MSc: Statistical Analysis (Spring Term)

Multivariate Analysis: Exercises 1 - Solutions

1. We have

$$\begin{aligned}
\text{cov}(Y_2, Y_1) &= \text{cov}(\mathbf{a}_2'\mathbf{X}, \mathbf{a}_1'\mathbf{X}) \\
&= E[(\mathbf{a}_2'\mathbf{X} - \mathbf{a}_2'\boldsymbol{\mu})(\mathbf{a}_1'\mathbf{X} - \mathbf{a}_1'\boldsymbol{\mu})] \\
&= E[\mathbf{a}_2'(\mathbf{X} - \boldsymbol{\mu})\mathbf{a}_1'(\mathbf{X} - \boldsymbol{\mu})] \\
&= E[\mathbf{a}_2'(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'\mathbf{a}_1] \\
&= \mathbf{a}_2' E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})']\mathbf{a}_1 \\
&= \mathbf{a}_2'\boldsymbol{\Sigma}\mathbf{a}_1.
\end{aligned}$$

2. We are given

$$R = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}.$$

An eigenvalue λ satisfies

$$\begin{vmatrix} 1-\lambda & r \\ r & 1-\lambda \end{vmatrix} = 0,$$

which $\Rightarrow \quad (1-\lambda)^2 - r^2 = 0 \Rightarrow (1-\lambda)^2 = r^2 \Rightarrow \lambda = 1 \pm r.$

If $r > 0$ then the largest eigenvalue is $1 + r$, accounting for $\frac{1+r}{2} \times 100\%$ of the total variance. This tends to 100% as r tends to 1 and 50% as r tends to 0.

To find the components we require corresponding eigenvectors. We solve

$$\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}\begin{pmatrix} e_1 \\ e_2 \end{pmatrix} = \quad (1+r) \quad \begin{pmatrix} e_1 \\ e_2 \end{pmatrix},$$

from which $\Rightarrow e_1 + re_2 = (1+r)e_1 \quad \Rightarrow \quad e_2 = e_1$ if $r \neq 0$.

Thus $\mathbf{a}_1' = (1/\sqrt{2}\,,\, 1/\sqrt{2})$, since $\mathbf{a}_1'\mathbf{a}_1 = 1$.

Hence, the first principal component is

$$\widetilde{y}_1 = \frac{1}{\sqrt{2}}\widetilde{y}_1 + \frac{1}{\sqrt{2}}\widetilde{y}_2.$$

Similarly we obtain the eigenvector corresponding to the eigenvalue $(1 - r)$ as

$$\widetilde{y}_2 = \frac{1}{\sqrt{2}}\widetilde{y}_1 - \frac{1}{\sqrt{2}}\widetilde{y}_2.$$

Thus $\widetilde{y}_1$ and $\widetilde{y}_2$ are independent of r, s_1^2 and s_2^2.

[Note that $\widetilde{x}_1$ and $\widetilde{x}_2$ represent the *standardized* data values, x_1/s_1 and x_2/s_2, from which the corresponding components $\widetilde{y}_1$ and $\widetilde{y}_2$ are calculated.]

Suppose now that the variables are *unstandardized* and $s_1 = 2$, $s_2 = 3$ and $r = 1/\sqrt{6}$. Then

$$S \;=\; \begin{pmatrix} 4 & \sqrt{6} \\ \sqrt{6} & 9 \end{pmatrix},$$

and

$$\begin{vmatrix} 4-\lambda & \sqrt{6} \\ \sqrt{6} & 9-\lambda \end{vmatrix} \;=\; 0$$

$$\Rightarrow \quad (4-\lambda)(9-\lambda) - 6 \;=\; 0 \;\Rightarrow\; \lambda^2 - 13\lambda + 30 \;=\; 0 \;\Rightarrow\; \lambda = 10 \text{ or } \lambda = 3.$$

To obtain the corresponding components we solve

$$\begin{pmatrix} 4 & \sqrt{6} \\ \sqrt{6} & 9 \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \;=\; \quad 10 \quad \begin{pmatrix} e_1 \\ e_2 \end{pmatrix},$$

with $e_1^2 + e_2^2 \;=\; 1$, yielding $e_1 = \frac{1}{\sqrt{7}}$, $\;e_1 = \frac{\sqrt{6}}{\sqrt{7}}$.

Thus

$$y_1 \;=\; \frac{1}{\sqrt{7}} x_1 \;+\; \frac{\sqrt{6}}{\sqrt{7}} x_2,$$

accounting for $(10/13) \times 100 \;\approx\; 77\%$ of the total sample variance.

Similarly we obtain

$$y_2 \;=\; \frac{6}{\sqrt{7}} x_1 \;-\; \frac{1}{\sqrt{7}} x_2,$$

accounting for $(3/13) \times 100 \;\approx\; 23\%$ of the sample variance.

