Problem 8.1.

First, we determine the population principal components $Y_1$ and $Y_2$ for the covariance matrix $\Sigma = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}$.

If we denote the original variables as $X_1$ and $X_2$ then the PCs are

$$Y_1 = -0.8944X_1 - 0.4472X_2$$
$$Y_2 = 0.4472X_1 - 0.8944X_2$$

Finally the proportion $P_1$ of the total population variance explained by the first principal component is

$$P_1 = \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{6}{7} \approx 86\%$$

SPLUS SYNTAX and results of the computations:

```splus
> Sigma <- matrix(c(5, 2, 2, 2), nrow = 2, byrow = F)
> eigen(Sigma)
$values:
 [1] 6 1
$vectors:
 [,1]      [,2]
[1,] -0.8944272 0.4472136
[2,] -0.4472136 -0.8944272
```

Problem 8.2.

Let

$$\rho = \begin{bmatrix} 1 & \frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}} \\ \frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}} & 1 \end{bmatrix} = \begin{bmatrix} 1 & \sqrt{2/5} \\ \sqrt{2/5} & 1 \end{bmatrix}$$

(a) Using $\rho$ we obtain the following PCs:

$$Y_1 = 0.7071Z_1 + 0.7071Z_2$$
$$Y_2 = 0.7071Z_1 - 0.7071Z_2,$$

where $Z$'s are the standardize variables. The proportion $P_1$ of the total population variance explained by the first principal component is

$$P_1 = \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{1.6324555}{1.6324555 + 0.3675445} \approx 81.6\%$$
(b) Comparing the components calculated in part a with those obtained in exercise 8.1 we notice that the results are not the same. This is not surprising since the eigenvalues and eigenvectors of the correlation and covariance matrices have no simple relationship. In general, we expect the PCs obtained from the covariance and correlations matrices to be different.

(c) Correlations between the variables and the PCs:

\[
\rho_{Y_1, Z_1} = corr(Y_1, Z_1) = e_{11}\sqrt{\lambda_1} = 0.707 \times \sqrt{1.632} \approx 0.903.
\]

\[
\rho_{Y_1, Z_2} = corr(Y_1, Z_2) = e_{12}\sqrt{\lambda_1} = 0.707 \times \sqrt{1.632} \approx 0.903.
\]

\[
\rho_{Y_2, Z_1} = corr(Y_2, Z_1) = e_{21}\sqrt{\lambda_2} = 0.707 \times \sqrt{0.368} \approx 0.429.
\]

**SPLUS SYNTAX and computation results:**

\[
> \text{Rho} <- \text{matrix(c(1, sqrt(2/5), sqrt(2/5), 1), nrow = 2, byrow = T)}
\]

```R
> eigen(Rho)
```

$\text{values}$:

[1] 1.6324555 0.3675445

$\text{vectors}$:

[,1] [,2]
[1,] 0.7071068 0.7071068
[2,] 0.7071068 -0.7071068

**Problem 8.5.**

Let

\[
A = \begin{bmatrix}
1 & \rho & \rho \\
\rho & 1 & \rho \\
\rho & \rho & 1
\end{bmatrix}
\]

we will find the eigenvalues by solving the equation \( \text{det}(\rho - \lambda I) = 0 \):

\[
\begin{vmatrix}
1 - \lambda & \rho & \rho \\
\rho & 1 - \lambda & \rho \\
\rho & \rho & 1 - \lambda
\end{vmatrix} = 1 - 3\lambda + 3\lambda^2 - 3\rho^2 - \lambda^3 + 3\lambda\rho^2 + 2\rho^3 = 0
\]

\[
\Leftrightarrow (2\rho + 1 - \lambda)(\rho - 1 + \lambda)^2 = 0
\]

\[
\Leftrightarrow \lambda_1 = 1 + 2\rho = 1 + (3 - 1)\rho \quad \lambda_2 = \lambda_3 = 1 - \rho
\]

Hence, the results are consistent with (8 - 16). Now, we turn to (8 - 17) and have to consider eigenvectors corresponding to each eigenvalue.

