Solutions to the exercises of ESL, Second Edition

December 7, 2019

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:

$$P(Y = \pm 1) = \frac{1}{2} \tag{3}$$

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

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:

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

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:

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

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:

$$\mathrm{EPE}(x_0) = \mathrm{Var}_{y_0|x_0}(y_0) + \mathrm{Bias}^2_{\mathcal{T},y_0|x_0}(y_0) + \mathrm{Var}_{\mathcal{T}}(\hat{y}_0) \tag{12}$$

$$\mathrm{Var}_{y_0|x_0}(y_0) \equiv \mathbb{E}_{y_0|x_0}\left[\left(y_0 - \mathbb{E}_{y_0|x_0}\left[y_0\right]\right)^2\right] \tag{13}$$

$$\mathrm{Bias}_{\mathcal{T},y_0|x_0}(y_0) \equiv \mathbb{E}_{y_0|x_0}\left[y_0\right] - \mathbb{E}_{\mathcal{T}}\left[\hat{y}_0\right] \tag{14}$$

$$\mathrm{Var}_{\mathcal{T}}(\hat{y}_0) \equiv \mathbb{E}_{\mathcal{T}}\left[\left(\mathbb{E}_{\mathcal{T}}\left[\hat{y}_0\right] - \hat{y}_0\right)^2\right] \tag{15}$$

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:

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

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

Ex. 2.6

One has:

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

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:

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

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:

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

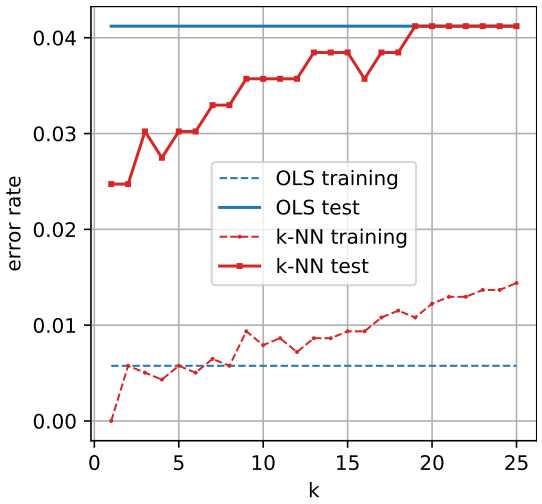
which is a simple consequence of the conditional expectations identity:

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

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

Ex. 2.8 (see Jupyter Notebook)

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 $\mathbb{E}_{Z_{te}}[R(Z_{tr}, Z_{te})]$ 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:

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

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 $\hat{\sigma}^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$. 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:

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

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).

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:

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

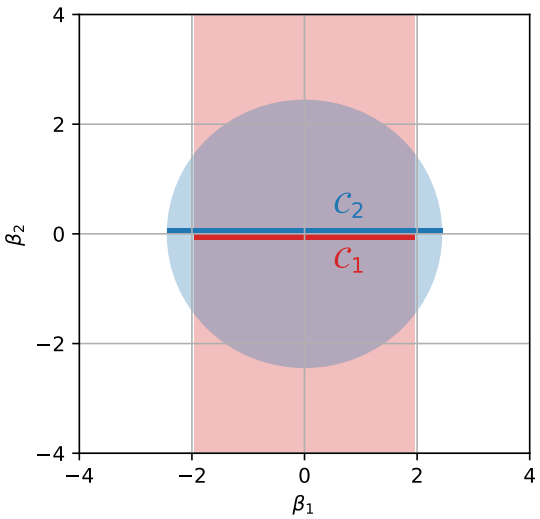


Figure 1: Confidence bands: single projection versus full vector. The two-dimensional vector β is taken to be centered and to have unit covariance. Both highlighted areas have 95% probability, but the band-shaped one has smaller projection on the β_1 axis.

