Some practice problems

Note: Some of hese problems are a bit more computationally intensive than what you might expect to see on the midterm but they should provide

1. Suppose that $\boldsymbol{X} = (X_1, \cdots, X_5)^T \sim \mathcal{N}_5(\boldsymbol{\mu}, C)$ where

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 2 \\ 1 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad C = \begin{pmatrix} 55 & 7 & -5 & -13 & -6 \\ 7 & 59 & -13 & 7 & -2 \\ -5 & -13 & 19 & -5 & -18 \\ -13 & 7 & -5 & 55 & -6 \\ -6 & -2 & -18 & -6 & 60 \end{pmatrix}$$

(a) What is the joint distribution of $(X_1, X_3, X_5)^T$?

(b) What is the distribution of $X_1 + X_2 + X_3 + X_4 + X_5$?

(c) What is the conditional distribution of X_1 given $X_5 = -1$? $X_1, \cdots, X_5$?

(d) What is the correlation matrix corresponding to this covariance matrix?

(e) The inverse of C is given by

$$\begin{pmatrix} 0.022 & 0.000 & 0.015 & 0.007 & 0.007 \\ 0.000 & 0.022 & 0.022 & 0.000 & 0.007 \\ 0.015 & 0.022 & 0.110 & 0.015 & 0.037 \\ 0.007 & 0.000 & 0.015 & 0.022 & 0.007 \\ 0.007 & 0.007 & 0.037 & 0.007 & 0.029 \end{pmatrix}$$

Give a graphical representation of the dependence structure of $\boldsymbol{X}$.

2. (a) Suppose that the correlation matrix for p-variate observations $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ is (rather improbably)

$$\widehat{R} = \begin{pmatrix} 1 & \rho & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \rho & \cdots & \rho \\ \vdots & \cdots & \ddots & \ddots & \cdots & \vdots \\ \rho & \rho & \rho & \cdots & \rho & 1 \end{pmatrix}$$

for some $0 < \rho < 1$.
Show that one of the principal components has equal loadings, that is, the coefficients of all p variables are equal.

(b) Suppose that p is even and the loadings for the first two principal components are

$$\begin{pmatrix} p^{-1/2} \\ p^{-1/2} \\ \vdots \\ p^{-1/2} \\ p^{-1/2} \\ \vdots \\ p^{-1/2} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} p^{-1/2} \\ p^{-1/2} \\ \vdots \\ p^{-1/2} \\ -p^{-1/2} \\ \vdots \\ -p^{-1/2} \end{pmatrix}$$

(so that half of the loadings for the second PC are $p^{-1/2}$ and the other half are $-p^{-1/2}$ and the variances of these first two PCs are $\lambda_1 \geq \lambda_2$ where $\lambda_1 + \lambda_2 \approx p$. Give an approximation for the correlation matrix.

(c) Assume the scenario of part (b) where $p = 16$. Suppose that $\boldsymbol{x} = (x_1, \cdots, x_{16})$ and $\boldsymbol{y} = (y_1, \cdots, y_{16})$. What is an approximation to

$$d(\boldsymbol{x}, \boldsymbol{y}) = \left\{ \sum_{i=1}^{16} (x_i - y_i)^2 \right\}^{1/2}$$

using the first two PCs?

3. The R output below gives the results of a principal component analysis using the correlation matrix. However, some of the values have been replaced by the letters `A`, `B`, and `C`.

```
> summary(r,loadings=T)
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4 Comp.5
Standard deviation     1.9453216         A 0.60282550 0.39877095      0
Proportion of Variance 0.7568552 0.1386614 0.07267972 0.03180365      0
Cumulative Proportion  0.7568552 0.8955166          B 1.00000000      1


Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
x1 -0.449 -0.359 -0.557  0.479  0.361
x2 -0.472  0.366        -0.625  0.498
x3 -0.377 -0.691  0.588 -0.186
x4 -0.504        -0.299 -0.175 -0.788
x5     C   0.506  0.499  0.561
```

(a) Find the values of `A` and `B`.

(b) What are the possible values of `C`? What information would you need to determine `C` exactly?

(c) The standard deviation (and variance) for the 5th principal component is approximately 0. What exactly does this mean?