Chapter 8

8.1 Eigenvalues of $\boldsymbol{\Sigma}$ are $\lambda_1 = 6$, $\lambda_2 = 1$. The principal components are

$$Y_1 = .894X_1 + .447X_2$$
$$Y_2 = .447X_1 - .894X_2$$

$\mathrm{Var}(Y_1) = \lambda_1 = 6$. Therefore, proportion of total population variance explained by Y_1 is $6/(6+1) = .86$.

8.2

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & .6325 \\ .6325 & 1 \end{bmatrix}$$

(a) $Y_1 = .707Z_1 + .707Z_2$ $\qquad$ $\mathrm{Var}(Y_1) = \lambda_1 = 1.6325$

$Y_2 = .707Z_1 - .707Z_2$ $\qquad$ Proportion of total population variance explained by Y_1 is $1.6325/(1+1) = .816$

(b) No. The two (standardized) variables contribute equally to the principal components in 8.2(a). The two variables contribute unequally to the principal components in 8.1 because of their unequal variances.

(c) $\rho_{Y_1 Z_1} = .903$; $\quad \rho_{Y_1 Z_2} = .903$; $\quad \rho_{Y_2 Z_1} = .429$

8.3 Eigenvalues of $\boldsymbol{\Sigma}$ are 2, 4, 4. Eigenvectors associated with the eigenvalues 4, 4 are not unique. One choice is $\mathbf{e}_2' = [0 \ \ 1 \ \ 0]$ and $\mathbf{e}_3' = [0 \ \ 0 \ \ 1]$. With these assignments the principal components are $Y_1 = X_1$, $Y_2 = X_2$ and $Y_3 = X_3$.

8.4 Eigenvalues of $\boldsymbol{\Sigma}$ are solutions of $|\boldsymbol{\Sigma}-\lambda I| = (\sigma^2-\lambda)^3 - 2(\sigma^2-\lambda)(\sigma^2\rho)^2 = 0$
Thus $(\sigma^2-\lambda)[(\sigma^2-\lambda)^2 - 2\sigma^4\rho^2] = 0$ so $\lambda = \sigma^2$ or $\lambda = \sigma^2(1\pm\rho\sqrt{2})$. For $\lambda_1 = \sigma^2$, $\mathbf{e}_1' = [1/\sqrt{2}, 0, -1/\sqrt{2}]$. For $\lambda_2 = \sigma^2(1+\rho\sqrt{2})$; $\mathbf{e}_2' = [1/2, 1/\sqrt{2}, 1/2]$. For $\lambda_3 = \sigma^2(1-\rho\sqrt{2})$, $\mathbf{e}_3' = [1/2, -1/\sqrt{2}, 1/\sqrt{2}]$

Principal Component	Variance	Proportion of Total Variance Explained
$Y_1 = \frac{1}{\sqrt{2}} X_1 - \frac{1}{\sqrt{2}} X_3$	σ^2	$1/3$
$Y_2 = \frac{1}{2} X_1 + \frac{1}{\sqrt{2}} X_2 + \frac{1}{2} X_3$	$\sigma^2(1+\rho\sqrt{2})$	$\frac{1}{3}(1+\rho\sqrt{2})$
$Y_3 = \frac{1}{2} X_1 - \frac{1}{\sqrt{2}} X_2 + \frac{1}{2} X_3$	$\sigma^2(1-\rho\sqrt{2})$	$\frac{1}{3}(1-\rho\sqrt{2})$

8.5 (a) Eigenvalues of $\underset{\sim}{\rho}$ satisfy

$$|\underset{\sim}{\rho}-\lambda I| = (1-\lambda)^3 + 2\rho^3 - 3(1-\lambda)\rho^2 = 0$$

or $(1+2\rho-\lambda)(1-\rho-\lambda)^2 = 0$. Hence $\lambda_1 = 1+2\rho$; $\lambda_2 = \lambda_3 = 1-\rho$ and results are consistent with (8-16) for $p = 3$.