\( \lambda_1 = 1 + 2\rho \):

\[
Ax = \lambda_1 x \Rightarrow \begin{bmatrix}
1 & \rho & \rho \\
\rho & 1 & \rho \\
\rho & \rho & 1
\end{bmatrix} \begin{bmatrix}
x_1 \\
x_2 \\
x_3
\end{bmatrix} = (1 + 2\rho) \begin{bmatrix}
x_1 \\
x_2 \\
x_3
\end{bmatrix}
\]
\[
\begin{align*}
\begin{cases}
    x_1 + \rho x_2 + \rho x_3 = (1 + 2\rho)x_1 \\
    \rho x_1 + x_2 + \rho x_3 = (1 + 2\rho)x_2 \\
    \rho x_1 + \rho x_2 + x_3 = (1 + 2\rho)x_3
\end{cases} \iff
\begin{cases}
    2x_1 = x_2 + x_3 \\
    2x_2 = x_1 + x_3 \\
    2x_3 = x_1 + x_2
\end{cases}
\end{align*}
\]

To get the eigenvector of unit length and equal components, we need \( e'_1 = \left[ \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}} \right] \). Hence, the results consistent with \((8 - 17)\).

**Problem 8.10.**

(a) Let \( x_1, x_2, x_3, x_4 \) and \( x_5 \) denote the Allied Chemical, Du Pont, Union Carbide, Exxon, and Texaco, respectively.

Then the PCs are

\[
\begin{align*}
    PC_1 &= 0.561x_1 + 0.470x_2 + 0.547x_3 + 0.291x_4 + 0.284x_5 \\
    PC_2 &= -0.739x_1 + 0.093x_2 + 0.654x_3 + 0.113x_4 - 0.071x_5 \\
    PC_3 &= -0.126x_1 - 0.467x_2 - 0.114x_3 + 0.610x_4 + 0.617x_5 \\
    PC_4 &= -0.284x_1 + 0.688x_2 - 0.500x_3 + 0.438x_4 - 0.062x_5 \\
    PC_5 &= 0.208x_1 - 0.281x_2 + 0.096x_3 + 0.582x_4 - 0.728x_5
\end{align*}
\]

(b) Let \( P_3 \) denotes the proportion of the total sample variance explained by the first three principal components, then

\[
P_3 = \frac{\lambda_1 + \lambda_2 + \lambda_3}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5} \approx 85.74%.
\]

This indicates that the first three principal components explain substantial proportion of the total data variance.

**INTERPRETATION OF THE PCs:** PC1 is essentially an average of all the stock returns. It may be thought of as the market component. I do not see a clear interpretation of PC2. PC3 is a contrast between the chemical and oil stock returns.

**SPLUS SYNTAX and computation results:**

```splus
> S <- var(data)
> S

V1       V2       V3       V4       V5
V1 0.0016299269 0.0008166676 0.0008100713 0.0004422405 0.0005139715
V2 0.0008166676 0.0012293759 0.0008276330 0.0003868550 0.0003109431
V3 0.0008100713 0.0008276330 0.0015560763 0.0004872816 0.0004624767
V4 0.0004422405 0.0003868550 0.0004872816 0.0008023323 0.0004084734
V5 0.0005139715 0.0003109431 0.0004624767 0.0004084734 0.0007587370
> eigen(S)

$values:
[1] 0.0035953867 0.0007921798 0.0007364426 0.0005086686 0.0003437707

$vectors:
[1,] 0.5605914 -0.73884565 -0.1260222 -0.28373183 0.20846832
[2,] 0.4698673 0.09286987 -0.4675066 0.68793190 -0.28069055
```

Problem 8.11.

(a) First, we multiply the fifth column in the data $X_5$, by 10 so that we make the median home value to be recorded in thousands of dollars.

SPLUS SYNTAX:

```splus
> data[, 5] <- 10 * data[, 5]
> S <- var(data)
> S
```

We could also modify the covariance matrix given in example 8.3 where $X_5$ is in $10,000$ units. The modification would be done by multiplying the off-diagonal elements in the fifth column and row by 10 and the diagonal element $s_{55}$ by 100 because $Cov(X, aY) = aCov(X, Y)$ and $Var(aX) = a^2Var(X)$ respectively.

