to be an affine equation:

$$\begin{aligned}
&\sum_{(k,b)} L_{(j,a),(k,b)}\,B_{(k,b)} - C_{j,a} = 0\\
&L_{(j,a),(k,b)} = \sum_i \Sigma^{-1}_{i,ab}\,X_{ij}\,X_{ik}\\
&C_{j,a} = \sum_{i,b}\Sigma^{-1}_{i,ab}\,X_{ij}Y_{ib}
\end{aligned}$$

This can be solved using standard linear system methods.

Ex. 3.12

Denote by $\mathbf{X}_a$ and $\mathbf{y}_a$ the matrices $\mathbf{X}$ and $\mathbf{y}$ after the indicated rows have been added. One has:

$$\begin{aligned}
\mathbf{X}_a^T\mathbf{X}_a &= \left(\begin{array}{cc} \mathbf{X}^T & \sqrt{\lambda}\,\mathbb{I} \end{array} \right) \left(\begin{array}{c} \mathbf{X} \\ \sqrt{\lambda}\,\mathbb{I} \end{array} \right) = \mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\,\mathbb{I} \\
\mathbf{X}_a^T\mathbf{Y}_a &= \left(\begin{array}{cc} \mathbf{X}^T & \sqrt{\lambda}\,\mathbb{I} \end{array} \right) \left(\begin{array}{c} \mathbf{y} \\ 0 \end{array} \right) = \mathbf{X}^T\,\mathbf{y}
\end{aligned}$$

therefore $\hat{\beta}_a = \left(\mathbf{X}_a^T\mathbf{X}_a\right)^{-1}\mathbf{X}_a^T\mathbf{y} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y} = \hat{\beta}_r(\lambda)$.

Ex. 3.13

After centering predictors, the PCR fitted function is:

$$\hat{f}(x) = \bar{y} + \sum_{m=1}^{M} \hat{\theta}_m\, z_m = \bar{y} + \sum_{m=1}^{M} \hat{\theta}_m\, (x \cdot v_m) = \bar{y} + \left(\sum_{m=1}^{M} \hat{\theta}_m\, v_m \right) \cdot x = \bar{y} + \hat{\beta}^{pcr} \cdot x$$

which proves the first point.

The proof of the second point follows from the fact that the span of the $\mathbf{x}$'s is equal to the span of all the $\mathbf{z}$'s. Remembering that $\mathbf{X} = \mathbf{Z}V^T$ where $\mathbf{z}_m$ are the columns of $\mathbf{Z}$ and V is orthogonal:

$$\hat{\beta}^{ls} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = V\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{y}$$

Since the columns of $\mathbf{Z}$ are orthogonal, this reads:

$$\hat{\beta}^{ls} = \sum_{m=1}^{p} \frac{\langle \mathbf{z}_m, \mathbf{y}\rangle}{\langle \mathbf{z}_m, \mathbf{z_m}\rangle}\, v_m = \hat{\beta}^{pcr}(p)$$

Ex. 3.14

When predictors are orthogonal and normalised ($\langle \mathbf{x}_j, \mathbf{x}_k\rangle = N\,\delta_{jk}$), the vector $\hat{\theta}_1\,\mathbf{z}_1$ found at the first step of the PLS procedure is the orthogonal projection of $\mathbf{y}$ on the span of the $\mathbf{x}$'s:

$$\begin{aligned}
\mathbf{z}_1 &= \sum_{j=1}^{p} \langle \mathbf{x}_j, \mathbf{y}\rangle\, \mathbf{x}_j \\
\hat{\theta}_1 &= \frac{\langle \mathbf{z}_1, \mathbf{y}\rangle}{\langle \mathbf{z}_1, \mathbf{z}_1\rangle} = \frac{\sum_{j=1}^{p} \langle \mathbf{x}_j, \mathbf{y}\rangle^2}{N\sum_{j=1}^{p} \langle \mathbf{x}_j, \mathbf{y}\rangle^2} = 1/N \\
\hat{\theta}_1\mathbf{z}_1 &= \sum_{j=1}^{p} \frac{\langle \mathbf{x}_j, \mathbf{y}\rangle}{N}\, \mathbf{x}_j = \sum_{j=1}^{p} \frac{\langle \mathbf{x}_j, \mathbf{y}\rangle}{\langle \mathbf{x}_j, \mathbf{x}_j\rangle}\, \mathbf{x}_j = \hat{\mathbf{y}}^{ls}
\end{aligned}$$

In other words, at step $m = 1$ we find the OLS fit of $\mathbf{y}$ against the $\mathbf{x}$'s. In preparation for the next step, the $\mathbf{x}_j$ are residualised against $\mathbf{z}_1$: we obtain a set of vectors $\mathbf{x}_j^{(1)}$ which still lie in the span of the $\mathbf{x}$'s, but are orthogonal to the projection of $\mathbf{y}$ onto the same subspace. Therefore, at the next step, we'll have:

$$\hat{\varphi}_{2j} = \langle \mathbf{x}_j^{(1)}, \mathbf{y} \rangle = \langle \mathbf{x}_j^{(1)}, \hat{\mathbf{y}}^{ls} \rangle = 0$$

This makes step $m = 2$ and all other steps void, since all $\mathbf{z}_m$'s are zero for $m \geq 2$.

Ex. 3.15

I believe that the p-dimensional vectors $\hat{\varphi}_m$ produced by Algorithm 3.3 in the text are distinct from the $\hat{\phi}_m$ produced by:

$$\begin{aligned} \hat{\phi}_m = \operatorname{argmax}_\alpha \operatorname{Corr}^2(\mathbf{y}, \mathbf{X}\alpha)\ \operatorname{Var}(\mathbf{X}\alpha) \\ \text{subject to}\, ||\alpha|| = 1, \quad \alpha^T S \hat{\phi}_l = 0,\ l = 1, \ldots, m-1 \end{aligned} \tag{30}$$

Indeed, the $\hat{\varphi}_m$ not only are not normalized, but do not satisfy $\hat{\varphi}_m S \hat{\varphi}_l = 0$ for $l \neq m$. For example one can show that:

$$\hat{\varphi}_2 = \hat{\varphi}_1 - \frac{||\hat{\varphi}_1||^2}{\hat{\varphi}_1^T S \hat{\varphi}_1} S \hat{\varphi}_1$$

hence:

$$\hat{\varphi}_1^T S \hat{\varphi}_2 = \hat{\varphi}_1^T S \hat{\varphi}_1 - ||\hat{\varphi}_1||^2 \frac{\hat{\varphi}_1^T S^2 \hat{\varphi}_1}{\hat{\varphi}_1^T S \hat{\varphi}_1}$$

The rhs is generally different from zero.

I believe this is only a notation collision, as the statement can be easily corrected. Since the $\mathbf{x}_j^{(m)}$ are constructed as linear combinations of the original $\mathbf{x}_j$'s, one can write the $\mathbf{z}_m$'s as linear combinations of the $\mathbf{x}_j$'s:

$$\mathbf{z}_m = \sum_{j=1}^{p} (\hat{\varphi}_m)_j\, \mathbf{x}_j^{(m-1)} = \sum_{j=1}^{p} \left(\hat{\phi}_m\right)_j \mathbf{x}_j \tag{31}$$

I believe the correct statement is: the $\hat{\phi}_m$'s in this equation can be obtained via (30). Indeed, the sample covariance between $\mathbf{z}_m$ and $\mathbf{z}_l$ is given by $\hat{\phi}_m^T S \hat{\phi}_l$, not by $\hat{\varphi}_m^T S \hat{\varphi}_l$. This is because the $\hat{\varphi}_m$ coordinate vectors correspond to different basis vectors $(\mathbf{x}^{(m-1)})_j \neq \mathbf{x}_j$, whose covariance is not given by S.

To recap, consider the two sets of vectors:

$$\begin{aligned} \hat{\phi}_m &= \operatorname{argmax}_\alpha \operatorname{Corr}^2(\mathbf{y}, \mathbf{X}\alpha)\ \operatorname{Var}(\mathbf{X}\alpha) = \operatorname{argmax}_\alpha \left(\alpha^T \mathbf{X}^T \mathbf{y}\right)^2 && (32) \\ & \quad \text{subject to}\, ||\alpha|| = 1, \quad \alpha^T S \hat{\phi}_l = 0,\ l = 1, \ldots, m-1 \\ \mathbf{v}_m &\equiv \mathbf{X}\, \hat{\phi}_m && (33) \end{aligned}$$

and the one produced by Algorithm 3.3:

$$\begin{aligned} \hat{\varphi}_m &= \mathbf{X}^{(m-1)\,T} \mathbf{y} && (34) \\ \mathbf{z}_m &\equiv \mathbf{X}^{(m-1)} \hat{\varphi}_m && (35) \\ \mathbf{X}^{(m)} &= \mathbf{X}^{(m-1)} - \frac{\mathbf{z}_m \mathbf{z}_m^T \mathbf{X}^{(m-1)}}{||\mathbf{z}_m||^2} && (36) \\ \mathbf{X}^{(0)} &\equiv \mathbf{X} && (37) \end{aligned}$$

We will prove that $\mathbf{z}_m \propto \mathbf{v}_m$, *assuming that S is positive definite.* The proof is very lengthy and it seems likely that a shorter proof should be possible...

Result A. *The sequence $\hat{\phi}_m$ contains p vectors, which are all non-zero and may or may not be unique.*
The vectors $\hat{\phi}_m$ are unit-norm hence not equal to zero. The matrix S defines a positive definite scalar product on $\mathbb{R}^p$, so the subspace:

$$\mathcal{O}_m \equiv \left\{ \alpha : \ \alpha^T S \hat{\phi}_1 = \alpha^T S \hat{\phi}_{m-1} = 0 \right\}$$

has dimension $(p - m + 1)$, which proves that the sequence stops at $m = p$.

Result B. *The vectors $\mathbf{v}_m$ are non-zero and mutually orthogonal.*
These two property follow immediately from the definition of $\mathbf{v}_m$:

$$\mathbf{v}_m^T \mathbf{v}_n \equiv \hat{\phi}_m^T S \hat{\phi}_n$$

For $m \neq n$, this product is zero because of the constraints in (32). For $m = n$, it needs to be positive since S is positive definite and $\hat{\phi}_m \neq 0$.

Result C. *Let $s \equiv \mathbf{X}^T \mathbf{y}$ and consider the sequence of vectors:*

$$s, Ss, \ldots, S^{p-1}s$$

Let $\bar{m}$ be the largest m such that S^{m-1} is linearly independent from the previous vectors $s, \ldots, S^{m-2}s$, with $\bar{m} = 0$ if $s = 0$. Then, for each $m = 1, \ldots, \bar{m}$:

a) $\hat{\phi}_m$ is unique up to a sign and proportional to the component of $S^{m-1}s$ which is S-orthogonal to $s, \ldots, S^{m-2}s$.

b) $(\hat{\phi}_m \cdot s)^2 = (\mathbf{v}_m \cdot \mathbf{y})^2 > 0$.

c) $\mathbf{v}_m$ is proportional to the component of $\mathbf{X}S^{m-1}s$ which is euclidean-orthogonal to $Span\left(\mathbf{X}s, \mathbf{X}Ss, \mathbf{X}S^{m-2}s\right) = Span\left(\mathbf{v}_1, \ldots, \mathbf{v}_{m-1}\right)$.