(b) By direct multiplication

$$\underset{\sim}{\rho}\left(\frac{1}{\sqrt{p}}\underset{\sim}{1}\right) = (1+(p-1)\rho)\left(\frac{1}{\sqrt{p}}\underset{\sim}{1}\right)$$

thus varifying the first eigenvalue-eigenvector pair. Further $\underset{\sim}{\rho}\,\underset{\sim}{e}_i = (1-\rho)\underset{\sim}{e}_i$, $i = 2,3,\ldots,p$.

8.6 (a)

$$\hat{y}_1 = .999x_1 + .041x_2 \quad \text{Sample variance of } \hat{y}_1 = \hat{\lambda}_1 = 7488.8$$

$$\hat{y}_2 = -.041x_1 + .999x_2 \quad \text{Sample variance of } \hat{y}_2 = \hat{\lambda}_2 = 13.8$$

(b) Proportion of total sample variance explained by $\hat{y}_1$ is $\hat{\lambda}_1/(\hat{\lambda}_1 + \hat{\lambda}_2) = .9982$

(c) Center of constant density ellipse is (155.60, 14.70). Half length of major axis is 102.4 in direction of $\hat{y}_1$. Half length of perpendicular minor axis is 4.4 in direction of $\hat{y}_2$.

(d) $r_{\hat{y}_1,x_1} = 1.000$, $r_{\hat{y}_1,x_2} = .687$ The first component is almost completely determined by x_1 = sales since its variance is approximately 285 times that of x_2 = profits. This is confirmed by the correlation coefficient $r_{\hat{y}_1,x_1} = 1.000$.

8.7 (a)

$$\hat{y}_1 = .707z_1 + .707z_2 \quad \text{Sample variance of } \hat{y}_1 = \hat{\lambda}_1 = 1.6861$$

$$\hat{y}_2 = .707z_1 - .707z_2 \quad \text{Sample variance of } \hat{y}_2 = \hat{\lambda}_2 = .3139$$

(b) Proportion of total sample variance explained by $\hat{y}_1$ is $\hat{\lambda}_1/(\hat{\lambda}_1 + \hat{\lambda}_2) = .8431$

(c) $r_{\hat{y}_1,z_1} = .918$, $r_{\hat{y}_1,z_2} = .918$ The standardized "sales" and "profits" contribute equally to the first sample principal component.

(d) The sales numbers are much larger than the profits numbers and consequently, sales, with the larger variance, will dominate the first principal component obtained from the sample covariance matrix. Obtaining the principal components from the sample correlation matrix (the covariance matrix of the standardized variables) typically produces components where the importance of the variables, as measured by correlation coefficients, is more nearly equal. It is usually best to use the correlation matrix or equivalently, to put the all the variables on similar numerical scales.

8.8 (a) $r_{\hat{y}_i, z_k} = \hat{e}_{ik}\sqrt{\hat{\lambda}_i} \quad i=1,2 \;\; k=1,2,\ldots,5$

Correlations:

i\k	1	2	3	4	5
1	.732	.831	.726	.604	.564
2	-.437	-.280	-.374	.694	.719

The correlations seem to reinforce the interpretations given in Example 8.5.

(b) Using (8-34) and (8-35) we have

k	$\bar{r}_k$
1	.353
2	.435
3	.354
4	.326
5	.299

$\bar{r} = .353$ $\qquad$ $\hat{\gamma} = 2.485$

$T = 103.1 > \chi^2_9(.01) = 21.67$ so would reject H_0 at the 1% level. This test assumes a large random sample and a multivariate normal parent population.

8.9 (a) By (5-10)

$$\max_{\underset{\sim}{\mu},\Sigma} L(\underset{\sim}{\mu},\Sigma) = \frac{e^{-\frac{np}{2}}}{(2\pi)^{\frac{pn}{2}} \left(\frac{n-1}{n}\right)^{\frac{pn}{2}} |S|^{\frac{n}{2}}}$$

The same result applied to each variable independently gives

$$\max_{\mu_i,\sigma_{ii}} L(\mu_i,\sigma_{ii}) = \frac{e^{-\frac{n}{2}}}{(2\pi)^{\frac{n}{2}} \left(\frac{n-1}{n}\right)^{\frac{n}{2}} s_{ii}^{\frac{n}{2}}}$$