If we now express the components $\widetilde{y}_1$ and $\widetilde{y}_1$, calculated from the correlation matrix, in terms of the original variables, we obtain

$$\begin{aligned} \widetilde{y}_1 \;&=\; \frac{1}{\sqrt{2}}\,\widetilde{x}_1 + \frac{1}{\sqrt{2}}\,\widetilde{x}_2 \\ &=\; \frac{1}{\sqrt{2}}\,\frac{x_1}{2} + \frac{1}{\sqrt{2}}\,\frac{x_2}{3} \\ &=\; \frac{1}{\sqrt{8}}\,x_1 + \frac{1}{\sqrt{18}}\,x_2. \end{aligned}$$

From this expression it clear that $\widetilde{y}_1$ is *not* proportional to

$$y_1 \;=\; \frac{1}{\sqrt{7}} x_1 \;+\; \frac{\sqrt{6}}{\sqrt{7}} x_2.$$

3. The first principal component accounts for 76.2% of the overall variance and can be interpreted as a measure of overall *size*, since all coefficients have the same sign and are of a comparable order of magnitude. [Note that the analysis was carried out on the correlation matrix so the variables have all been standardized and have the same importance in terms of sample variance.] Although it accounts for by far the largest percentage of the variance, the remaining components, accounting respectively for 11.8%, 6.8%, 2.8%, 1.3% and 1% of the variance, are nearly all relatively interpretable and may be of interest. It is often true that an overall 'size' component needs to be removed before features that may be of more interest are revealed. Here we may roughly interpret the first four components.

   PC1: Overall size, accounting for 76.2% of the variance.

   PC2: This component compares head size with overall body size. Thus it measures relative head to body size and accounts for about 11.8% of the variance.

   PC3: This measures head shape, the skull length relative to skull breadth; it accounts for 6.8% of the variance.

   PC4: This compares wing size to leg size and accounts for 2.8% of the variance.

   Components 5 and 6 are less interpretable.

   PC5: leg shape?

   PC6: wing shape ??

4. (a)
```
> ass1 <- c(8, 12, 14, 12, 9, 10, 11, 11, 12, 14)
> ass2 <- c(14, 13, 11, 13, 10, 12, 10, 15, 13, 10)
> ass3 <- c(7, 13, 8, 10, 12, 11, 10, 12, 10, 9)
> chem.data <- data.frame(ass1, ass2, ass3)

> var(chem.data)
           ass1       ass2       ass3
ass1  3.7888889 -0.9222222 -0.2888889
ass2 -0.9222222  3.2111111  0.3111111
ass3 -0.2888889  0.3111111  3.5111111

> cor(chem.data)
           ass1       ass2       ass3
ass1  1.0000000 -0.2643942 -0.0792050
ass2 -0.2643942  1.0000000  0.0926543
ass3 -0.0792050  0.0926543  1.0000000
```

   The first matrix is the sample covariance matrix, and the second is the sample correlation matrix. Note that there is an unusual correlation structure among the three assignment results. Assignments one and three are essentially uncorrelated, while assignments one and two are negatively correlated. This suggests that these assignments are testing very

different aspects of a students knowledge and ability in chemistry. Note also that the variances do not differ much, suggesting that there is no good reason not to use the covariance matrix in the principal components analysis.

(b) (i)
```
> chem.eig.cov <- eigen(var(chem.data))

> chem.eig.cov$values
[1] 4.623209 3.361408 2.526494

> chem.eig.cov$vectors
           [,1]       [,2]        [,3]
[1,]  0.7463196  0.3365007 -0.57425985
[2,] -0.5649449 -0.1359171 -0.81385737
[3,] -0.3519153  0.9318229  0.08866676

> percov <- (chem.eig.cov$values/sum(chem.eig.cov$values)) * 100
> percov
[1] 43.98401 31.97957 24.03642
```

PCs are:

$y_1 = 0.7463x_1 - 0.5649x_2 - 0.3519x_3$

$y_2 = \quad 0.3365x_1 - 0.1359x_2 + 0.9318x_3$

$y_3 = \quad -0.57426x_1 - 0.81386x_2 + 0.08867x_3.$

Note that the respective amounts of variance accounted for by each of $y_1,\ y_2,\ y_3$ are 44%, 32% and 24%. These are artificial data. However - rough interpretations?

y_1 A contrast between assignment 1 versus 2 and 3. High values indicate high ability in areas tested by assignments 2 and 3 and low ability in area tested by assignment 1.

y_2 Mainly assignment 3, plus perhaps a contrast between assignments 1 and 2.

y_3 Rough average of assignments 1 and 2, with perhaps more emphasis on assignment 2.

(ii)
```
> chem.eig.cor <- eigen(cor(chem.data))

> chem.eig.cor$values
[1] 1.3117837 0.9529924 0.7352239

> chem.eig.cor$vectors
           [,1]       [,2]        [,3]
[1,]  0.6545354  0.2856441 -0.69999344
[2,] -0.6630159 -0.2280570 -0.71302170
[3,] -0.3633088  0.9308047  0.04011519
```

```
> percor <- (chem.eig.cor$values/3) * 100
> percor
[1] 43.72612 31.76641 24.50746
```

PCs are:

$\widetilde{y}_1 = 0.6545\widetilde{x}_1 - 0.6630\widetilde{x}_2 - 0.3633\widetilde{x}_3$

$\widetilde{y}_2 = \quad 0.2856\widetilde{x}_1 - 0.2281\widetilde{x}_2 + 0.9308\widetilde{x}_3$

$\widetilde{y}_3 = \quad -0.69999\widetilde{x}_1 - 0.71302\widetilde{x}_2 + 0.04012\widetilde{x}_3$

where the $\widetilde{x}_i$ are 'standardized' versions of the x_i, i.e. $\widetilde{x}_i = x_i/S_i$, $i=1,2,3$.

Note that the percent variances accounted for and the 'interpretations' are very little changed.