Each vector $\hat{\phi}_m$ must satisfy the Lagrange equation:

$$0 = \left(\hat{\phi}_m^T \mathbf{X}^T \mathbf{y}\right) \mathbf{X}^T \mathbf{y} - \lambda \hat{\phi}_m - \sum_{l=1}^{m-1} \nu_l S \hat{\phi}_l = (s \cdot \hat{\phi}_m) s - \lambda \hat{\phi}_m - \sum_{l=1}^{m-1} \nu_l S \hat{\phi}_l \tag{38}$$

for some (m-specific) values of the Lagrange multipliers $\lambda, (\nu_l)_l$. Taking the scalar product of (38) with $\hat{\phi}_m$ and using the constraints it is easy to fix the value of λ:

$$0 = (s \cdot \hat{\phi}_m) s - (s \cdot \hat{\phi}_m)^2 \hat{\phi}_m - \sum_{l=1}^{m-1} \nu_l S \hat{\phi}_l$$

This equation implies that either $\hat{\phi}_m$ is orthogonal to s, or:

$$\hat{\phi}_m \in \text{Span}\left(s, S\hat{\phi}_1, \ldots, S\hat{\phi}_{m-1}\right) \tag{39}$$

Note that if $\hat{\phi}_m$ is orthogonal to s, the maximizand in Equation (32) is zero:

$$\left(\hat{\phi}_m^T \mathbf{X}^T \mathbf{y}\right)^2 = \left(\hat{\phi}_m \cdot s\right)^2$$

Notice that if $\bar{m} = 0$ $(s = 0)$, there is nothing to prove, hence we assume $\bar{m} \geq 1$. With this in mind, we can now prove Result C by induction on $m = 1, \ldots, \bar{m}$:

- $\mathbf{m = 1}$: If $\hat{\phi}_m$ is not orthogonal to s, Equation (39) requires it to be proportional to s, hence equal to $s/||s||$ up to a sign. This vector satisfies property b), hence it is chosen by the maximization (32) over vectors orthogonal to s, which have zero value of the maximizand. This proves its uniqueness and point a). Point c) is straightforward.
- $(\mathbf{1}, \ldots, \mathbf{m}) \Rightarrow \mathbf{m + 1}$: The inductive hypothesis a) for $\hat{\phi}_1, \ldots, \hat{\phi}_m$ implies:

  $$\text{Span}\left(\hat{\phi}_1, \ldots, \hat{\phi}_m\right) = \text{Span}\left(s, \ldots, S^{m-1}s\right)$$

  which also yields:

  $$\text{Span}\left(s, S\hat{\phi}_1, \ldots, S\hat{\phi}_m\right) = \text{Span}\left(s, \ldots, S^{m-1}s, S^m s\right)$$

  Therefore, unless it is orthogonal to s, $\hat{\phi}_{m+1}$ must both belong to Span $(s, \ldots, S^m s)$ and be S-orthogonal to Span $\left(s, \ldots, S^{m-1}s\right)$. The constraint imply that this vector is unique up to a sign, and proportional to the component of $S^m s$ which is S-orthogonal to $s, \ldots, S^{m-1}s$. Let's denote this candidate vector by $\phi_{m+1,c}$. Note that $\phi_{m+1,c} \neq 0$ since $m + 1 < \bar{m}$ and $S^m s$ is linearly independent from $s, \ldots, S^{m-1}s$. To prove that $\hat{\phi}_{m+1} = \phi_{m+1,c}$ it suffices to show that $\phi_{m+1,c}$ satisfies b), so that it is chosen by the maximization (32) over any vector orthogonal to s. Note that, by construction, $\phi_{m+1,c}$ satisfies:

  $$\phi_{m+1,c}^T Ss = \ldots = \phi_{m+1,c}^T S^m s = 0$$

  i.e.:

  $$\phi_{m+1,c} \perp Ss, \ldots, S^m s$$

  where the orthogonality symbol refers to the euclidean scalar product. We see that $\phi_{m+1,c}$ cannot be orthogonal to s too, because in that case it could not belong to:

  $$\text{Span}\left(s, \ldots, S^m s\right)$$

  Therefore, $\phi_{m+1,c} \cdot s \neq 0$ and the maximizand takes a strictly positive value:

  $$\left(\phi_{m+1,c} \cdot s\right)^2 > 0$$

  This proves points a) and b). Point c) follows mechanically from the inductive hypothesis.

We are now left with the description of the $\hat{\phi}_m$ sequence for $\bar{m} < m \leq p$ (if any).

Result D. *When $\bar{m} < p$, the subspace:*

$$\mathcal{O}_{\bar{m}+1} \equiv \left\{\alpha : \ \alpha^T S \hat{\phi}_1 = \ldots = \alpha^T S \hat{\phi}_{\bar{m}} = 0\right\}$$

is a $(p - \bar{m})$-dimensional subspace of:

$$\left\{\alpha : \ \alpha \cdot s = \alpha^T \mathbf{X}^T \mathbf{y} = 0\right\}$$

Therefore, for $\bar{m} < m \leq p$ the maximization criterion (32) is void, and the vectors $\hat{\phi}_m$ can be chosen as arbitrary unit-norm, mutually S-orthogonal vectors in $\mathcal{O}_{\bar{m}+1}$. The vectors $\mathbf{v}_m$ are orthogonal to $\mathbf{y}$ and will not appear in the regression, so the sequence can be thought of as terminating at $m = \bar{m}$.
We have already established that:

$$\text{Span}\left(\hat{\phi}_1, \ldots, \hat{\phi}_{\bar{m}}\right) = \text{Span}\left(s, \ldots, S^{\bar{m}-1} s\right)$$

therefore:

$$\mathcal{O}_{\bar{m}+1} = \left\{\alpha : \ \alpha^T S s = \ldots = \alpha^T S^{\bar{m}} s = 0\right\} \tag{40}$$

By definition of $\bar{m}$, we know that $S^{\bar{m}} s$ can be expressed as a linear combination of $s, \ldots, S^{\bar{m}-1} s$:

$$S^{\bar{m}} s - a_1 S s - \ldots - a_{\bar{m}-1} S^{\bar{m}-1} s = a_0 s \tag{41}$$

Note that $a_0 \neq 0$, otherwise one could multiply (41) by S^{-1} and write $S^{\bar{m}-1}$ as a linear combination of the $s, \ldots, S^{\bar{m}-2} s$, which violates the definition of $\bar{m}$. If α belongs to $\mathcal{O}_{\bar{m}+1}$, the scalar product of the lhs with α is zero because of (40), therefore $\alpha \cdot s = 0$.

This completes the characterization of the sequence generated by (32). Now we move to the one generated by Algorithm 3.3.

Result E. *The sequence $\mathbf{z}_m$ contains at most p vectors, after which the recurrence is undefined.*
Let $m^\star$ be the smallest $m \geq 0$ such that $\mathbf{z}_{m^\star+1} = 0$. The vectors $\mathbf{z}_1, \ldots, \mathbf{z}_{m^\star}$ are mutually orthogonal and non-zero, hence linearly independent. This implies $m^\star \leq p$, since all $\mathbf{z}_m$'s belong to the p-dimensional Span of the columns of $\mathbf{X}$. For $m > m^\star + 1$, the recursive procedure (36) becomes undefined. Therefore, the sequence effectively stops at $m = m^\star$.

Result F. For $m = 1, \ldots, \bar{m}$ one has $\mathbf{z}_m \propto \mathbf{v}_m$, i.e. Algorithm 3.3 and (32) produce equivalent sequences.
To prove this key point, we adopt two inductive hypotheses:

$$\hat{\varphi}_m \in \text{Span}\left(s, Ss, \ldots, S^{m-1} s\right) - \text{Span}\left(s, Ss, \ldots, S^{m-2} s\right) \tag{42}$$

$$\mathbf{z}_m \propto \mathbf{v}_m \tag{43}$$

The first hypothesis simply means that $\hat{\varphi}_m$ can be written as a linear combination of the $s, Ss, \ldots, S^{m-1}s$, which are linearly independent according to previous results, with a non-zero coefficient for $S^{m-1}s$. If $\bar{m} = 0$ there is nothing to prove, so we can assume $\bar{m} \geq 1$:

- $\mathbf{m = 1}$: Since $\mathbf{X}^{(0)} \equiv \mathbf{X}$ this follows mechanically:

$$\begin{aligned} \hat{\varphi}_1 &= \mathbf{X}^T\mathbf{y} \equiv s \\ \mathbf{z}_1 &= \mathbf{X}\hat{\varphi}_1 = \mathbf{X}s \propto \mathbf{X}\hat{\phi}_1 \equiv \mathbf{v}_1 \end{aligned}$$

- $(\mathbf{1}, \ldots, \mathbf{m}) \Rightarrow \mathbf{m+1}$: First, we prove that $\hat{\varphi}_{m+1}$ belongs to Span$\,(s, \ldots, S^m s)$. Recalling the definition:

$$\hat{\varphi}_{m+1} \equiv \mathbf{X}^{(m)\,T}\mathbf{y}$$

we notice that the recurrence for $\mathbf{X}^{(m)}$ can be re-written as:

$$\mathbf{X}^{(m)} = \mathbf{X}^{(m-1)} - \frac{\mathbf{z}_m\mathbf{z}_m^T\mathbf{X}}{||\mathbf{z}_m||^2} \tag{44}$$

Indeed, the difference between $\mathbf{X}^{(m-1)}$ and $\mathbf{X}$ is a linear combination of the vectors $\mathbf{z}_1, \ldots, \mathbf{z}_{m-1}$, which are by construction orthogonal to $\mathbf{z}_m$. Therefore:

$$\hat{\varphi}_{m+1} = \hat{\varphi}_m - \frac{\mathbf{z}_m \cdot \mathbf{y}}{||\mathbf{z}_m||^2}\mathbf{X}^T\mathbf{z}_m$$

The first term belongs to Span$\,\big(s, \ldots, S^{m-1}s\big)$ by the inductive hypothesis. Also from the inductive hypothesis, we know that $\mathbf{z}_m \propto \mathbf{v}_m$ and we established in Result C that $\mathbf{v}_m \cdot \mathbf{y} \neq 0$. Therefore, the coefficient of the term $\mathbf{X}^T\mathbf{z}_m \propto \mathbf{X}^T\mathbf{v}_m$ is different from zero. We established in Result C that $\hat{\phi}_m$ is the component of $S^{m-1}s$ which is S-orthogonal to $s, \ldots, S^{m-2}s$, therefore it admits a linear expansion:

$$\hat{\phi}_m = a_0 s + \ldots + a_{m-1}S^{m-1}s$$

with $a_{m-1} \neq 0$. The identity $\mathbf{X}^T\mathbf{v}_m = S\hat{\phi}_m$ implies that the expansion of $\hat{\varphi}_{m+1}$ contains a non-zero term in $S^m s$ and proves the first inductive hypothesis (42) for $m+1$.

To complete the proof, note that (44) can be unwrapped as:

$$\mathbf{X}^{(m)} = \mathbf{X} - \frac{\mathbf{z}_1\mathbf{z}_1^T\mathbf{X}}{||\mathbf{z}_1||^2} - \ldots - \frac{\mathbf{z}_m\mathbf{z}_m^T\mathbf{X}}{||\mathbf{z}_m||^2}$$

Therefore:

$$\mathbf{z}_{m+1} = \mathbf{X}\hat{\varphi}_{m+1} - \frac{\mathbf{z}_1^T\mathbf{X}\hat{\varphi}_{m+1}}{||\mathbf{z}_1||^2}\mathbf{z}_1 - \ldots - \frac{\mathbf{z}_m^T\mathbf{X}\hat{\varphi}_{m+1}}{||\mathbf{z}_m||^2}\mathbf{z}_m \tag{45}$$

By the inductive hypothesis, $\mathbf{z}_l \propto \mathbf{v}_l$ for $l = 1, \ldots, m$. Therefore, remembering Result C, all the terms in the rhs of (45) except the first one belong to:

$$\text{Span}\left(\mathbf{X}s, \ldots, \mathbf{X}S^{m-1}s\right)$$

What about the first term? We have just proven that $\hat{\varphi}_{m+1}$ can be written as a linear combination of $s, \ldots, S^m s$, with a non-zero coefficient for this last term. Therefore, $\mathbf{z}_{m+1}$ satisfies:

$$\mathbf{z}_{m+1} \in \text{Span}\left(\mathbf{X}s, \ldots, \mathbf{X}S^m s\right)$$

and it is non-zero, because the $\mathbf{X}S^m s$ term is linearly independent from the others ($m + 1 \leq \bar{m}$) and has non-zero coefficient. By construction, however, $\mathbf{z}_{m+1}$ is orthogonal to each $\mathbf{z}_l \propto \mathbf{v}_l, l = 1, \ldots, m$. This implies that $\mathbf{z}_m$ is proportional to the component of $\mathbf{X}S^m s$ that is orthogonal to $\mathbf{v}_1, \ldots, \mathbf{v}_m$, i.e. $\mathbf{z}_m \propto \mathbf{v}_m$.

Result G. *The two sequences can always be thought of as terminating together, $\bar{m} = m^\star$. More specifically, for $m = 1, \ldots, \bar{m}$ one has $\mathbf{z}_m \propto \mathbf{v}_m$. Also $\mathbf{z}_{\bar{m}+1} = 0$, and the following $\mathbf{z}_m$'s are undefined. On the other hand, for $\bar{m} < p$ the vectors $\mathbf{v}_{\bar{m}+1}, \ldots, \mathbf{v}_p$ can be still defined, but they are all orthogonal to $\mathbf{y}$ and if $p - \bar{m} > 1$ they are not uniquely specified by (32).*
In Result F we have proven that, for $m = 1, \ldots, \bar{m}$, one has $\mathbf{z}_m \propto \mathbf{v}_m$. Repeating the steps of the inductive proof for $m = \bar{m}$ and using the fact that $S^{\bar{m}}s$ is linearly dependent on $s, \ldots, S^{\bar{m}-1}s$, it is easy to prove that $\mathbf{z}_{\bar{m}+1}$ can be written as linear combination of $\mathbf{X}s, \ldots, \mathbf{X}S^{\bar{m}-1}s$. However, we know by construction that $\mathbf{z}_{\bar{m}+1}$ needs to be orthogonal to the subspace generated by these same vectors, hence $\mathbf{z}_{\bar{m}+1} = 0$. The remaining parts of the proof follow from the previous results.