Under H_0 , $\max\limits_{\underset{\sim}{\mu},\Sigma_0} L(\underset{\sim}{\mu},\Sigma_0) = \prod\limits_{i=1}^{p} L(\mu_i,\sigma_{ii})$

and the likelihood ratio statistic becomes

$$\Lambda = \frac{\max\limits_{\underset{\sim}{\mu},\Sigma_0} L(\underset{\sim}{\mu},\Sigma_0)}{\max\limits_{\underset{\sim}{\mu},\Sigma} L(\underset{\sim}{\mu},\Sigma)} = \frac{|S|^{\frac{n}{2}}}{\prod\limits_{i=1}^{p} s_{ii}^{\frac{n}{2}}}$$

(b) When $\Sigma = \sigma^2 I$, using (4-16) and (4-17) we get

$$\max_{\underset{\sim}{\mu}} L(\underset{\sim}{\mu},\sigma^2 I) = \frac{1}{(2\pi)^{\frac{np}{2}} (\sigma^2)^{\frac{np}{2}}} e^{-\frac{1}{2\sigma^2}\{tr[(n-1)S]\}}$$

8.9 (Continued)

so

$$\max_{\underset{\sim}{\mu},\sigma^2} L(\underset{\sim}{\mu},\sigma^2 I) = \frac{(np)^{np/2}\, e^{-np/2}}{(2\pi)^{np/2}(n-1)^{np/2}(tr[S])^{np/2}}$$

$$= \frac{e^{-np/2}}{(2\pi)^{np/2}\,\left(\frac{n-1}{n}\right)^{np/2}\,\left(\frac{1}{p}\,tr\,(S)\right)^{np/2}}$$

and the result follows. Under H_0 there are p μ_i's and one variance so the dimension of the parameter space is $\gamma_0 = p + 1$. The unrestricted case has dimension $p + p(p+1)/2$ so the χ^2 has $p(p+1)/2 - 1 = (p+2)(p-1)/2$ d.f.

8.10 (a) **Covariances: JPMorgan, CitiBank, WellsFargo, RoyDutShell, ExxonMobil**

	JPMorgan	CitiBank	WellsFargo	RoyDutShell	ExxonMobil
JPMorgan	0.00043327				
CitiBank	0.00027566	0.00043872			
WellsFargo	0.00015903	0.00017999	0.00022398		
RoyDutShell	0.00006410	0.00018144	0.00007341	0.00072251	
ExxonMobil	0.00008897	0.00012325	0.00006055	0.00050828	0.00076568

Principal Component Analysis: JPMorgan, CitiBank, WellsFargo, RoyDutShell, Exxon

Eigenanalysis of the Covariance Matrix
103 cases used

Eigenvalue	0.0013677	0.0007012	0.0002538	0.0001426	0.0001189
Proportion	0.529	0.271	0.098	0.055	0.046
Cumulative	0.529	0.801	0.899	0.954	1.000

Variable	PC1	PC2	PC3	PC4	PC5
JPMorgan	0.223	-0.625	-0.326	0.663	-0.118
CitiBank	0.307	-0.570	0.250	-0.414	0.589
WellsFargo	0.155	-0.345	0.038	-0.497	-0.780
RoyDutShell	0.639	0.248	0.642	0.309	-0.149
ExxonMobil	0.651	0.322	-0.646	-0.216	0.094

(b) From part (a),

$\hat{\lambda}_1 = .00137 \quad \hat{\lambda}_2 = .00070 \quad \hat{\lambda}_3 = .00025 \quad \hat{\lambda}_4 = .00014 \quad \hat{\lambda}_5 = .00012,$

so the total sample variance is $\sum_{i=1}^{5} \hat{\lambda}_i = .00258$ and the proportion of total variance explained by the first three components is $\sum_{i=1}^{3} \hat{\lambda}_i / \sum_{i=1}^{5} \hat{\lambda}_i = .899$. As in Example 8.5, the first component might be interpreted as a market component, the second component as an industry component, and the third component is difficult to interpret.

(c) Using (8-33), Bonferroni 90% simultaneous confidence intervals for $\lambda_1 \; \lambda_2 \; \lambda_3$ are

λ_1: (.00106, .00195)
λ_2: (.00054, .00100)
λ_3: (.00019, .00036)

(d) Stock returns are probably best summarized in two dimensions with 80% of the total variation accounted for by a "market" component and an "industry" component.