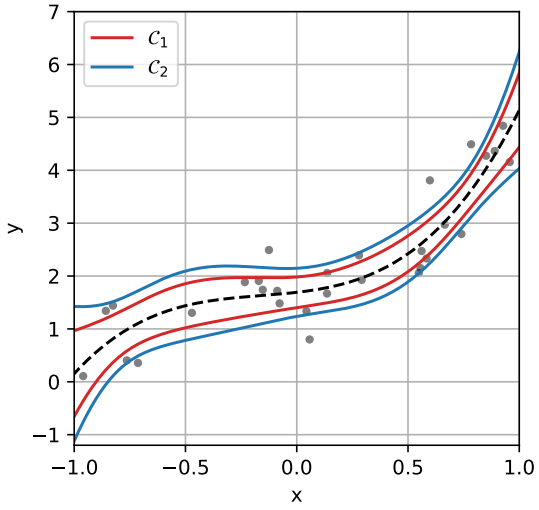


Figure 2: Confidence bands for a cubic univariate model. Values of x have been drawn from a uniform distribution in $[-1, 1]$, and y has been generated as $\beta^T x^{(3)} + \epsilon$, where β is a centered, unit-covariance gaussian vector, $x^{(3)} \equiv (1, x, x^2, x^3)$ and ϵ is a centered gaussian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

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:

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

Its variance-covariance matrix is:

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

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

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

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 $\langle \mathbf{q}_i, \mathbf{q}_j \rangle = \delta_i j$.
   - Compute the regression coefficient $\hat{\beta}_j^q$ of $\mathbf{y}$ against $\mathbf{q}_j$ as $\langle \mathbf{q}_j, \mathbf{y} \rangle$. Notice that, the $\mathbf{q}_j$'s being orthogonal, $\hat{\beta}_j^q$ is the coefficient of $\mathbf{q}_j$ in the regression of $\mathbf{y}$ against all of the $\mathbf{q}_j$'s:
     $$\mathbf{y} \sim \hat{\beta}_1^q \mathbf{q}_1 + \ldots + \hat{\beta}_p^q \mathbf{q}_p$$
   - 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:
     $$\begin{aligned} \hat{\beta}_j^q \mathbf{q}_j &= \hat{\beta}_j^q \left((R^{-1})_{1j} \mathbf{x}_1 + \ldots + (R^{-1})_{jj} \mathbf{x}_j \right) \\ \hat{\beta}_i &\to \hat{\beta}_i + (R^{-1})_{ij} \hat{\beta}_j^q \quad i = 1, \ldots, j \end{aligned}$$

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:

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

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 $\beta \equiv (\beta_j)_{j=1,\dots,p}$, 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

Let's denote $(\mathbf{q}_0, \mathbf{q}_1, \dots, \mathbf{q}_p)$ the columns of $\mathbf{Q}$: the matrix $\mathbf{Q}_2$ contains the variables $\mathbf{q}_1, \dots, \mathbf{q}_p$. By construction, these are orthogonal to the variable $\mathbf{q}_0$ which, considering that $(\mathbf{e}_0)_i = 1$ and remembering the Gram-Schmidt procedure, has all elements equal to $1/\sqrt{N}$. This means that $\mathbf{e} \cdot \mathbf{q}_j \propto \langle \mathbf{q}_j, \mathbf{q}_1 \rangle = 0$ for $j = 1, \dots, p$, i.e. the variables $\mathbf{q}_j$ are demeaned. Also, since the $\mathbf{q}_j$ are mutually orthogonal by construction, we can obtain them by first demeaning the $\mathbf{x}_j$'s to obtain $\tilde{\mathbf{X}}$, then applying the Gram-Schmidt procedure to perform the remaining residualisations. This shows that $\mathbf{Q}_2$ is part of the QR decomposition of $\tilde{\mathbf{X}}$, and their columns span the same subspace. The latter is also equal to the subspace spanned by the columns of $\mathbf{U}$ since DV^T is a non-singular $p \times p$ matrix.

Coming to the second part of the question, we try to determine under which circumstances the set of columns of $\mathbf{Q}_2$ in $\tilde{\mathbf{X}} = \mathbf{Q}_2 R_2$ is equal, up to sign flips and regardless of the column sorting, to the set of columns of $\mathbf{U}$, the matrix appearing in the SVD decomposition $\tilde{\mathbf{X}} = \mathbf{U}DV^T$. We remember that, provided the eigenvalues of $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$ are unique,

the SVD decomposition is unique, i.e. the set of columns of $\mathbf{U}$ is uniquely determined up to sign flips. Moreover, starting from a rotated version of the original predictors, and in particular shuffling them or flipping their sign, does not change this set of columns, since any such change can be re-absorbed in the matrix V. On the other hand, the QR decomposition depends explicitly on the initial order of predictors.