Finally, one may wonder how generic $\bar{m} < p$ is. For fixed predictor values and as long as S is non-singular (which in particular requires $N \geq p$) the vector s can take any values:

$$s \equiv \mathbf{X}^T\mathbf{y} : \quad \mathbf{y} = \mathbf{X}S^{-1}s + \mathbf{y}_\perp$$

where $\mathbf{y}_\perp$ is any vector with $\mathbf{X}^T\mathbf{y}_\perp = 0$. Therefore, unless S has some special structure, the probability of finding:

$$S^{m-1}s - a_0 s - a_1 S s - a_{m-2}S^{m-2}s = 0$$

for $m \leq p$ should be zero, as we have p equations and $m - 1 < p$ parameters. The special structure could be:

$$S^{m-1} - a_0\mathbb{I} - a_1 S - \ldots - a_{m-2}S^{m-2} = 0$$

Using the SVD of $\mathbf{X} = \mathbf{U}\Lambda^{1/2}V^T$, this reads:

$$\Lambda^{m-1} - a_0\mathbb{I} - a_1\Lambda - \ldots - a_{m-2}\Lambda^{m-2} = 0 \qquad (46)$$

This is means that the vector of the $(m-1)$-th powers of the PCA components variances should be linearly dependent on the vectors corresponding to lower powers. Equation (46) requires the Vandermonde matrix:

$$\begin{pmatrix} 1 & \lambda_1 & \lambda_1^2 & \dots & \lambda_1^{m-1} \\ 1 & \lambda_2 & \lambda_2^2 & \dots & \lambda_2^{m-1} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & \lambda_p & \lambda_p^2 & \dots & \lambda_p^{m-1} \end{pmatrix}$$

to have rank smaller than m. The Vandermonde matrix has rank m if and only if at least m of the λ_j's are distinct. We conclude that $\bar{m}$ is generally equal to the number of distinct PCA components' variances. Even though predictors are normalised before PLS and barring special cases such as orthogonal predictors ($\bar{m} = 1$), one should expect $\bar{m} = p$.

Ex. 3.16

When $\mathbf{X}$ has orthonormal colums:

$$\begin{aligned} \hat{\beta}_j &= \mathbf{y}^T \mathbf{x}_j & (47) \\ RSS(\beta) &= ||\mathbf{y}||^2 - 2\beta \cdot \hat{\beta} + ||\beta||^2 & (48) \end{aligned}$$

Hence:

- **Best subset:** Minimizing (48) for a subset of active predictors $\mathcal{S} \subseteq \{1, \ldots, p\}$ one can easily see that the coefficient of predictor $\mathbf{x}_j$ is always $\hat{\beta}_j$ no matter which subset is being considered. Also:

  $$RSS(\hat{\beta}, \mathcal{S}) = ||\mathbf{y}||^2 - \sum_{j \in \mathcal{S}} \hat{\beta}_j^2$$

  For fixed $M = |\mathcal{S}|$, this quantity is minimized when $\mathcal{S}$ contains the predictors corresponding to the M largest $|\hat{\beta}_j|$, which proves the first formula in the table.

- **Ridge:** Using (48):

  $$RSS(\beta, \lambda) \equiv RSS(\beta) + \lambda ||\beta||^2 = ||\mathbf{y}||^2 - 2\beta \cdot \hat{\beta} + (1 + \lambda)\, ||\beta||^2$$

  The minimum of this quantity is reached for $\hat{\beta}_\lambda = \hat{\beta}/(1 + \lambda)$.

- **Lasso:** Using again (48):

  $$RSS(\beta, \lambda) \equiv RSS(\beta) + 2\lambda ||\beta||_1 = ||\mathbf{y}||^2 - 2\beta \cdot \hat{\beta} + ||\beta||^2 + 2\lambda ||\beta||_1$$

  Notice that the β-dependent part of this expression can be written as a sum of p terms, each containing only one of the β_j's. Each term can thus be minimized separately:

  $$\text{argmin}_\beta \left(-2\hat{\beta}_j \beta + \beta^2 + 2\lambda |\beta| \right)$$

  The solution is easily seen to be:

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

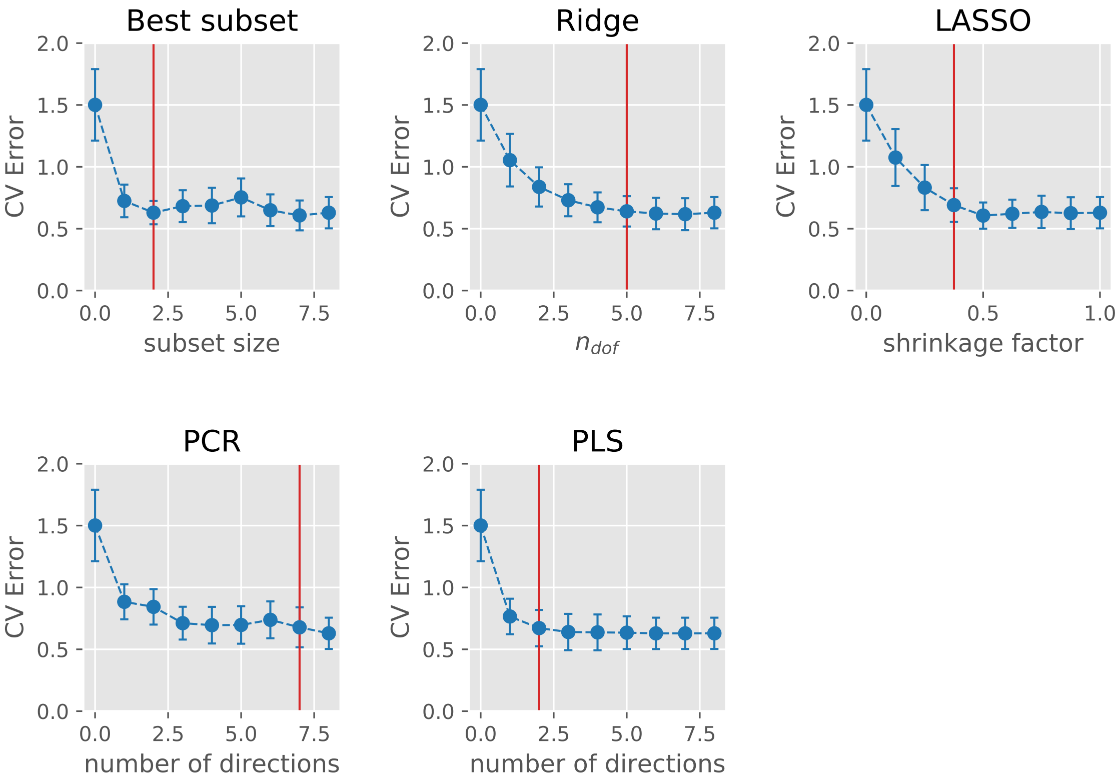
Ex. 3.17

(see Jupyter Notebook)

The notebook for this exercise reproduces the results for the `prostate` dataset before repeating the analysis on the `spam` dataset.

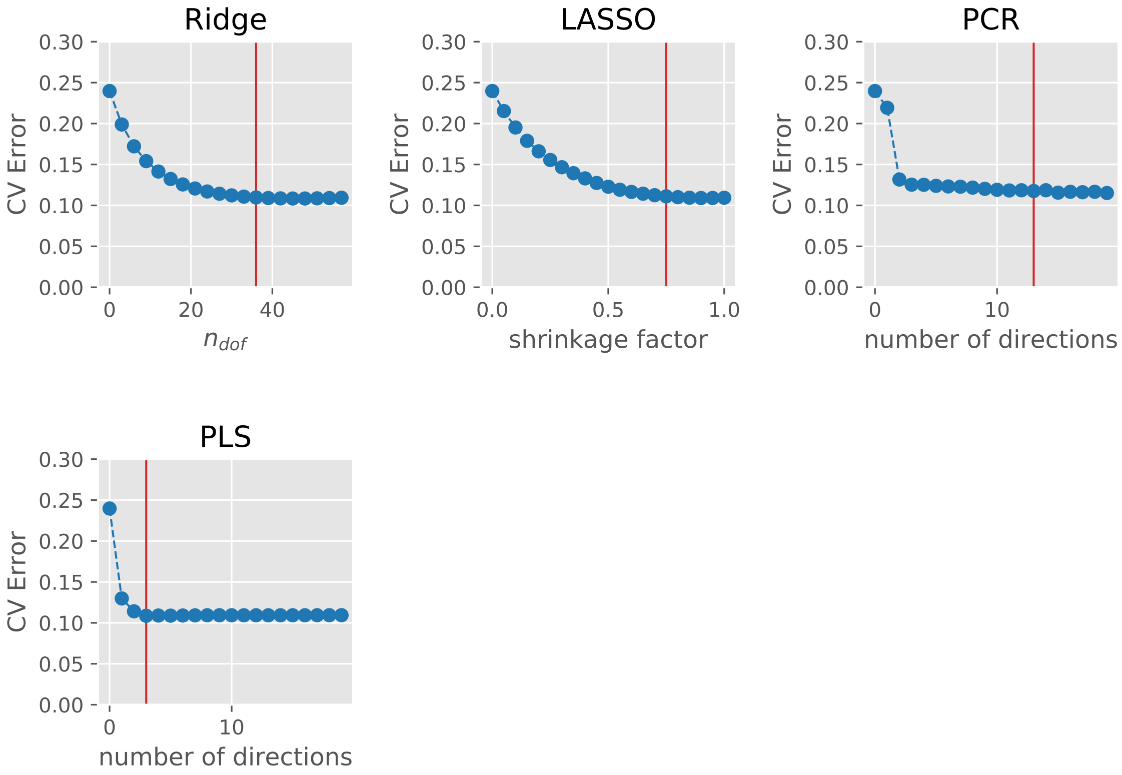
A few notes:

- I interpreted the exercise as requiring one to perform *regression* on the binary response variable.
- I found some significant deviations from the values of Table 3.3 in the text, even for simple OLS - it isn't clear to me where such deviations come from.
- The cross-validated parameter values also seem to differ, for some model families, from the ones implicitly used in Table 3.3. This is less surprising: the fold indices used by the authors are not known.
- In the `spam` case, which has 57 predictors, exhaustive best subset regression is unfeasible, so I dropped the model family from the study.
- `scikit-learn` was used where possible, only providing an implementation where the regularization was parametrized in a `scikit`-friendly way in the book.



Term	OLS	Best subset	Ridge	LASSO	PCR	PLS
Intercept	2.452	2.452	2.452	2.452	2.452	2.452
`lcavol`	0.711	0.774	0.433	0.558	0.566	0.436
`lweight`	0.290	0.349	0.252	0.187	0.321	0.360
`age`	-0.141		-0.046		-0.153	-0.021
`lbph`	0.210		0.168	0.002	0.214	0.243
`svi`	0.307		0.234	0.094	0.320	0.259
`lcp`	-0.287		0.003		-0.050	0.086
`gleason`	-0.021		0.042		0.227	0.006
`pgg45`	0.275		0.134		-0.063	0.084
Test Error	0.521	0.492	0.492	0.479	0.448	0.536
Test Error Std	0.179	0.143	0.162	0.164	0.104	0.149

Figure 3: `prostate` dataset: reproducing Figure 3.7 and Table 3.3



	OLS	Ridge	LASSO	PCR	PLS
Test Error	0.121	0.117	0.122	0.126	0.123
Test Error Std	0.007	0.005	0.007	0.005	0.008

Figure 4: `spam` dataset results.

Ex. 3.18

Review of PLS

Let's review the PLS algorithm:

$$\begin{aligned}
\mathbf{z}_m &= \mathbf{X}^{(m-1)}\mathbf{X}^{(m-1)\,T}\mathbf{y} && (49)\\
\hat{\mathbf{y}}_m &= \hat{\mathbf{y}}_{m-1} + \frac{\mathbf{z}_m\,\mathbf{z}_m^T}{||\mathbf{z}_m||^2}\,\mathbf{y} && (50)\\
& && (51)\\
\mathbf{X}^{(m)} &= \mathbf{X}^{(m-1)} - \frac{\mathbf{z}_m\,\mathbf{z}_m^T}{||\mathbf{z}_m||^2}\mathbf{X}^{(m-1)} && (52)
\end{aligned}$$

Denoting P_m the euclidean projector over the subspace generated by $\mathbf{z}_m$:

$$P_m \equiv \frac{\mathbf{z}_m\,\mathbf{z}_m^T}{||\mathbf{z}_m||^2} : \quad P_m^2 = P_m, \;\; P_m\,\mathbf{z}_m = \mathbf{z}_m$$

One has then:

$$\begin{aligned}
\hat{\mathbf{y}}_m &= (P_1 + \ldots P_m)\,\mathbf{y} && (53)\\
\mathbf{X}^{(m)} &= (1 - P_m)\,\mathbf{X}^{(m-1)} && (54)
\end{aligned}$$

$\mathbf{X}^{(m)}$ is orthogonal not only to $\mathbf{z}_m$, but also to all previous $\mathbf{z}$'s:

$$\mathbf{X}^{(m)\,T}\mathbf{z}_n \;\; \forall\, n \leq m$$

We can prove this by recursion over m. Observe that:

$$\mathbf{X}^{(m)\,T}\mathbf{z}_n = \mathbf{X}^{(m-1)\,T}\,(1 - P_m)\,\mathbf{z}_n$$

For $n = m$ the rhs vanishes because $P_m\,\mathbf{z}_m = \mathbf{z}_m$, while for $n \leq m-1$ one has:

$$P_m\,\mathbf{z_n} = \frac{\mathbf{z}_m}{||\mathbf{z}_m||^2}\mathbf{z}_m^T\mathbf{z}_n = \frac{\mathbf{z}_m}{||\mathbf{z}_m||^2}\mathbf{y}^T\mathbf{X}^{(m-1)}\mathbf{X}^{(m-1)\,T}\mathbf{z}_n = 0$$

by recursive hypothesis on the product of the last two terms.

