Some practice problems – the solutions

1. (a) The joint distribution of $(X_1, X_3, X_5)^T$ is 3-variate normal with mean vector $(1, 1, 0)$ and covariance matrix

$$C_{135} = \begin{pmatrix} 55 & -5 & -6 \\ -5 & 19 & -18 \\ -6 & -18 & 60 \end{pmatrix}$$

(b) The distribution of $X_1 + X_2 + X_3 + X_4 + X_5$ is normal with mean $\mu = 1 + 2 + 1 + 0 + 0 = 4$ and variance

$$\sigma^2 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 55 & 7 & -5 & -13 & -6 \\ 7 & 59 & -13 & 7 & -2 \\ -5 & -13 & 19 & -5 & -18 \\ -13 & 7 & -5 & 55 & -6 \\ -6 & -2 & -18 & -6 & 60 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = 140$$

(c) The joint distribution of $(X_1, X_5)^T$ is bivariate normal with mean vector $(1, 0)^T$ and covariance matrix

$$C_{15} = \begin{pmatrix} 55 & -6 \\ -6 & 60 \end{pmatrix}$$

and so $\text{Corr}(X_1, X_5) = \rho_{15} = -6/\sqrt{55 \times 60}$. The conditional distribution of X_1 given $X_5 = x_5$ is normal with mean

$$\mu_{1|5} = \mu_1 + \rho_{15}\frac{\sigma_1}{\sigma_5}(x - \mu_5) = 1 - \frac{6}{60}x$$

and variance $\sigma_1^2(1 - \rho_{15}^2) = 54.4$. Substituting $x = -1$, the conditional mean is 1.1.

(d) The correlation matrix is

$$R = \begin{pmatrix} 1.000 & 0.123 & -0.155 & -0.236 & -0.104 \\ 0.123 & 1.000 & -0.388 & 0.123 & -0.034 \\ -0.155 & -0.388 & 1.000 & -0.155 & -0.533 \\ -0.236 & 0.123 & -0.155 & 1.000 & -0.104 \\ -0.104 & -0.034 & -0.533 & -0.104 & 1.000 \end{pmatrix}$$

(e) There are no links between variables 1 and 2 nor between variables 2 and 4 since the corresponding elements of the concentration matrix are 0.

2. (a) The loadings are just eigenvectors of the correlation, standardized so that their sums of squares equal 1. If one of the principal components has equal loadings then (for example), the

vector $\boldsymbol{v} = (1\ 1 \cdots 1)^T$ is an eigenvector of $\widehat{R}$. Therefore, we just need to verify that $\widehat{R}\boldsymbol{v} = \lambda\boldsymbol{v}$ for some λ:

$$\begin{aligned}
\widehat{R}\boldsymbol{v} &= \begin{pmatrix} 1 & \rho & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \rho & \cdots & \rho \\ \vdots & \cdots & \ddots & \ddots & \cdots & \vdots \\ \rho & \rho & \rho & \cdots & \rho & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \\
&= (1 + (p-1)\rho) \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.
\end{aligned}$$

(b) $\widehat{R} = V\Lambda V^T$ where the columns of V are the PC loadings and the diagonal elements of Λ are the eigenvalues of $\widehat{R}$. Define

$$\boldsymbol{v}_1 = \begin{pmatrix} p^{-1/2} \\ p^{-1/2} \\ \vdots \\ p^{-1/2} \\ p^{-1/2} \\ \vdots \\ p^{-1/2} \end{pmatrix} \quad \text{and} \quad \boldsymbol{v}_2 = \begin{pmatrix} p^{-1/2} \\ p^{-1/2} \\ \vdots \\ p^{-1/2} \\ -p^{-1/2} \\ \vdots \\ -p^{-1/2} \end{pmatrix}$$

(which are the first two columns of V) and $A = (\lambda_1^{1/2}\boldsymbol{v}_1\ \ \lambda_2^{1/2}\boldsymbol{v}_2)$ (A is a $p \times 2$ matrix). If $\lambda_1 + \lambda_2 \approx p$ then $\lambda_3 + \cdots + \lambda_p \approx 0$ and so $\widehat{R} \approx AA^T$.

(c) The problem as stated is 100% well-posed! We need to assume that $\boldsymbol{x}$ and $\boldsymbol{y}$ are standardize variables. From part (b), it follows that $\boldsymbol{x}$ and $\boldsymbol{y}$ can be represented very well by their first two PC scores:

$$\begin{aligned}
s_1(\boldsymbol{x}) &= \frac{1}{4}(x_1 + \cdots + x_{16}) \\
s_2(\boldsymbol{x}) &= \frac{1}{4}(x_1 + \cdots x_8 - x_9 - \cdots - x_{16}) \\
s_1(\boldsymbol{y}) &= \frac{1}{4}(y_1 + \cdots + y_{16}) \\
s_2(\boldsymbol{y}) &= \frac{1}{4}(y_1 + \cdots y_8 - y_9 - \cdots - y_{16})
\end{aligned}$$

and so

$$d(\boldsymbol{x}, \boldsymbol{y}) \approx \left\{ (s_1(\boldsymbol{x}) - s_1(by))^2 + (s_2(\boldsymbol{x}) - s_2(\boldsymbol{y}))^2 \right\}^{1/2}.$$

3. (a) The standard deviations are the square roots of the eigenvalues and the some of the eigenvalues equals the number of variables (for PCA using the correlation matrix), which in this case is 5. Thus $1.9453216^2 + \mathtt{A}^2 + 0.60282550^2 + 0.39877095^2 + 0^2 = 5$. Therefore $\mathtt{A} = \sqrt{5 - 4.306693} = 0.8326506$.

$\texttt{B} = 0.8955166 + 0.07267972 = 0.9681963.$

(b) We know that the sum of squares of each loading is 1; therefore, $(-0.449)^2 + (-0.472)^2 + (-0.377)^2 + (-0.504)^2 + \texttt{C}^2 = 1$ and so $\texttt{C} = \pm\sqrt{1 - 0.82053} = \pm 0.423639$.
The sign of C can be determined in a number of ways. We know that the PC loadings are orthogonal vectors so (for example),

$$(-0.449) \times (-0.359) + (-0.472) \times 0.366 + (-0.377) \times (-0.691) + (-0.504) \times 0 + \texttt{C} \times 0.506 = 0$$

and so $\texttt{C} = -0.248946/0.506 = -0.4919881$, which suggests that the sign of C is negative. (The effects of rounding here are quite large!)
Alternatively, if you are given the correlation matrix, you can determine the value of C since you know that the loading vector is an eigenvector of $\widehat{R}$ with eigenvalue 1.9453216^2.

(c) If the variance of a PC is 0 or approximately 0, it means that the particular linear combination of variables (given by the PC loadings) takes the same value (or approximately the same value) for all observations. Therefore, in this case, we have $0.361 \times x_1 + 0.498 \times x_2 - 0.788 \times x_4$ approximately equal to some constant for all observations. (Note that this would hold for the raw variables as well.)