Remebering about how the columns of $\mathbf{Q}_2$ are obtained in the Gram-Schmidt procedure, one can conclude that the QR decomposition will result in the same set of vectors as the SVD decomposition if and only if:

$$j = 1, \dots, p: \quad \text{Span}(\mathbf{x}_1, \dots, \mathbf{x}_j) = \text{Span}(\mathbf{u}_{\sigma(1)}, \dots, \mathbf{u}_{\sigma(j)}) \tag{22}$$

for some permutation σ of $(1, \dots, p)$. Indeed, the Gram-Schmidt procedure at step j sets $\mathbf{q}_j$ to be the normal to $\text{Span}(\mathbf{x}_1, \dots, \mathbf{x}_{j-1})$ in $\text{Span}(\mathbf{x}_1, \dots, \mathbf{x}_j)$, which is equal to $\mathbf{u}_{\sigma(j)}$ up to a sign when (22). The converse is obviously true, since the QR decomposition is such that $\text{Span}(\mathbf{x}_1, \dots, \mathbf{x}_j) = \text{Span}(\mathbf{q}_1, \dots, \mathbf{q}_j)$ for every j. The condition (22) can be written as:

$$\tilde{\mathbf{X}} = \mathbf{U}DV^T = \mathbf{U}S_\sigma A$$

where S_σ is a shuffling matrix, and A is an upper-triangular matrix. A particularly simple case is V^T, i.e. when the columns of $\tilde{\mathbf{X}}$ are orthogonal: the two decompositions give the same result since one can take $S_\sigma = \mathbb{I}$, $A = D$.

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 $\mathbf{X}_1 = \mathbf{Q}R$:

$$\begin{aligned} \mathbf{X}_2^{(r)} &= \mathbf{X}_2 - \mathbf{X}_1 \left(\mathbf{X}_1^T \mathbf{X}_1 \right)^{-1} \mathbf{X}_1^T \mathbf{X}_2 & (23) \\ &= \left(\mathbb{I} - \mathbf{Q}\mathbf{Q}^T \right) \mathbf{X}_2 & (24) \end{aligned}$$

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

$$\bar{j} = \text{argmax} \frac{|\langle \mathbf{y}, \mathbf{x}_{2,j}^{(r)} \rangle|}{||\mathbf{x}_{2,j}^{(r)}||} = \text{argmax} \frac{|\langle \mathbf{r}, \mathbf{x}_{2,j}^{(r)} \rangle|}{||\mathbf{x}_{2,j}^{(r)}||} = \text{argmax} \frac{|\langle \mathbf{r}, \mathbf{x}_{2,j} \rangle|}{||\mathbf{x}_{2,j}^{(r)}||} \tag{25}$$

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 (23).

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 (25) is not equivalent to:

$$\bar{j} = \text{argmax} \frac{|\langle \mathbf{r}, \mathbf{x}_{2,j} \rangle|}{||\mathbf{x}_{2,j}||} \tag{26}$$

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 (26), 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 3.10, 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:

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

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

$$B = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}\mathbf{Y} \tag{27}$$

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

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

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 (27). 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:

$$\begin{aligned}
\mathbf{X}_a^T\mathbf{X}_a &= \begin{pmatrix} \mathbf{X}^T & \sqrt{\lambda}\,\mathbb{I} \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda}\,\mathbb{I} \end{pmatrix} = \mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\,\mathbb{I} \\
\mathbf{X}_a^T\mathbf{Y}_a &= \begin{pmatrix} \mathbf{X}^T & \sqrt{\lambda}\,\mathbb{I} \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ 0 \end{pmatrix} = \mathbf{X}^T\,\mathbf{y}
\end{aligned}$$

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:

$$\begin{aligned}
\hat{\phi}_m = \mathrm{argmax}_\alpha\, \mathrm{Corr}^2\,(\mathbf{y}, \mathbf{X}\alpha)\ \mathrm{Var}\,(\mathbf{X}\alpha) \\
\text{subject to}\, ||\alpha|| = 1, \quad \alpha^T S \hat{\phi}_l = 0,\ l = 1, \ldots, m-1
\end{aligned} \tag{28}$$