Since $\mathbf{z}_{m+1}$ is a linear combination of the columns of $\mathbf{X}^{(m)}$, this also implies that the $\mathbf{z}$'s are mutually orthogonal:

$$\begin{aligned}
\mathbf{z}_m^T\,\mathbf{z}_n &= 0 \;\; \forall\, n \neq m\\
P_m\,P_n &= 0 \;\; \forall\, n \neq m
\end{aligned}$$

The recurrence in (54) can therefore be 'unrolled' as:

$$\mathbf{X}^{(m)} = (1 - P_1 - \ldots - P_m)\,\mathbf{X} \qquad (55)$$

Connection with CGA: notation

The connection with conjugate gradients algorithms (CGA) is at the level of the regression coefficients β. As explained in the book, all $\mathbf{X}^{(m)}$'s are linear combinations of the original $\mathbf{X}$, therefore:

$$\hat{\mathbf{y}}_m = \mathbf{X}\beta_m \tag{56}$$

for some vector β_m. If we compare with (50), we see that:

$$\hat{\mathbf{y}}_m - \hat{\mathbf{y}}_{m-1} = \mathbf{X}\left(\beta_m - \beta_{m-1}\right) \propto \mathbf{z}_m$$

In CGA, the value of a function $L(\beta)$ is minimized by successively moving β along some directions p_m:

$$\beta_m - \beta_{m-1} \propto p_{m-1}$$

Therefore, in order to establish a connection, we denote:

$$\mathbf{X}p_{m-1} \equiv \mathbf{z}_m \tag{57}$$

Notice that this equation determines p_{m-1} uniquely as long as the $\mathbf{X}$ are linearly independent.

Finally, we denote $l(\beta)$ the L2 loss and $g(\beta)$ its gradient:

$$\begin{aligned} l(\beta) &\equiv ||\mathbf{y} - \mathbf{X}\beta||^2 \\ g(\beta) &= \nabla l(\beta) = -\mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\beta\right) \end{aligned}$$

Review of CGA

The CGA algorithm tries to minimize a quadratic form:

$$f(\beta) = c^T\beta + \frac{1}{2}x^TGx, \quad x \in \mathbb{R}^p \;\; G = G^T$$

by recursively finding the optimal update of x along increasingly larger subspaces of $\mathbb{R}^p$. In practice, one starts from a set of vectors:

$$p_0, p_1, \ldots \in \mathbb{R}^p$$

and recursively defines:

$$\begin{aligned} x_{m+1} &= x_m + \text{argmin}_{w \in \mathcal{P}_m} f(x_m + w) \\ \mathcal{P}_m &\equiv \text{Span}\left(p_0, \ldots, p_m\right) \end{aligned}$$

In particular, in conjugate gradient methods one takes:

$$\mathcal{P}_m = \text{Span}\left(g(x_0), \ldots, g(x_m)\right) \tag{58}$$

$$p_0 = -g(x_0) \tag{59}$$

$$p_m = -g(x_m) + \gamma_{m-1}\,p_{m-1} \tag{60}$$

where $g(x)$ is the gradient of f, in accordance with the notation of the previous section, and the coefficients γ are chosen in such a way that the p_m are all mutually *conjugate*:

$$p_m^T G \, p_n = 0 \quad \forall \, m \neq n$$

The fact that the second term in (60) is sufficient to guarantee orthogonality is not trivial, see [2], Section 4.8.3. Thanks to conjugacy, the update $x_m \to x_{m+1}$ consists in moving x along the p_m direction only:

$$x_{m+1} = x_m + (\text{argmin}_\alpha f(x_m + \alpha \, p_m)) \; p_m \tag{61}$$

Connection between PLS and CGA

We now show that PLS is nothing but CGA on the regression coefficients applied to the L2 loss as a function of the regression coefficients β, with initial condition $\beta_0 = 0$:

$$\begin{aligned} f(\beta) &= c^T \beta + \frac{1}{2} \beta^T G \beta \\ c^T &= -\mathbf{y}^T \mathbf{X} \\ G &= \mathbf{X}^T \mathbf{X} \end{aligned}$$

The update directions are specified by (57):

$$\mathbf{X} p_m = \mathbf{z}_{m+1} = \mathbf{X}_m \mathbf{X}_m^T \mathbf{y}$$

and they are indeed mutually conjugate:

$$p_m \, G \, p_n = p_m \mathbf{X}^T \mathbf{X} \, p_n = \mathbf{z}_{m+1}^T \mathbf{z}_{n+1} = 0, \;\; m \neq n \tag{62}$$

Moreover, the coefficient of the $\mathbf{z}_m$ term at iteration m of PLS is the result of a least squares regression of $\mathbf{y}$ against $\mathbf{z}_m$ itself, which is the equivalent of (61).

To complete the connection, we show that

$$p_m = -g(\beta_m) + p_{m,||}$$

where $p_{m,||} \in \text{Span}\,(p_0, \ldots, p_{m-1})$, so that indeed $\text{Span}\,(p_0, \ldots, p_m) = \text{Span}\,(g_0, \ldots, g_m)$. We note that:

$$\begin{aligned} -\mathbf{X} g(\beta_m) &= \mathbf{X}\mathbf{X}^T \, (y - \mathbf{X}\beta_m) && (63) \\ &= \mathbf{X}\mathbf{X}^T \, (1 - P_1 - \ldots - P_m) \, \mathbf{y} && (64) \\ &= \mathbf{X} \left((1 - P_1 - \ldots P_m) \mathbf{X} \right)^T \mathbf{y} && (65) \\ &= \mathbf{X}\mathbf{X}_m^T \, \mathbf{y} && (66) \end{aligned}$$

Remember that:

$$\mathbf{X}_m = \mathbf{X} - \mathbf{z}_1 \frac{\mathbf{z}_1^T \mathbf{X}}{||\mathbf{z}_1||^2} - \ldots - \mathbf{z}_m \frac{\mathbf{z}_m^T \mathbf{X}}{||\mathbf{z}_m||^2} \tag{67}$$

Plugging in the previous equation in (66) we obtain:

$$\begin{aligned} -\mathbf{X}g(\beta_m) &= \mathbf{X}_m\mathbf{X}_m^T\,\mathbf{y} + c_1\,\mathbf{z_1} + \ldots + c_m\,\mathbf{z}_m \\ &= \mathbf{z}_{m+1} + c_1\,\mathbf{z_1} + \ldots + c_m\,\mathbf{z}_m \end{aligned}$$

or, in other terms:

$$\mathbf{X}p_m = -\mathbf{X}\,g(\beta_m) + \mathbf{X}p_{m,\|}$$

which completes our proof.

Ex. 3.19

Remember that, after predictor demeaning:

$$\beta^{\text{ridge}}(\lambda) \quad = \quad \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

Hence:

$$||\beta^{\text{ridge}}(\lambda)||^2 \quad = \quad \mathbf{y}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I}\right)^{-1}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

The following identity holds for symmetric, invertible matrices:

$$\frac{d}{d\lambda}\left(A(\lambda)\right)^{-1} = -A^{-1}\frac{dA}{d\lambda}A^{-1}$$

We get:

$$\begin{aligned}
\frac{d}{d\lambda}||\beta^{\text{ridge}}(\lambda)||^2 &= -2\,\mathbf{y}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I}\right)^{-1}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I}\right)^{-1}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I}\right)^{-1}\mathbf{X}^T\mathbf{y} \\
&= -2\,\mathbf{z}^T\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I}\right)^{-1}\mathbf{z} \\
\mathbf{z} &\equiv \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I}\right)^{-1}\mathbf{X}^T\mathbf{y}
\end{aligned}$$

The matrix $\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I}\right)^{-1}$ is positive definite, being the inverse of a symmetric, positive definite matrix. Hence:

$$\frac{d}{d\lambda}||\beta^{\text{ridge}}(\lambda)||^2 \quad = \quad -2\,\mathbf{z}^T\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I}\right)^{-1}\mathbf{z} \leq 0$$

Ex. 3.20

The canonical-correlation problem is:

$$\begin{aligned} u_m, v_m &= \mathrm{argmax}_{u,v} \mathrm{Corr}^2\left(\mathbf{Y}u, \mathbf{X}v\right) = \mathrm{argmax}_{u,v} \frac{\left(u^T\mathbf{Y}^T\mathbf{X}v\right)^2}{\left[\left(u^T\mathbf{Y}^T\mathbf{Y}u\right)\left(v^T\mathbf{X}^T\mathbf{X}v\right)\right]^{1/2}} \\ 0 &= u^T\mathbf{Y}^T\mathbf{Y}u_1 = \ldots = u^T\mathbf{Y}^T\mathbf{Y}u_{m-1} \\ 0 &= v^T\mathbf{X}^T\mathbf{X}v_1 = \ldots = v^T\mathbf{Y}^T\mathbf{Y}v_{m-1} \end{aligned}$$

The maximizand and the constraints are invariant under rescaling of u, v, so without loss of generality we can set $u^T\mathbf{Y}^T\mathbf{Y}u = v^T\mathbf{X}^T\mathbf{X}v = 1$:

$$\begin{aligned} u_m, v_m &= \mathrm{argmax}_{u,v}\left(u^T\mathbf{Y}^T\mathbf{X}v\right)^2 \\ 1 &= u^T\mathbf{Y}^T\mathbf{Y}u = v^T\mathbf{X}^T\mathbf{X}v \\ 0 &= u^T\mathbf{Y}^T\mathbf{Y}u_1 = \ldots = u^T\mathbf{Y}^T\mathbf{Y}u_{m-1} \\ 0 &= v^T\mathbf{X}^T\mathbf{X}v_1 = \ldots = v^T\mathbf{Y}^T\mathbf{Y}v_{m-1} \end{aligned}$$

Flipping the relative sign of u and v does not change the constraints but flips the sign of $u^T\mathbf{Y}^T\mathbf{X}v$, so without loss of generality we can remove the square in the first equation:

$$u_m, v_m = \mathrm{argmax}_{u,v}\left(u^T\mathbf{Y}^T\mathbf{X}v\right)$$

This completes the first part of the exercise. Now let:

$$\begin{aligned} A \equiv\ & \left(\mathbf{Y}^T\mathbf{Y}\right)^{-1/2}\left(\mathbf{Y}^T\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{X}\right)^{-1/2} = U^\star D^\star V^{\star T} \\ & U^\star \in \mathbb{R}^{K,q}, \quad V^\star \in \mathbb{R}^{p,q}, \quad q \leq \min\left(p, K\right) \\ & U^{\star T}U^\star = \mathbb{I}_q, \quad V^{\star T}V^\star = \mathbb{I}_q \\ & D^\star_{jj} > 0 \quad \forall\, j = 1, \ldots, q \end{aligned}$$

We further assume that the eigenvalues of $D^\star$ are all distinct and sorted decreasingly to simplify the discussion:

$$D^\star_{11} > D^\star_{22} > \ldots > D^\star_{qq} > 0$$

Denoting $\Sigma_X \equiv \mathbf{X}^T\mathbf{X}$, $\Sigma_Y \equiv \mathbf{Y}^T\mathbf{Y}$, we have:

$$\begin{aligned} u_m,\ v_m &= \underset{\substack{u,v \\ u^T\Sigma_X u=1, \quad v^T\Sigma_Y v=1 \\ u^T\Sigma_X u_1 = \ldots = u^T\Sigma_X u_{m-1} = 0 \\ v^T\Sigma_Y v_1 = \ldots = v^T\Sigma_Y v_{m-1} = 0}}{\mathrm{argmax}} \left(u^T\mathbf{Y}^T\mathbf{X}v\right) = \Sigma_Y^{-1/2}u^\star_m,\ \Sigma_X^{-1/2}v^\star_m \\ u^\star_m,\ v^\star_m &= \underset{\substack{u^\star, v^\star \\ \|u^\star\| = \|v^\star\| = 1 \\ u^{\star T}u^\star_1 = \ldots = u^{\star T}u^\star_{m-1} = 0 \\ v^{\star T}v^\star_1 = \ldots = v^{\star T}v^\star_{m-1} = 0}}{\mathrm{argmax}} \left(u^{\star T}U^\star D^\star V^{\star T}v^\star\right) \qquad (68) \end{aligned}$$

We can prove recursively that $u_n^\star = \pm U_n^\star$ and $v_n^\star = \pm V_n^\star$ with a common relative sign. Assuming that this is true for $n = 1, \ldots, m-1$, one can see that for any pair $(u^\star, v^\star)$ which maximizes (68), $u^\star$ must belong to the span of $U_m^\star, \ldots, U_q^\star$ and similarly for $v^\star$ with $V_m^\star \ldots, V_q^\star$. This is because, on the one hand, only the orthogonal projection of $u_m^\star$ on $U_1^\star, \ldots, U_q^\star$ contributes to the maximizand and, on the other hand, the inductive hypothesis requires $u_m^\star$ to be orthogonal to $U_1^\star, \ldots, U_{m-1}^\star$. Hence, we can write:

$$\begin{aligned} u_m^\star,\ v_m^\star &= U_{m\to q}^\star \alpha_m^\star,\ V_{m\to q}^\star \beta_m^\star \\ \alpha_m^\star,\ \beta_m^\star &= \underset{\substack{\alpha,\beta \\ ||\alpha||=||\beta||=1}}{\text{argmax}} \left(\alpha^{\star T} D_{m\to q}^\star \beta^\star\right) \end{aligned}$$

where $U_{m\to q}^\star$ is the submatrix obtained by taking columns m through q of $U^\star$, and similarly for $V^\star$. This last bit of maximization can be carried out explicitly using Lagrange multipliers, to show that $\alpha_m^\star = \beta_m^\star = (\pm 1, 0, \ldots, 0)$, i.e. $u_m^\star = \pm U_m^\star$, $v_m^\star = \pm V_m^\star$ with a common sign.