8.11 (a)

$$\mathbf{S} = \begin{bmatrix} 3.397 & -1.102 & 4.306 & -2.078 & .270 \\ & 9.673 & -1.513 & 10.953 & 12.030 \\ & & 55.626 & -28.937 & -.440 \\ & & & 89.067 & 9.570 \\ & \text{(Symmetric)} & & & 31.900 \end{bmatrix}$$

(b)

$\hat{\lambda}_1 = 108.27 \quad \hat{\lambda}_2 = 43.15 \quad \hat{\lambda}_3 = 31.29 \quad \hat{\lambda}_4 = 4.60 \quad \hat{\lambda}_5 = 2.35$

$\hat{e}_1$	$\hat{e}_2$	$\hat{e}_3$	$\hat{e}_4$	$\hat{e}_5$
-0.037630	-0.062264	0.040076	0.554515	0.828018
0.118931	-0.249442	-0.259861	-0.769147	0.514314
-0.479670	-0.759246	0.431404	-0.027909	-0.081081
0.858905	-0.315978	0.393975	0.068822	-0.049884
0.128991	-0.507549	-0.767815	0.308887	-0.202000

$$\hat{y}_1 = -.038x_1 + .119x_2 - .480x_3 + .859x_4 + .129x_5$$
$$\hat{y}_2 = -.062x_1 - .249x_2 - .759x_3 - .316x_4 - .508x_5$$

(c) Correlations between variables and components:

	x_1	x_2	x_3	x_4	x_5
$r_{\hat{y}_1, x_i}$	-.212	.398	-.669	.947	.238
$r_{\hat{y}_2, x_i}$	-.222	-.527	-.669	-.220	-.590

The proportion of total sample variance explained by the first two principal Components is (108.27+43.15)/(108.27+43.15+31/29+4.60+2.35)=.80.

The first component appears to be a weighted difference between percent total employment and percent employed by government. We might call this component an employment contrast. The second component appears to be influenced most by roughly equal contributions from percent with professional degree (x_2), percent employment (x_3) and median home value (x_5). We might call this an achievement component. The change in scale for x_5 did not appear to have much affect on the first sample principal component (see Example 8.3) but did change the nature of the second component. Variable x_5 now has much more influence in the second principal component.

8.12

$$S = \begin{bmatrix} 2.500 & -2.768 & -.378 & -.464 & -.586 & -2.235 & .171 \\ & 300.516 & 3.914 & -1.395 & 6.779 & 30.779 & .624 \\ & & 1.522 & .673 & 2.316 & 2.822 & .142 \\ & & & 1.182 & 1.089 & -.811 & .177 \\ & & & & 11.364 & 3.133 & 1.045 \\ & & & & & 30.978 & .593 \\ \text{(Symmetric)} & & & & & & .479 \end{bmatrix}$$

$$R = \begin{bmatrix} 1.0 & -.101 & -.194 & -.270 & -.110 & -.254 & .156 \\ & 1.0 & .183 & -.074 & .116 & .319 & .052 \\ & & 1.0 & .502 & .557 & .411 & .166 \\ & & & 1.0 & .297 & -.134 & .235 \\ & & & & 1.0 & .167 & .448 \\ & & & & & 1.0 & .154 \\ \text{(Symmetric)} & & & & & & 1.0 \end{bmatrix}$$

<u>Using S</u>:

$\hat{\lambda}_1 = 304.26$; $\hat{\lambda}_2 = 28.28$; $\hat{\lambda}_3 = 11.46$; $\hat{\lambda}_4 = 2.52$; $\hat{\lambda}_5 = 1.28$;

$\hat{\lambda}_6 = .53$; $\hat{\lambda}_7 = .21$

The first sample principal component

$$\hat{y}_1 = -.010x_1 + .993x_2 + .014x_3 - .005x_4 + .024x_5 + .112x_6 + .002x_7$$

accounts for 87% of the total sample variance. The first component is essentially "solar radiation". (Note the large sample variance for x_2 in S).