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:

$$\mathbf{z}_m = \sum_{j=1}^{p} \left(\hat{\varphi}_m\right)_j\, \mathbf{x}_j^{(m-1)} = \sum_{j=1}^{p} \left(\hat{\phi}_m\right)_j\, \mathbf{x}_j \tag{29}$$

I believe the correct statement is: the $\hat{\phi}_m$'s in this equation can be obtained via (28). 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:

$$\begin{aligned} \hat{\phi}_m &= \text{argmax}_\alpha \, \text{Corr}^2(\mathbf{y}, \mathbf{X}\alpha) \, \text{Var}(\mathbf{X}\alpha) = \text{argmax}_\alpha \, \left(\alpha^T \mathbf{X}^T \mathbf{y}\right)^2 \qquad (30) \\ &\quad \text{subject to } ||\alpha|| = 1, \quad \alpha^T S \hat{\phi}_l = 0, \; l = 1, \ldots, m-1 \\ \mathbf{v}_m &\equiv \mathbf{X}\hat{\phi}_m \qquad (31) \end{aligned}$$

and the one produced by Algorithm 3.3:

$$\begin{aligned} \hat{\varphi}_m &= \mathbf{X}^{(m-1)\,T} \mathbf{y} \qquad (32) \\ \mathbf{z}_m &\equiv \mathbf{X}^{(m-1)} \hat{\varphi}_m \qquad (33) \\ \mathbf{X}^{(m)} &= \mathbf{X}^{(m-1)} - \frac{\mathbf{z}_m \mathbf{z}_m^T \mathbf{X}^{(m-1)}}{||\mathbf{z}_m||^2} \qquad (34) \\ \mathbf{X}^{(0)} &\equiv \mathbf{X} \qquad (35) \end{aligned}$$

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 (30). 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:

$$0 = \left(\hat{\phi}_m^T \mathbf{X}^T \mathbf{y}\right) \mathbf{X}^T \mathbf{y} - \lambda \hat{\phi}_m - \sum_{l=1}^{m-1} \nu_l S \hat{\phi}_l = (s \cdot \hat{\phi}_m) s - \lambda \hat{\phi}_m - \sum_{l=1}^{m-1} \nu_l S \hat{\phi}_l \tag{36}$$

for some (m-specific) values of the Lagrange multipliers $\lambda, (\nu_l)_l$. Taking the scalar product of (36) 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:

$$\hat{\phi}_m \in \text{Span}\left(s, S\hat{\phi}_1, \ldots, S\hat{\phi}_{m-1}\right) \tag{37}$$

Note that if $\hat{\phi}_m$ is orthogonal to s, the maximizand in Equation (30) 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 (37) 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 (30) 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 $\text{Span}\left(s, \ldots, S^m s\right)$ and be S-orthogonal to $\text{Span}\left(s, \ldots, S^{m-1}s\right)$. 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 (30) 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:

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

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

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

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 (30) 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:

$$\mathcal{O}_{\bar{m}+1} = \left\{\alpha :\ \alpha^T Ss = \ldots = \alpha^T S^{\bar{m}} s = 0\right\} \tag{38}$$

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

$$S^{\bar{m}}s - a_1 Ss - \ldots - a_{\bar{m}-1}S^{\bar{m}-1}s = a_0 s \tag{39}$$

Note that $a_0 \neq 0$, otherwise one could multiply (39) 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 (38), therefore $\alpha \cdot s = 0$.

This completes the characterization of the sequence generated by (30). 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 (34) 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 (30) produce equivalent sequences.
To prove this key point, we adopt two inductive hypotheses:

$$\hat{\varphi}_m \in \text{Span}\left(s, Ss, \ldots, S^{m-1}s\right) - \text{Span}\left(s, Ss, \ldots, S^{m-2}s\right) \tag{40}$$

$$\mathbf{z}_m \propto \mathbf{v}_m \tag{41}$$

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:

$$\mathbf{X}^{(m)} = \mathbf{X}^{(m-1)} - \frac{\mathbf{z}_m\mathbf{z}_m^T\mathbf{X}}{||\mathbf{z}_m||^2} \tag{42}$$

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 $\left(s, \ldots, S^{m-1}s\right)$ 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 (40) for $m+1$.