Ex. 3.21 and 3.22

The reduced-rank regression problem can be written as:

$$\hat{B}(m) = \underset{\text{rank}(B)=m}{\text{argmin}} \ \text{Tr}\left[(\mathbf{Y}-\mathbf{X}B)\,\Sigma_Y^{-1}\,(\mathbf{Y}-\mathbf{X}B)^T\right] \tag{69}$$

From now on, we only assume that Σ_Y is symmetric and positive-definite, without specifying how it is obtained from the data. We also assume $\Sigma_X \equiv \mathbf{X}^T\mathbf{X}$ to be positive-definite to simplify the discussion. One has:

$$\begin{aligned}
\hat{B}(m) &= \Sigma_X^{-1/2}\hat{B}^\star(m)\,\Sigma_Y^{1/2} \\
\hat{B}^\star(m) &= \underset{\text{rank}(B^\star)=m}{\text{argmin}} \ \text{Tr}\left[(\mathbf{Y}_r-\mathbf{X}_nB^\star)\,(\mathbf{Y}_r-\mathbf{X}_nB^\star)^T\right] \\
\mathbf{Y}_r &\equiv \mathbf{Y}\,\Sigma_Y^{-1/2} \\
\mathbf{X}_n &\equiv \mathbf{X}\,\Sigma_X^{-1/2}: \quad \mathbf{X}_n^T\mathbf{X}_n = \mathbb{I}_p
\end{aligned}$$

For any $\mathcal{V} \in \mathbb{R}^{p,m}$, $\mathcal{V}^T\mathcal{V} = \mathbb{I}_m$ and $\mathcal{L} \in \mathbb{R}^{m,K}$, the matrix:

$$B^\star = \mathcal{V}\mathcal{L} \tag{70}$$

has rank m. Conversely, any matrix of rank m can be written this way. Indeed, the rank condition is equivalent to the span of the columns of $B^\star$ having dimension m, in which case all columns of $B^\star$ can be written as linear combinations of a set of m orthonormal vectors $\mathcal{V}_1, \ldots, \mathcal{V}_m$. Notice that the decomposition (70) is not unique, since any rotation of the columns of $\mathcal{V}$ leaves $B^\star$ invariant:

$$\begin{aligned}
\mathcal{V} &\rightarrow \mathcal{V}\mathcal{R} &(71)\\
\mathcal{L} &\rightarrow \mathcal{R}^T\mathcal{L} &(72)\\
\mathcal{R}^T\mathcal{R} &= \mathcal{R}\mathcal{R}^T = \mathbb{I}_m &(73)
\end{aligned}$$

Keeping this degeneracy in mind, we can write:

$$\begin{aligned}
\hat{B}^\star(m) &= \mathcal{V}^\star(m)\mathcal{L}^\star(m) \\
\mathcal{V}^\star(m),\ \mathcal{L}^\star(m) &= \underset{\substack{\mathcal{V}^\star,\mathcal{L}^\star \\ \mathcal{V}^{\star T}\mathcal{V}^\star=\mathbb{I}_m}}{\text{argmin}} \ \text{Tr}\left[(\mathbf{Y}_r-\mathbf{X}_n\mathcal{V}^\star\mathcal{L}^\star)\,(\mathbf{Y}_r-\mathbf{X}_n\mathcal{V}^\star\mathcal{L}^\star)^T\right]
\end{aligned}$$

In the absence of constraints, the minimization over $\mathcal{L}^\star$ is a simple OLS procedure, which gives:

$$\mathcal{L}^\star(m) = \mathcal{V}^{\star T}(m)\mathbf{X}_n^T\mathbf{Y}_r$$

One gets:

$$\begin{aligned}
\mathcal{V}^\star(m) &= \underset{\mathcal{V}^\star:\ \mathcal{V}^{\star T}\mathcal{V}^\star\,=\,\mathbb{I}_m}{\text{argmax}} \ \text{Tr}\left(\mathcal{V}^\star\mathcal{V}^{\star T}\mathbf{X}_n^T\mathbf{Y}_r\mathbf{Y}_r^T\mathbf{X}_n\right) \\
&= \underset{\mathcal{V}^\star:\ \mathcal{V}^{\star T}\mathcal{V}^\star\,=\,\mathbb{I}_m}{\text{argmax}} \ \text{Tr}\left(\mathcal{V}^\star\mathcal{V}^{\star T}V^\star D^{\star 2}V^{\star T}\right) &(74)
\end{aligned}$$

The last equality is obtained by plugging in the the generalized SVD:

$$\Sigma_Y^{-1/2}(\mathbf{Y}^T\mathbf{X})\Sigma_X^{-1/2} = \mathbf{Y}_r^T\mathbf{X}_n = U^\star D^\star V^{\star T} \tag{75}$$

The maximization of (74) can be carried out using the lagrangian:

$$\mathrm{Tr}\left(\mathcal{V}^\star\mathcal{V}^{\star T}V^\star D^{\star 2}V^{\star T}\right) + \mathrm{Tr}\left(\Lambda\mathcal{V}^{\star T}\mathcal{V}\right)$$

The lagrange equations are equivalent to:

$$\left(\mathbb{I} - \mathcal{V}^\star(m)\mathcal{V}^{\star T}(m)\right) V^\star D^{\star 2}V^{\star T}\mathcal{V}^\star(m) = 0$$

This implies that the subspace generated by the columns of $\mathcal{V}^\star(m)$ is stable under the action of the self-adjoint operator $V^\star D^{\star 2}V^{\star T}$. This implies that the m-dimensional span of the columns of $\mathcal{V}^\star(m)$ is generated my m eigenvectors of $V^\star D^{\star 2}V^{\star T}$, i.e. by m columns of $V^\star_{c_1}, \ldots, V^\star_{c_m}$ of $V^\star$. Without loss of generality, we can take $\mathcal{V}^\star(m)$ to be the matrix with columns $V^\star_{c_1}, \ldots, V^\star_{c_m}$. The corresponding value of the maximizand is:

$$\mathrm{Tr}\Big[\mathcal{V}^\star(m)\mathcal{V}^{\star T}(m)V^\star D^{\star 2}V^{\star T}\Big] = \sum_{i=1}^{m} D^2_{c_i,c_i}$$

This quantity is clearly maximized by choosing $\{c_i\}_{i=1,\ldots,m} = \{1, \ldots, m\}$. Hence, without loss of generality:

$$\mathcal{V}^\star(m) = V^\star_{(m)} \tag{76}$$

Putting everything together:

$$\begin{aligned}
\hat{B}(m) &= \Sigma_X^{-1/2}V^\star_{(m)}V^{\star T}_{(m)}V^\star D^\star U^{\star T}\Sigma_Y^{1/2} && (77)\\
&= \Sigma_X^{-1/2}V^\star_{(m)}D^\star_{(m)}U^{\star T}_{(m)}\Sigma_Y^{1/2} && (78)\\
&= \Sigma_X^{-1/2}V^\star D^\star U^{\star T}U^\star_{(m)}U^{\star T}_{(m)}\Sigma_Y^{1/2} && (79)
\end{aligned}$$

Notice how (77) and (79) can be re-written as:

$$\begin{aligned}
\hat{B}(m) &= \Sigma_X^{-1/2}V^\star_{(m)}V^{\star T}_{(m)}\Sigma_X^{-1/2}\mathbf{X}^T\mathbf{Y} && (80)\\
&= \Sigma_X^{-1}\mathbf{X}^T\mathbf{Y}\,\Sigma_Y^{-1/2}U^\star_{(m)}U^{\star T}_{(m)}\Sigma_Y^{1/2} && (81)
\end{aligned}$$

The second equation makes the connection with the OLS coefficients $\hat{B}$:

$$\hat{B}(m) = \hat{B}\,\Sigma_Y^{-1/2}U^\star_{(m)}U^{\star T}_{(m)}\Sigma_Y^{1/2} \tag{82}$$

Now:

- For $\Sigma_Y = \mathbf{Y}^T\mathbf{Y}$, the matrix $U^\star$ is precisely the one used in CCA, so $\Sigma_Y^{-1/2}U^\star_{(m)}$ is the matrix $U_{(m)}$ of the top m left CCA vectors and $\Sigma_Y^{1/2}U^\star_{(m)}$ is its pseudo-inverse $U^-_{(m)}$:

$$U^T_{(m)}U^-_{(m)} = \mathbb{I}_m$$

- Equation (82) shows that $\hat{B}(m)$ depends on Σ_Y both explicitly and implicitly via the dependence of $U^\star$ through the generalized SVD decomposition (75). Equation (80) shows that this dependence is fully encoded in the dependence of $V^\star$ on Σ_Y. As a side note, notice that multiplying Σ_Y by a constant does not change $U^\star$ or $V^\star$, but merely rescales $D^\star$. Therefore, (80) shows that $\hat{B}(m)$ is invariant under such rescaling.

We now show that $\hat{B}(m)$ has the same value for $\Sigma_Y = \Sigma_0 \equiv \mathbf{Y}^T\mathbf{Y}$ and $\Sigma_Y = \Sigma_{ols} \equiv \left(\mathbf{Y} - \mathbf{X}\hat{B}\right)^T \left(\mathbf{Y} - \mathbf{X}\hat{B}\right)$, by showing that $V^\star$ is the same for these two choices. Let's fix the notation for the two SVD decompositions:

$$\begin{align} (\mathbf{Y}^T\mathbf{X})\Sigma_X^{-1/2} &= \Sigma_0^{1/2}\, U_0^\star D_0^\star\, V_0^{\star T} \tag{83}\\ &= \Sigma_{ols}^{1/2}\, U_{ols}^\star\, D_{ols}^\star\, V_{ols}^{\star T} \tag{84}\end{align}$$

Using the expression of $\hat{B}$ and the first of these two SVD decompositions, one can easily derive:

$$\Sigma_{ols} = \Sigma_0^{1/2}\, \left(\mathbb{I} - U_0^\star\, D_0^{\star 2}\, U_0^{\star T}\right) \Sigma_0^{1/2} \tag{85}$$

Now consider the matrix:

$$A = \Sigma_{ols}^{-1/2}\Sigma_0^{1/2}U_0^\star$$

Using (85), one can see that:

$$\begin{aligned} A^T A &= U_0^{\star T}\, \left(\mathbb{I}_K - U_0^\star D_0^{\star 2} U_0^{\star T}\right)^{-1} U_0^\star \\ &= \left(\mathbb{I}_q - D_0^{\star 2}\right)^{-1}\end{aligned}$$

Therefore, the matrix:

$$U_{cand}^\star = A\, \left(\mathbb{I}_q - D_0^{\star 2}\right)^{1/2} = \Sigma_{ols}^{-1/2}\Sigma_0^{1/2}U_0^\star \left(\mathbb{I}_q - D_0^{\star 2}\right)^{1/2}$$

is orthogonal and:

$$\begin{aligned} (\mathbf{Y}^T\mathbf{X})\Sigma_X^{-1/2} &= \Sigma_0^{1/2}\, U_0^\star D_0^\star\, V_0^{\star T} \\ &= \Sigma_{ols}^{1/2} U_{cand}^\star \left(\mathbb{I}_q - D_0^{\star 2}\right)^{-1/2} D_0^\star\, V_0^{\star T}\end{aligned}$$

Comparing this expression with (84), we conclude:

$$\begin{aligned} U_{ols}^\star &= U_{cand}^\star \\ D_{ols}^\star &= \left(\mathbb{I}_q - D_0^{\star 2}\right)^{-1/2} D_0^\star \\ V_{ols}^\star &= V_0^\star\end{aligned}$$

Equation (80) lets us immediately deduce $\hat{B}_{ols}(m) = \hat{B}_0(m)$.