(b) The eigenvalue-eigenvectors pairs are:

```
values:
50.9693577 6.6498899 1.41999710 0.22978608 0.01416317

vectors:
-0.0576598 -0.78220765 -0.02307185 -0.541282426 0.302171422
0.03281610 -0.35032828 -0.76422159 0.540439691 0.009137870
-0.03467366 -0.32717525 0.10109097 -0.051177198 -0.937504990
-0.07557524 -0.39085238 0.63207658 0.642117054 0.172301138
0.99432619 -0.07491367 0.07545108 -0.002204195 -0.002375237
```

Let $x_1, x_2, x_3, x_4$ and $x_5$ denote the Total Population, Median school years, Total employment, Health services employment, and Median home value, respectively. Then the first two PCs are

\[
PC_1 = -0.058x_1 + 0.033x_2 - 0.035x_3 - 0.076x_4 + 0.994x_5 \\
PC_2 = 0.782x_1 - 0.350x_2 - 0.327x_3 - 0.391x_4 - 0.075x_5
\]

(c) The proportion of total variance explained by the first two principal components is

\[
P_2 = 97.19\%.
\]
the correlation coefficients between the PCs and the variables are presented in the table below. Vi’s represent the variables.

<table>
<thead>
<tr>
<th></th>
<th>Comp.1</th>
<th>Comp.2</th>
<th>Comp.3</th>
<th>Comp.4</th>
<th>Comp.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>-0.1983408</td>
<td>-0.97188266</td>
<td>-0.01324679</td>
<td>-0.125017363</td>
<td>0.01732679107</td>
</tr>
<tr>
<td>V2</td>
<td>0.1762240</td>
<td>-0.67952602</td>
<td>-0.68499389</td>
<td>0.194864516</td>
<td>0.00081799158</td>
</tr>
<tr>
<td>V3</td>
<td>-0.2766482</td>
<td>-0.94289056</td>
<td>0.13462624</td>
<td>-0.027416498</td>
<td>-0.12468879082</td>
</tr>
<tr>
<td>V4</td>
<td>-0.3844669</td>
<td>-0.71819899</td>
<td>0.53670819</td>
<td>0.219331425</td>
<td>0.01461144422</td>
</tr>
<tr>
<td>V5</td>
<td>0.9995498</td>
<td>-0.02720128</td>
<td>0.01265990</td>
<td>-0.000148776</td>
<td>-0.00003980229</td>
</tr>
</tbody>
</table>

Comparing our results with example 8.3 we notice differences. In particular the proportion of total variance explained by the first two principal components has changed from 93.2% to 97.19%. This change is the effect of scale change that made values of $X_5$ HUGE compared to the values of the other variables. That implied that the variance of $X_5$ was also much larger than the variances of the rest of the variables and it dominated the total variance. The first PC is essentially $X_5$, so its variance is "almost" the total variance and thus the percent of the total variance explained by the first PC is large.

**SPLUS SYNTAX and results:**

```r
> pcomp <- princomp(data, scores = T)
> pcomp$loadings[, 1]
V1  V2  V3  V4  V5
-0.05765977 0.0328161 -0.03467366 -0.07557524 0.9943262

> pcomp$loadings[, 2]
V1  V2  V3  V4  V5
-0.7822076 -0.3503283 -0.3271752 -0.3908524 -0.07491367

> pcomp$correlations
Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
V1 -0.1983408 -0.97188266 -0.01324679 -0.125017363 0.01732679107
V2 0.1762240 -0.67952602 -0.68499389 0.194864516 0.00081799158
V3 -0.2766482 -0.94289056 0.13462624 -0.027416498 -0.12468879082
V4 -0.3844669 -0.71819899 0.53670819 0.219331425 0.01461144422
V5 0.9995498 -0.02720128 0.01265990 -0.000148776 -0.00003980229

> (pcomp$sdev)^2
Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
47.32869 6.174898 1.318569 0.2133728 0.01315151

> (cumsum((pcomp$sdev)^2)[2]/cumsum((pcomp$sdev)^2)[5]) * 100
Comp.2
97.19322
```