To complete the proof, note that (42) 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:

$$\mathbf{z}_{m+1} = \mathbf{X}\hat{\varphi}_{m+1} - \frac{\mathbf{z}_1^T \mathbf{X} \hat{\varphi}_{m+1}}{||\mathbf{z}_1||^2} \mathbf{z}_1 - \ldots - \frac{\mathbf{z}_m^T \mathbf{X} \hat{\varphi}_{m+1}}{||\mathbf{z}_m||^2} \mathbf{z}_m \tag{43}$$

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 (43) 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 (30).*
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:

$$\Lambda^{m-1} - a_0\mathbb{I} - a_1\Lambda - \ldots - a_{m-2}\Lambda^{m-2} = 0 \tag{44}$$

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 (44) requires the Vandermonde matrix:

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

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:

$$\hat{\beta}_j = \mathbf{y}^T \mathbf{x}_j \quad (45)$$

$$RSS(\beta) = ||\mathbf{y}||^2 - 2\beta \cdot \hat{\beta} + ||\beta||^2 \quad (46)$$

Hence:

- **Best subset:** Minimizing (46) 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 (46):

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

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

- **Lasso:** Using again (46):

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

  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

Ex. 3.19

Remember that, after predictor demeaning:

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

Hence:

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

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:

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

Ex. 3.20

The canonical-correlation problem is:

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

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

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

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:

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

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

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

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:

$$\begin{aligned} u_m,\, v_m &= \underset{\substack{u,v \\ u^T\Sigma_X u=1, \quad v^T\Sigma_Y v=1 \\ u^T\Sigma_X u_1 = \ldots = u^T\Sigma_X u_{m-1} = 0 \\ v^T\Sigma_Y v_1 = \ldots = v^T\Sigma_Y v_{m-1} = 0}}{\operatorname{argmax}} \left(u^T\mathbf{Y}^T\mathbf{X}v\right) = \Sigma_Y^{-1/2}u^\star_m,\, \Sigma_X^{-1/2}v^\star_m \\ u^\star_m,\, v^\star_m &= \underset{\substack{u^\star,v^\star \\ ||u^\star|| = ||v^\star|| = 1 \\ u^{\star T}u^\star_1 = \ldots = u^{\star T}u^\star_{m-1} = 0 \\ v^{\star T}v^\star_1 = \ldots = v^{\star T}v^\star_{m-1} = 0}}{\operatorname{argmax}} \left(u^{\star T}U^\star D^\star V^{\star T}v^\star\right) \end{aligned} \tag{47}$$

We can prove recursively that $u^\star_n = \pm U^\star_n$ and $v^\star_n = \pm V^\star_n$ 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 (47), $u^\star$ must belong to the span of $U^\star_m,\ldots,U^\star_q$ and similarly for $v^\star$ with $V^\star_m\ldots,V^\star_q$. This is because, on the one hand, only the orthogonal projection of $u^\star_m$ on $U^\star_1,\ldots,U^\star_q$ contributes to the maximizand and, on the other hand, the inductive hypothesis requires $u^\star_m$ to be orthogonal to $U^\star_1,\ldots,U^\star_{m-1}$. Hence, we can write:

$$\begin{aligned} u^\star_m,\, v^\star_m &= U^\star_{m\to q}\alpha^\star_m,\, V^\star_{m\to q}\beta^\star_m \\ \alpha^\star_m,\, \beta^\star_m &= \underset{\substack{\alpha,\beta \\ ||\alpha||=||\beta||=1}}{\operatorname{argmax}} \left(\alpha^{\star T}D^\star_{m\to q}\beta^\star\right) \end{aligned}$$

where $U^\star_{m\to q}$ 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 $\alpha^\star_m = \beta^\star_m = (\pm 1, 0, \ldots, 0)$, i.e. $u^\star_m = \pm U^\star_m$, $v^\star_m = \pm V^\star_m$ with a common sign.