Ex. 3.23

The initial assumption can be re-written as:

$$\frac{1}{N}\left|\mathbf{X}^T\mathbf{y}\right| = \lambda\, e_p \tag{86}$$

Denote $\hat{\mathbf{y}} \equiv X\hat{\beta}$. In the following, we will make use of the following identity:

$$\mathbf{X}^T\left(\mathbf{y}-\hat{\mathbf{y}}\right) = 0 \tag{87}$$

which tells us that the OLS estimate $\hat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto the subspace generated by the columns of $\mathbf{X}$. (87) also implies:

$$\hat{\mathbf{y}}^T\left(\mathbf{y}-\hat{\mathbf{y}}\right) = 0 \tag{88}$$

Using (87):

$$\frac{1}{N}\left|\mathbf{X}^T\left(\mathbf{y}-\alpha\hat{\mathbf{y}}\right)\right| = \frac{1}{N}\left|(1-\alpha)\mathbf{X}^T\mathbf{y}\right| = (1-\alpha)\lambda\, e_p$$

which proves (a). Moreover:

$$\begin{aligned} RSS &\equiv ||\mathbf{y}-\hat{\mathbf{y}}||^2 \\ &= ||\mathbf{y}||^2 - 2\mathbf{y}^T\hat{\mathbf{y}} + ||\hat{\mathbf{y}}||^2 \\ &= N - \mathbf{y}^T\hat{\mathbf{y}} \end{aligned}$$

In the last step, we made use of (88). Hence:

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

This result can be combined with the expression (89) of the absolute covariance to give (b). As for (c), LAR segments are indeed of the form $\mathbf{y} \to \mathbf{y} - \alpha\hat{\mathbf{y}}$ where $\mathbf{y} = \mathbf{r}_k$ are the residuals at beginning of the segment, and the regression is performed w.r.t. the set of active predictors, which all have the same absolute covariance with $\mathbf{r}_k$ just like in the exercise.

Ex. 3.24

By definition, all predictors in LAR are normalised and all predictors that are active at step k have the same covariance with the current residuals:

$$\frac{1}{N}|\mathbf{X}_{\mathcal{A}_k}^T \mathbf{r}_k| = \lambda\, e$$

The cosine of the angles between the LAR direction $\mathbf{u_k}$ and the active predictors is proportional to their scalar products:

$$\mathbf{X}_{\mathcal{A}_k}^T \mathbf{u}_k = \mathbf{X}_{\mathcal{A}_k}^T \mathbf{X}_{\mathcal{A}_k} \delta_k = \mathbf{X}_{\mathcal{A}_k}^T \mathbf{X}_{\mathcal{A}_k} (\mathbf{X}_{\mathcal{A}_k}^T \mathbf{X}_{\mathcal{A}_k})^{-1} \mathbf{X}_{\mathcal{A}_k}^T \mathbf{r}_k = \mathbf{X}_{\mathcal{A}_k}^T \mathbf{r}_k = N\lambda\, e.$$

Ex. 3.25

Denote λ the scalar product between the active set predictors and the residuals at the beginning of step k:

$$\frac{1}{N}|\mathbf{X}_{\mathcal{A}_k}^T \mathbf{r}_k| = \lambda\, e$$

This quantity decreases linearly along the LAR segment (see Ex. 3.23):

$$\begin{aligned} \mathbf{r}_k &\to \mathbf{r}_k - \alpha \mathbf{X}_{\mathcal{A}_k} \delta_k \\ \frac{1}{N}|\mathbf{X}_{\mathcal{A}_k}^T \mathbf{r}_k| &\to (1-\alpha)\lambda\, e \end{aligned}$$

Denote $\mathcal{I}_k$ the set of predictors that are *inactive* at step k. The LAR segment ends when the absolute correlation of one of these predictors with the residuals matches that of the active predictors. The norm of the residuals is immaterial because it affects all correlations as a common multiplicative factor, so we can just compute:

$$\begin{aligned} \frac{1}{N}|\mathbf{X}_{\mathcal{I}_k}^T \mathbf{r}_k| &\to \frac{1}{N}|\mathbf{X}_{\mathcal{I}_k}^T (\mathbf{r}_k - \alpha \mathbf{X}_{\mathcal{A}_k} \delta_k)| \\ &= |\lambda_{\mathcal{I},k} - \alpha \lambda'_{\mathcal{I},k}| \\ \lambda_{\mathcal{I},k} &= \frac{1}{N}\mathbf{X}_{\mathcal{I}_k}^T \mathbf{r}_k \\ \lambda'_{\mathcal{I},k} &= \frac{1}{N}\mathbf{X}_{\mathcal{I}_k}^T \mathbf{X}_{\mathcal{A}_k} \delta_k \end{aligned}$$

The crossing points $\alpha^\star$ are therefore determined by:

$$|\lambda_{\mathcal{I},k} - \alpha^\star \lambda'_{\mathcal{I},k}| = (1-\alpha^\star)\lambda\, e$$

and the LAR segments ends at the smallest of the values of $\alpha^\star$.

Ex. 3.26

Denote $\mathbf{x}_{j,\mathcal{A}}$ the residual of $\mathbf{x}_j$ against the active predictors $\mathbf{X}_{\mathcal{A}}$:

$$\begin{aligned}
\mathbf{x}_j &= \mathbf{x}_{j,\mathcal{A}} + \mathbf{X}_{\mathcal{A}}\gamma_j \\
\mathbf{X}_{\mathcal{A}}^T\mathbf{x}_{j,\mathcal{A}} &= 0
\end{aligned}$$

Using the last relationship, the RSS with j included simplifies to:

$$\begin{aligned}
||\mathbf{y} - \mathbf{X}_{\mathcal{A}}\beta_{\mathcal{A}} - \mathbf{x}_j\beta_j||^2 &= ||\mathbf{y} - \mathbf{X}_{\mathcal{A}}\beta'_{\mathcal{A}}||^2 + ||\mathbf{y} - \mathbf{x}_{j,\mathcal{A}}\beta_j||^2 - ||\mathbf{y}||^2 \\
\beta'_{\mathcal{A}} &\equiv \beta_{\mathcal{A}} + \beta_j\gamma_j
\end{aligned}$$

This shows that the OLS minimization can be carried over on $\beta'_{\mathcal{A}}$ and β_j independently. Therefore, the OLS coefficient $\hat{\beta}_j$ is the single variable regression coefficient of $\mathbf{y}$ against $\mathbf{x}_{j,\mathcal{A}}$, which had already been shown in Section 3.2.3 of the text. At the OLS value:

$$||\mathbf{y} - \mathbf{x}_{j,\mathcal{A}}\hat{\beta}_j||^2 - ||\mathbf{y}||^2 = ||\mathbf{y}||^2\left(1 - \rho_{j,\mathcal{A}}^2\right)$$

where $\rho_{j,\mathcal{A}}$ is the correlation between $\mathbf{y}$ and $\mathbf{x}_{j,\mathcal{A}}$. Therefore, in forward stepwise regression, the largest decrease in RSS is provided by the predictor whose residual w.r.t. the active predictors is mostly correlated to $\mathbf{y}$ in absolute value, as pointed out by the hint.

In a way, LAR residualises the response against the active predictors before computing correlations, while forward stepwise residualises the candidate predictors themselves. These two operations result in potentially very different correlations. Denoting $\mathbf{P}_{\mathcal{A}} \equiv \mathbf{X}_{\mathbf{A}}\left(\mathbf{X}_{\mathbf{A}}^T\mathbf{X}_{\mathbf{A}}\right)^{-1}\mathbf{X}_{\mathcal{A}}$ the projector onto the subspace generated by the columns of $\mathbf{X}_{\mathcal{A}}$, one has:

$$\begin{aligned}
\text{LAR}: &\quad \text{Corr}\left(\mathbf{x}_j, \mathbf{y}_{\mathcal{A}}\right) = \frac{\mathbf{x}_j^T P_{\mathcal{A}}\mathbf{y}}{||\mathbf{x}_j||\,||\mathbf{y}_{\mathcal{A}}||} = \frac{\mathbf{y}^T P_{\mathcal{A}}\mathbf{x}_j}{||\mathbf{x}_j||\,||\mathbf{y}_{\mathcal{A}}||} = \frac{\mathbf{y}^T\mathbf{x}_{j,\mathcal{A}}}{||\mathbf{x}_j||\,||\mathbf{y}_{\mathcal{A}}||} \\
\text{Forward stepwise}: &\quad \text{Corr}\left(\mathbf{x}_{j,\mathcal{A}}, \mathbf{y}\right) = \frac{\mathbf{y}^T\mathbf{x}_{j,\mathcal{A}}}{||\mathbf{x}_{j,\mathcal{A}}||\,||\mathbf{y}||}
\end{aligned}$$

While the numerator is the same, the denominator is different between the two expressions. In particular, the denominator in the second expression tends to more aggressively 'squeeze' out of all predictors their idiosyncratic part that is correlated to the response. Consider the following example:

$$\begin{aligned}
\mathbf{y} &= \mathbf{X}_{\mathcal{A}}\beta_{\mathcal{A}} + \eta\,\mathbf{v} + \boldsymbol{\epsilon}_1 \\
\mathbf{x}_1 &= \mathbf{X}_{\mathcal{A}}\gamma_{\mathcal{A}} + \mathbf{v} + \boldsymbol{\epsilon}_2 \\
\mathbf{x}_2 &= \mathbf{v} + \sqrt{1 + ||\mathbf{X}_{\mathcal{A}}\gamma_{\mathcal{A}}||^2}\,\boldsymbol{\epsilon}_3 \\
\boldsymbol{\epsilon}_i^T\mathbf{X}_{\mathcal{A}} &= \boldsymbol{\epsilon}_i^T\mathbf{v} = 0 \\
||\boldsymbol{\epsilon}_1||^2 &= ||\boldsymbol{\epsilon}_2||^2
\end{aligned}$$

The two predictors have the same norm and their residuals have the same covariance with $\mathbf{y}$:

$$\begin{aligned}\mathbf{x}_{1,\mathcal{A}} &= \mathbf{v} + \boldsymbol{\epsilon}_1 \\ \mathbf{x}_{2,\mathcal{A}} &= \mathbf{x}_2 = \mathbf{v} + \sqrt{1 + ||\mathbf{X}_{\mathcal{A}}\gamma_{\mathcal{A}}||^2}\,\boldsymbol{\epsilon}_2\end{aligned}$$

hence they behave similarly w.r.t. LAR. On the other hand:

$$||\mathbf{x}_{1,\mathcal{A}}||^2 = ||\mathbf{v}||^2 + ||\boldsymbol{\epsilon}_1||^2 < ||\mathbf{x}_{2,\mathcal{A}}||^2 = ||\mathbf{v}||^2 + \left(1 + ||\mathbf{X}_{\mathcal{A}}\gamma_{\mathcal{A}}||^2\right)||\boldsymbol{\epsilon}_2||^2$$

Ex. 3.27

The KKT conditions apply to solutions of convex optimization problems[4]:

$$\underset{\substack{\beta \\ f_i(\beta)=0,\, i=1,\ldots,M \\ g_a(\beta)\leq 0,\, a=1,\ldots,N}}{\operatorname{argmax}} \qquad L(\beta) \tag{89}$$

The Lagrange dual function is:

$$L(\beta,\eta,\lambda) \equiv L(\beta) + \sum_i \eta_i\, f_i(\beta) + \sum_a \lambda_a\, g_a(\beta)$$

and the KKT conditions read:

$$\begin{aligned} g_a(\beta) &\leq 0 && (90) \\ \lambda_a &\geq 0 && (91) \\ \lambda_a\, g_a(\beta) &= 0 && (92) \\ \frac{\partial L}{\partial \beta}(\beta,\eta,\lambda) &= 0 && (93) \end{aligned}$$

The LASSO regression problem reads:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(L(\beta) + \lambda \sum_a |\beta_a| \right)$$

where $L(\beta) \equiv \frac{1}{2}||\mathbf{y} - \mathbf{X}\beta||^2$. Now consider the following:

$$\hat{\beta}^+,\, \hat{\beta}^- \equiv \underset{\beta_a^+\geq 0,\, \beta_a^-\geq 0}{\operatorname{argmin}} \left(L(\beta^+ - \beta^-) + \lambda \sum_a (\beta_a^+ + \beta_a^-) \right) \tag{94}$$

It is fairly easy to show that:

- For each a, either $\hat{\beta}_a^+ = 0$ or $\hat{\beta}_a^- = 0$, so that $\hat{\beta}_a^+ + \hat{\beta}_a^- = |\hat{\beta}_a^+ - \hat{\beta}_a^-|$.
- $\hat{\beta} = \hat{\beta}^+ - \hat{\beta}^-$.

For the first point, notice that subtracting from both $\hat{\beta}_a^+$ and $\hat{\beta}_a^-$ the smallest number between the two does not change the first term in the minimizand (94), but decreases the second. For the second, it is sufficient to notice that any β can be written as a difference between a positive and negative part.

The minimization problem (94) is of the general form (89)[5]. The dual lagrangian reads like in the text:

$$L(\beta,\lambda^+,\lambda^-) = L(\beta^+ - \beta^-) + \lambda \sum_a (\beta_a^+ + \beta_a^-) - \sum_a \left(\lambda_a^+ \beta_a^+ + \lambda_a^- \beta_a^- \right)$$

[4] See e.g. `https:ocw.mit.edu/courses/sloan-school-of-management15097prediction-machine-learningandstatisticsspring2012/lecture-notesMIT15_097S12_lec11.pdf`

[5] Quadratic L with affine inequality constraints is one of the cases in which strong duality is guaranteed, see e.g. [1], Chapter 5.

The KKT conditions can be derived in a straightforward manner from (90) - (93):

$$\begin{aligned} \beta_a^{\pm} &\geq 0 && (95) \\ \lambda_a^{\pm} &\geq 0 && (96) \\ \lambda_a^{\pm}\,\beta_a^{\pm} &= 0 && (97) \\ \frac{\partial}{\partial\beta_a} L(\beta^+ - \beta^-) &= \mp(\lambda - \lambda_a^{\pm}) && (98) \end{aligned}$$

which proves (a). From the two versions last equation we have:

$$(+):\quad \frac{\partial L}{\partial\beta_a} = -\lambda + \lambda_a^+ \geq -\lambda \tag{99}$$

$$(-):\quad \frac{\partial L}{\partial\beta_a} = \lambda - \lambda_a^- \leq \lambda \tag{100}$$

hence $|\frac{\partial L}{\partial\beta_a}| \leq \lambda$, which automatically proves the first case in the text ($\lambda = 0$). Supposing now that $\lambda > 0$ and $\beta_a^+ > 0$, (97) implies $\lambda_a^+ = 0$. From (99) we deduce $\frac{\partial L}{\partial\beta_a} = -\lambda < 0$, hence $\lambda_a^- > 0$ from (100) and $\beta_a^- = 0$ from (97). The $\beta_a^- > 0$ case can be dealt with similarly. Therefore, if $\beta_a \neq 0$:

$$\frac{\partial L}{\partial\beta_a} = -\mathbf{x}_a^T\,(\mathbf{y} - \mathbf{X}\beta) = -\text{sign}(\beta_a)\,\lambda \tag{101}$$

If the predictors and the response are standardised, this implies:

$$\text{Corr}(\mathbf{x}_a, \mathbf{y} - \mathbf{X}\beta) = \text{sign}(\beta_a)\frac{\lambda}{N}$$

which completes part (b). As for part (c), as long as the set of active predictors does not change, the set of one can pack all the corresponding equations (101) as:

$$\mathbf{X}_{\mathcal{A}}^T\mathbf{y} - \mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}}\beta_{\mathcal{A}} = s\,\lambda$$

where s is the vector of $\text{sign}(\beta_a),\ a \in \mathcal{A}$. Hence:

$$\beta_{\mathcal{A}}(\lambda) = \left(\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}}\right)^{-1}\left(\mathbf{X}_{\mathcal{A}}^T\mathbf{y} - s\lambda\right)$$

which is indeed linear in λ. The derivative of $\beta_{\mathcal{A}}$ with respect to λ can be shown to be proportional to the OLS coefficients of the regression of the residuals $\mathbf{r}(\lambda_1) \equiv \mathbf{y} - \mathbf{X}_{\mathcal{A}}\beta(\lambda_1)$ against the active predictors, in accordance with the LAR/LASSO correspondence established in the text[6]:

$$\begin{aligned} (101):\quad & \mathbf{X}_{\mathcal{A}}^T\,\mathbf{r}(\lambda_1) = s\,\lambda_1 \\ & -\beta_{\mathcal{A}}'(\lambda) = \left(\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}}\right)^{-1} s \propto \left(\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}}\right)^{-1}\mathbf{X}_{\mathcal{A}}^T\,\mathbf{r}(\lambda_1) \end{aligned}$$

[6] Notice the sign difference: LAR moves in the direction of increasing coefficients magnitude, which is the opposite as increasing λ

Ex. 3.28

Denote $\mathbf{X}_{\setminus j}$ the matrix of predictor values except predictor $\mathbf{x}_j$. The modified LASSO minimization problem with the added predictor copy reads:

$$B^\star : \qquad \underset{\substack{\beta_{\setminus j}, \beta_j, \beta_j^\star \\ ||\beta_{\setminus j}||_1 + |\beta_j| + |\beta_j^\star| \leq t}}{\text{argmin}} \quad l(\beta_{\setminus j}, \beta_j + \beta_j^\star) \tag{102}$$

$$l(\beta_{\setminus j}, \beta_j) \equiv ||\mathbf{y} - \mathbf{X}_{\setminus j}\beta_{\setminus j} - \beta_j \mathbf{x}_j||^2 \tag{103}$$

Denote $B^\star$ the set of solutions $(\beta_{\setminus j}, \beta_j, \beta_j^\star)$ of (102), and B the set of solutions $(\beta_{\setminus j}, \beta_j)$ to the ordinary LASSO problem with the same parameter t:

$$B : \qquad \underset{\substack{\beta_{\setminus j}, \beta_j \\ ||\beta_{\setminus j}||_1 + |\beta_j| \leq t}}{\text{argmin}} \quad l(\beta_{\setminus j}, \beta_j) \tag{104}$$

First, we note that:

$$\min_{\substack{\beta_{\setminus j}, \beta_j, \beta_j^\star \\ ||\beta_{\setminus j}||_1 + |\beta_j| + |\beta_j^\star| \leq t}} l(\beta_{\setminus j}, \beta_j + \beta_j^\star) \geq \min_{\substack{\beta_{\setminus j}, \beta_j \\ ||\beta_{\setminus j}||_1 + |\beta_j| \leq t}} l(\beta_{\setminus j}, \beta_j) \tag{105}$$

Indeed, one can trivially re-express the r.h.s. as:

$$\min_{\substack{\beta_{\setminus j}, \beta_j \\ ||\beta_{\setminus j}||_1 + |\beta_j| \leq t}} l(\beta_{\setminus j}, \beta_j) = \min_{\substack{\beta_{\setminus j}, \beta_j \\ ||\beta_{\setminus j}||_1 + |\beta_j + \beta_j^\star| \leq t}} l(\beta_{\setminus j}, \beta_j + \beta_j^\star) \tag{106}$$

Since $|\beta_j + \beta_j^\star| \leq |\beta_j| + |\beta_j^\star|$, the modified LASSO problem on the left hand-side of (105) is just a more constrained version of the original one. Also, since $\beta = (\beta_{\setminus j}, \beta_j, \beta_j^\star = 0)$ respects the modified LASSO constraint whenever $(\beta_{\setminus j}, \beta_j)$ respects the original one, we can conclude that:

$$\min_{\substack{\beta_{\setminus j}, \beta_j, \beta_j^\star \\ ||\beta_{\setminus j}||_1 + |\beta_j| + |\beta_j^\star| \leq t}} l(\beta_{\setminus j}, \beta_j + \beta_j^\star) = \min_{\substack{\beta_{\setminus j}, \beta_j \\ ||\beta_{\setminus j}||_1 + |\beta_j| \leq t}} l(\beta_{\setminus j}, \beta_j) \tag{107}$$

Using this result:

- If $\beta = (\beta_{\setminus j}, \beta_j, \beta_j^\star)$ is such that $(\beta_{\setminus j}, \beta_j + \beta_j^\star) \in B$ and β respects the modified LASSO constraint, then necessarily $\beta \in B^\star$, otherwise the l.h.s. in (107) would be smaller than the r.h.s..
- Reasoning in the same way, it is easy to see that if $\beta = (\beta_{\setminus j}, \beta_j, \beta_j^\star) \in B^\star$, then $(\beta_{\setminus j}, \beta_j + \beta_j^\star) \in B$.

This allows us to conclude that:

$$B^\star = \left\{\beta = (\beta_{\setminus j}, \beta_j, \beta_j^\star) : \ (\beta_{\setminus j}, \beta_j + \beta_j^\star) \in B, \ ||\beta_{\setminus j}||_1 + |\beta_j| + |\beta_j^\star| \leq t\right\} \tag{108}$$

In particular, if the original LASSO solution is unique and such that:

$$||\hat{\beta}_{\setminus j}||_1 + |\hat{\beta}_j| = t$$

the set of solutions of the modified LASSO is:

$$B^\star = \left\{\beta = \left(\hat{\beta}_{\setminus j}, \beta_j, \beta_j^\star\right) : \beta_j + \beta_j^\star = \hat{\beta}_j,\ \beta_j \cdot \beta_j^\star \geq 0\right\}$$

The last condition ensures that $|\beta_j| + |\beta_j^\star| = |\hat{\beta}_j|$.

Ex. 3.29

Denote $\mathbf{X}^{(m)}$ the matrix containing m copies of $\mathbf{x}$:

$$\mathbf{X}^{(m)} = \mathbf{x}\, e_m^T$$

The Ridge regression coefficients are:

$$\begin{aligned}
\hat{\beta}(\lambda) &= \left(\mathbf{X}^{(m)\,T}\mathbf{X}^{(m)} + \lambda\right)^{-1} \mathbf{X}^{(m)\,T}\mathbf{y} \\
&= \left(||\mathbf{x}||^2\, e_m\, e_m^T + \lambda\right)^{-1} e_m\, \mathbf{x}^T\mathbf{y} \\
&= \frac{1}{\lambda}\left(1 - \frac{e_m\, e_m^T}{m + \lambda/||\mathbf{x}||^2}\right) e_m\, \mathbf{x}^T\mathbf{y} \\
&= \frac{||\mathbf{x}||^2}{m||\mathbf{x}||^2 + \lambda}\, \frac{\mathbf{x}^T\mathbf{y}}{||\mathbf{x}||^2}\, e_m = \frac{||\mathbf{x}||^2}{m||\mathbf{x}||^2 + \lambda}\, \hat{\beta}_{ols}\, e_m
\end{aligned}$$

This result is in accordance with Eq. (3.44) in the text: the PCA decomposition of $\mathbf{X}^{(m)}$ contains a single non-zero component with eigenvalue $m\,||\mathbf{x}||^2$ and eigenvector $\mathbf{x}/||\mathbf{x}||$.

Ex. 3.30

Consider the augmented matrices $\mathbf{y}'$, $\mathbf{X}'$ obtained by:

- $\mathbf{y}'$: adding p zero elements (rows) to $\mathbf{y}$;
- $\mathbf{X}'$: adding a $p \times p$ diagonal block at the bottom of $\mathbf{X}$, with diagonal elements $\sqrt{\lambda\alpha}$.

One can easily see that:

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

Therefore, the (λ, α) elastic-net problem on $\mathbf{X}, \mathbf{y}$ is the same as an ordinary $\lambda \cdot (1 - \alpha)$ LASSO problem on $\mathbf{X}', \mathbf{y}'$.

Chapter 4: Linear methods for Classification

Ex. 4.1

Without loss of generality, the matrix W can be assumed to be symmetric. As long as its eigenvalues are all zero, its square root is well-defined together with its inverse. Hence, using the change of variable $a \equiv W^{-1/2}b$:

$$\max_{a^T W a=1} a^T B\, a = \max_{||b||^2=1} b^T \tilde{B}\, b, \quad \tilde{B} \equiv W^{-1/2} B W^{-1/2}$$

Ex. 4.2

(a)

The linear discriminant functions are:

$$\begin{aligned}\hat{\delta}_k(x) &= x^T\hat{\Sigma}^{-1}\hat{\mu}_k - \frac{1}{2}\,\hat{\mu}_k^T\hat{\Sigma}^{-1}\hat{\mu}_k + \log\,\hat{\pi}_k \\ \hat{\pi}_k &= N_k/N\end{aligned}$$

So (a) follows directly from $\hat{\delta}_2(x) > \hat{\delta}_1(x)$.