Ex. 3.21 and 3.22

The reduced-rank regression problem can be written as:

$$\hat{B}(m) = \underset{\operatorname{rank}(B)=m}{\operatorname{argmin}} \operatorname{Tr}\left[(\mathbf{Y}-\mathbf{X}B)\,\Sigma_Y^{-1}\,(\mathbf{Y}-\mathbf{X}B)^T\right] \tag{48}$$

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:

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

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:

$$B^\star = \mathcal{V}\mathcal{L} \tag{49}$$

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 (49) is not unique, since any rotation of the columns of $\mathcal{V}$ leaves $B^\star$ invariant:

$$\mathcal{V} \rightarrow \mathcal{V}\mathcal{R} \tag{50}$$

$$\mathcal{L} \rightarrow \mathcal{R}^T\mathcal{L} \tag{51}$$

$$\mathcal{R}^T\mathcal{R} = \mathcal{R}\mathcal{R}^T = \mathbb{I}_m \tag{52}$$

Keeping this degeneracy in mind, we can write:

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

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:

$$\begin{aligned}
\mathcal{V}^\star(m) &= \underset{\mathcal{V}^\star:\ \mathcal{V}^{\star T}\mathcal{V}^\star\,=\,\mathbb{I}_m}{\mathrm{argmax}}\ \mathrm{Tr}\left(\mathcal{V}^\star\mathcal{V}^{\star T}\mathbf{X}_n^T\mathbf{Y}_r\mathbf{Y}_r^T\mathbf{X}_n\right) \\
&= \underset{\mathcal{V}^\star:\ \mathcal{V}^{\star T}\mathcal{V}^\star\,=\,\mathbb{I}_m}{\mathrm{argmax}}\ \mathrm{Tr}\left(\mathcal{V}^\star\mathcal{V}^{\star T}V^\star D^{\star 2}V^{\star T}\right)
\end{aligned} \tag{53}$$

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

$$\Sigma_Y^{-1/2}(\mathbf{Y}^T\mathbf{X})\Sigma_X^{-1/2} = \mathbf{Y}_r^T\mathbf{X}_n = U^\star D^\star V^{\star T} \tag{54}$$

The maximization of (53) 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:

$$\mathcal{V}^\star(m) = V^\star_{(m)} \tag{55}$$

Putting everything together:

$$\begin{aligned}
\hat{B}(m) &= \Sigma_X^{-1/2}V^\star_{(m)}V^{\star T}_{(m)}V^\star D^\star U^{\star T}\Sigma_Y^{1/2} && (56)\\
&= \Sigma_X^{-1/2}V^\star_{(m)}D^\star_{(m)}U^{\star T}_{(m)}\Sigma_Y^{1/2} && (57)\\
&= \Sigma_X^{-1/2}V^\star D^\star U^{\star T}U^\star_{(m)}U^{\star T}_{(m)}\Sigma_Y^{1/2} && (58)
\end{aligned}$$

Notice how (56) and (58) can be re-written as:

$$\begin{aligned}
\hat{B}(m) &= \Sigma_X^{-1/2}V^\star_{(m)}V^{\star T}_{(m)}\Sigma_X^{-1/2}\mathbf{X}^T\mathbf{Y} && (59)\\
&= \Sigma_X^{-1}\mathbf{X}^T\mathbf{Y}\,\Sigma_Y^{-1/2}U^\star_{(m)}U^{\star T}_{(m)}\Sigma_Y^{1/2} && (60)
\end{aligned}$$

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

$$\hat{B}(m) = \hat{B}\,\Sigma_Y^{-1/2}U^\star_{(m)}U^{\star T}_{(m)}\Sigma_Y^{1/2} \tag{61}$$

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 (61) shows that $\hat{B}(m)$ depends on Σ_Y both explicitly and implicitly via the dependence of $U^\star$ through the generalized SVD decomposition (54). Equation (59) 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, (59) 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:

$$\begin{aligned}(\mathbf{Y}^T\mathbf{X})\Sigma_X^{-1/2} &= \Sigma_0^{1/2}\, U_0^\star D_0^\star\, V_0^{\star T} &(62)\\ &= \Sigma_{ols}^{1/2}\, U_{ols}^\star\, D_{ols}^\star\, V_{ols}^{\star T} &(63)\end{aligned}$$

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

$$\Sigma_{ols} = \Sigma_0^{1/2}\ \left(\mathbb{I} - U_0^\star\, D_0^{\star 2}\, U_0^{\star T}\right) \Sigma_0^{1/2} \qquad (64)$$

  Now consider the matrix:

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

  Using (64), 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 (63), 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 (59) lets us immediately deduce $\hat{B}_{ols}(m) = \hat{B}_0(m)$.