(b), (c), (d)

Let's denote $\mathbf{e}_{(1)}$ and $\mathbf{e}_{(2)}$ the one-hot variables associated to the labels g. The coefficients $\hat{\beta}$ of the OLS regression with the bias term included can be obtained by regressing $\mathbf{y}$ against the demeaned version of the original predictors:

$$\begin{aligned}\hat{\beta} &= \left(\mathbf{X}_{dm}^T\mathbf{X}_{dm}\right)^{-1}\mathbf{X}_{dm}^T\mathbf{y} \\ \mathbf{X}_{dm} &\equiv \mathbf{X} - \mathbf{e}\,\hat{\mu}^T \\ &= \mathbf{X} - \left(\mathbf{e}_{(1)} + \mathbf{e}_{(2)}\right)\hat{\mu}^T \\ \hat{\mu} &\equiv \frac{1}{N}\mathbf{X}^T\mathbf{e}\end{aligned}$$

Notice that $\hat{\mu}$ is the vector of predictor means, regardless of the class. On the other hand, the LDA covariance matrix is obtained as:

$$\hat{\Sigma} = \frac{1}{N-2}\mathbf{X}_r^T\mathbf{X}_r$$

where $\mathbf{X}_r$ are the values of predictors centered around the class averages:

$$\begin{aligned}\mathbf{X}_r &= \mathbf{X} - \mathbf{e}_{(1)}\,\hat{\mu}_1^T - \mathbf{e}_{(2)}\,\hat{\mu}_2^T \\ \hat{\mu}_1 &\equiv \frac{1}{N_1}\mathbf{X}^T\mathbf{e}_{(1)} \\ \hat{\mu}_2 &\equiv \frac{1}{N_2}\mathbf{X}^T\mathbf{e}_{(2)}\end{aligned}$$

We have therefore:

$$\mathbf{X}_{dm} = \mathbf{X}_r + \mathbf{e}_{(1)}\left(\hat{\mu}_1 - \hat{\mu}\right)^T + \mathbf{e}_{(2)}\left(\hat{\mu}_2 - \hat{\mu}\right)^T$$

By construction, the residuals $\mathbf{X}_r$ are orthogonal to $\mathbf{e}_{(1)}$, $\mathbf{e}_{(2)}$:

$$\mathbf{X}_r^T\,\mathbf{e}_{(1)} = \mathbf{X}_r^T\,\mathbf{e}_{(2)} = 0$$

Also:

$$\begin{aligned}
\mathbf{e}_{(k)}^T \, \mathbf{e}_{(k)} &= N_k \\
\mathbf{e}_{(1)}^T \mathbf{e}_{(2)} &= 0 \\
\hat{\mu} &= \frac{N_1}{N}\hat{\mu}_1 + \frac{N_2}{N}\hat{\mu}_2
\end{aligned}$$

Hence:

$$\begin{aligned}
\mathbf{X}_{dm}^T \mathbf{X}_{dm} &= \mathbf{X}_r^T \mathbf{X}_r + N_1 \, (\hat{\mu}_1 - \hat{\mu})(\hat{\mu}_1 - \hat{\mu})^T + N_2 \, (\hat{\mu}_2 - \hat{\mu})(\hat{\mu}_2 - \hat{\mu})^T \\
&= (N-2)\,\hat{\Sigma} + N_1 \left(\frac{N_2}{N}\hat{\mu}_1 - \frac{N_2}{N}\hat{\mu}_2 \right)(\ldots) + N_2 \left(\frac{N_1}{N}\hat{\mu}_2 - \frac{N_1}{N}\hat{\mu}_1 \right)(\ldots) \\
&= (N-2)\,\hat{\Sigma} + \frac{N_1\,N_2}{N^2}(N_1 + N_2)\,(\hat{\mu}_1 - \hat{\mu}_2)\,(\hat{\mu}_1 - \hat{\mu}_2)^T \\
&= (N-2)\,\hat{\Sigma} + \frac{N_1\,N_2}{N}\,(\hat{\mu}_2 - \hat{\mu}_1)\,(\hat{\mu}_2 - \hat{\mu}_1)^T
\end{aligned}$$

Consider the general target encoding:

$$\mathbf{y} = y_1\,\mathbf{e}_{(1)} + y_2\,\mathbf{e}_{(2)}$$

We have:

$$\begin{aligned}
\mathbf{X}_{dm}^T \mathbf{y} &= N_1\,y_1\,(\hat{\mu}_1 - \hat{\mu}) + N_2\,y_2\,(\hat{\mu}_2 - \hat{\mu}) \\
&= \frac{N_1}{N}\,y_1\,(N\,\hat{\mu}_1 - N_1\hat{\mu}_1 - N_2\hat{\mu}_2) + \frac{N_2}{N}\,y_2\,(N\,\hat{\mu}_2 - N_1\hat{\mu}_1 - N_2\hat{\mu}_2) \\
&= \frac{N_1\,N_2}{N}\,(y_2 - y_1)\,(\hat{\mu}_2 - \hat{\mu}_1)
\end{aligned}$$

Therefore:

$$\begin{aligned}
\hat{\beta} &= \left((N-2)\,\hat{\Sigma} + \frac{N_1\,N_2}{N}\,(\hat{\mu}_1 - \hat{\mu}_2)\,(\hat{\mu}_1 - \hat{\mu}_2)^T \right)^{-1} \frac{N_1\,N_2}{N}\,(y_2 - y_1)\,(\hat{\mu}_2 - \hat{\mu}_1) \\
&= \left((N-2)\,\hat{\Sigma} + \frac{N_1\,N_2}{N}\hat{\Sigma}_B \right)^{-1} \frac{N_1\,N_2}{N}\,(y_2 - y_1)\,(\hat{\mu}_2 - \hat{\mu}_1) \\
\hat{\Sigma}_B &\equiv (\hat{\mu}_1 - \hat{\mu}_2)\,(\hat{\mu}_1 - \hat{\mu}_2)^T
\end{aligned}$$

This proves (d) and, in particular, substituting $y_1 = -N/N_1$, $y_2 = N/N_2$ yields the equation in the text. (c) is a direct consequence of this result.

(e)

After predictor demeaning:

$$\hat{f} = \hat{\beta}_0 + \hat{\beta}^T x_{dm}$$

the OLS value $\hat{\beta}_0$ is the sample average of the response, which is zero with the encoding indicated in the text. Therefore:

$$\begin{aligned}
\hat{f} &= \hat{\beta}^T \hat{\Sigma}^{-1} (x - \hat{\mu}) \\
&\propto (\hat{\mu}_2 - \hat{\mu}_1)^T \hat{\Sigma}^{-1} \left(x - \frac{N_1}{N} \hat{\mu}_1 - \frac{N_2}{N} \hat{\mu}_2 \right) \\
&= x^T \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1) - \frac{N_2}{N} \hat{\mu}_2^T \hat{\Sigma}^{-1} \hat{\mu}_2 + \frac{N_1}{N} \hat{\mu}_1^T \hat{\Sigma}^{-1} \hat{\mu}_1 + \frac{N_2 - N_1}{N} \hat{\mu}_1^T \hat{\Sigma}^{-1} \hat{\mu}_2
\end{aligned}$$

This expression coincides with the one for LDA only when $N_1 = N_2$.

Ex. 4.3

Denote $\mathcal{N}$ the diagonal matrix containing the sample counts for categories on the diagonal:

$$\mathcal{N} = \mathbf{Y}^T\,\mathbf{Y}$$

The matrix $\mathcal{M}$ containing the sample averages $\hat{\mu}_1, \ldots, \hat{\mu}_K$ as columns can be written as:

$$\mathcal{M} = \mathbf{X}^T\mathbf{Y}\mathcal{N}^{-1}$$

The matrix $\mathbf{X}_r$ containing the values of predictors centered around the category averages is:

$$\mathbf{X}_r = \mathbf{X} - \mathbf{Y}\mathcal{M}^T = \mathbf{X} - \mathbf{Y}\mathcal{N}^{-1}\mathbf{Y}^T\mathbf{X}$$

By construction, the columns of $\mathbf{X}_r$ are demeaned:

$$\mathbf{X}_r^T\mathbf{Y} = 0$$

Using these relationships, we deduce:

$$\begin{aligned}
\Sigma_X \equiv \mathbf{X}^T\mathbf{X} &= \mathbf{X}_r^T\mathbf{X}_r + \mathbf{X}^T\mathbf{Y}\mathcal{N}^{-1}\mathbf{Y}^T\mathbf{Y}\mathcal{N}^{-1}\mathbf{Y}^T\mathbf{X} \\
&= (N-K)\hat{\Sigma} + \mathbf{X}^T\mathbf{Y}\mathcal{N}^{-1}\mathbf{Y}^T\mathbf{X}
\end{aligned}$$

Now consider the product:

$$\begin{aligned}
\Sigma_X\hat{\Sigma}^{-1}\mathbf{X}^T\mathbf{Y} &= \left((N-K)\hat{\Sigma} + \mathbf{X}^T\mathbf{Y}\mathcal{N}^{-1}\mathbf{Y}^T\right)\hat{\Sigma}^{-1}\mathbf{X}^T\mathbf{Y} \\
&= \mathbf{X}^T\mathbf{Y}\left((N-K) + \mathcal{N}^{-1}\mathbf{Y}^T\hat{\Sigma}^{-1}\mathbf{X}^T\mathbf{Y}\right)
\end{aligned}$$

Multiplying both sides by Σ_X^{-1} on the left, and by $\left((N-K) + \mathcal{N}^{-1}\mathbf{Y}^T\hat{\Sigma}^{-1}\mathbf{X}^T\mathbf{Y}\right)^{-1}$ on the right, we obtain:

$$\begin{aligned}
\hat{B} &= \Sigma_X^{-1}\mathbf{X}^T\mathbf{Y} \\
&= \hat{\Sigma}^{-1}\mathbf{X}^T\mathbf{Y}\left((N-K) + \mathcal{N}^{-1}\mathbf{Y}^T\hat{\Sigma}^{-1}\mathbf{X}^T\mathbf{Y}\right)^{-1} \equiv \hat{\Sigma}^{-1}\mathbf{X}^T\mathbf{Y}\,\Sigma_Y^{-1} \qquad (109) \\
\Sigma_Y &\equiv (N-K) + \mathcal{N}^{-1}\mathbf{Y}^T\hat{\Sigma}^{-1}\mathbf{X}^T\mathbf{Y} \qquad (110)
\end{aligned}$$

Now, consider the following $(1,p)$ matrix:

$$g(x) \equiv x^T\hat{\Sigma}^{-1}\mathcal{M} = x^T\hat{\Sigma}^{-1}\mathbf{X}^T\mathbf{Y}\mathcal{N}^{-1} \qquad (111)$$

Aside from the $\log\hat{\pi}_k$ terms, which do not change when we move from x to $\hat{B}^Tx$, the two terms in the LDA discriminant functions are just elements taken from this matrix, for different values of x:

$$x^T\hat{\Sigma}^{-1}\mu_k - \frac{1}{2}\hat{\mu}_k^T\hat{\Sigma}^{-1}\mu_k = (f(x))_k - \frac{1}{2}(g(\hat{\mu}_k))_k$$

Indeed, $\mathcal{M}$ contains the sample averages $\hat{\mu}_k$ as columns. Therefore, in order to prove that LDA gives the same results in $B^T x$-space, it suffices to show that $g(x)$ does not change with this transformation. Since the transformation is linear, all the changes are pretty straightforward:

$$\begin{aligned} x &\rightarrow \hat{B}^T\mathbf{X} && (112)\\ \mathbf{X} &\rightarrow \mathbf{X}\hat{B} && (113)\\ \mathcal{M} &\rightarrow \hat{B}^T\mathcal{M} && (114)\\ \hat{\Sigma} &\rightarrow \hat{B}^T\hat{\Sigma}\hat{B} && (115)\end{aligned}$$

Hence:

$$g'(x' = \hat{B}^T\mathbf{X}) = x^T\hat{B}\left(\hat{B}^T\hat{\Sigma}\hat{B}\right)^{-1}\hat{B}^T\mathbf{X}^T\mathbf{Y}\mathcal{N}^{-1}$$

Notice that $\hat{B}$ is not a square matrix, hence we cannot apply the inverse to each matrix in the product in the parentheses. If we replace twice $\hat{B}$ according to (109), we obtain:

$$\begin{aligned} g'(x' = \hat{B}^T\mathbf{X}) &= x^T\hat{B}\left(\hat{B}^T\mathbf{X}^T\mathbf{Y}\,\Sigma_Y^{-1}\right)^{-1}\hat{B}^T\mathbf{X}^T\mathbf{Y}\mathcal{N}^{-1}\\ &= x^T\hat{B}\,\Sigma_Y\left(\hat{B}^T\mathbf{X}^T\mathbf{Y}\right)^{-1}\hat{B}^T\mathbf{X}^T\mathbf{Y}\mathcal{N}^{-1}\\ &= x^T\hat{B}\,\Sigma_Y\mathcal{N}^{-1}\\ &= x^T\hat{\Sigma}^{-1}\mathbf{X}^T\mathbf{Y}\,\Sigma_Y^{-1}\,\Sigma_Y\mathcal{N}^{-1}\\ &= x^T\hat{\Sigma}^{-1}\mathbf{X}^T\mathbf{Y}\mathcal{N}^{-1}\end{aligned}$$

which agrees with the expression (111) in the original feature space.

Bibliography

[1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press, USA, 2004.

[2] Philip E. Gill, Walter Murray, and Margaret H. Wright. *Practical optimization.* Academic Press Inc. [Harcourt Brace Jovanovich Publishers], London, 1981.