<u>Using R</u>:

$\hat{\lambda}_1 = 2.34$; $\hat{\lambda}_2 = 1.39$; $\hat{\lambda}_3 = 1.20$; $\hat{\lambda}_4 = .73$; $\hat{\lambda}_5 = .65$;
$\hat{\lambda}_6 = .54$; $\hat{\lambda}_7 = .16$

The first three sample principle components are

$$\hat{y}_1 = .237z_1 - .205z_2 - .551z_3 - .378z_4 - .498z_5 - .324z_6 - .319z_7$$

$$\hat{y}_2 = -.278z_1 + .527z_2 + .007z_3 - .435z_4 - .199z_5 + .567z_6 - .308z_7$$

$$\hat{y}_3 = .644z_1 + .225z_2 - .113z_3 - .407z_4 + .197z_5 + .159z_6 + .541z_7$$

These components account for 70% of the total sample variance.

The first component contrasts "wind" with the remaining variables. It might be some general measure of the pollution level. The second component is largely composed of "solar radiation" and the pollutants "NO" and "O_3". It might represent the effects of solar radiation since solar radiation is involved in the production of NO and O_3 from the other pollutants. The third component is composed largely of "wind" and certain pollutants (e.g. "NO" and "HC"). It might represent a wind transport effect. A "better" interpretation of the components would depend on more extensive subject matter knowledge.

The data can be effectively summarized in three or fewer dimensions. The choice of S or R makes a difference.

8.13

(a) Covariance Matrix

	X1	X2	X3
X1	4.654750889	0.931345370	0.589699088
X2	0.931345370	0.612821160	0.110933412
X3	0.589699088	0.110933412	0.571428861
X4	0.276915309	0.118469052	0.087004959
X5	1.074885659	0.388886434	0.347989910
X6	0.158150852	-0.024851988	0.110131391

	X4	X5	X6
X1	0.276915309	1.074885659	0.158150852
X2	0.118469052	0.388886434	-0.024851988
X3	0.087004959	0.347989910	0.110131391
X4	0.110409072	0.217405649	0.021814433
X5	0.217405649	0.862172372	-0.008817694
X6	0.021814433	-0.008817694	0.861455923

Correlation Matrix

	X1	X2	X3	X4	X5	X6
X1	1.0000	0.5514	0.3616	0.3863	0.5366	0.0790
X2	0.5514	1.0000	0.1875	0.4554	0.5350	-.0342
X3	0.3616	0.1875	1.0000	0.3464	0.4958	0.1570
X4	0.3863	0.4554	0.3464	1.0000	0.7046	0.0707
X5	0.5366	0.5350	0.4958	0.7046	1.0000	-.0102
X6	0.0790	-.0342	0.1570	0.0707	-.0102	1.0000

(b) We will work with **R** since the sample variance of x1 is approximately 40 times larger than that of x4.

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	2.86431	1.78786	0.477385	0.47738
PRIN2	1.07645	0.29881	0.179408	0.65679
PRIN3	0.77764	0.12733	0.129607	0.78640
PRIN4	0.65031	0.26228	0.108386	0.89479
PRIN5	0.38803	0.14478	0.064672	0.95946
PRIN6	0.24326		0.040543	1.00000

Eigenvectors

	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5	PRIN6
X1	0.444858	-.026660	0.339330	-.551149	-.600851	0.146492
X2	0.429300	-.291738	0.498607	-.061367	0.687297	0.076408
X3	0.358773	0.380135	-.628157	-.421060	0.331839	0.211635
X4	0.462854	-.020959	-.124585	0.665604	-.207413	0.532689
X5	0.521276	-.073690	-.203339	0.200526	-.103175	-.794127
X6	0.055877	0.873960	0.429880	0.178715	0.053090	-.116262

(c) It is not possible to summarize the radiotherapy data with a single component. We need the first four components to summarize the data.

(d) Correlations between principal components and $X1 - X6$ are

	PRIN1	PRIN2	PRIN3	PRIN4
X1	0.75289	-0.02766	0.29923	-0.44446
X2	0.72656	-0.30268	0.43969	-0.04949
X3	0.60720	0.39440	-0.55393	-0.33955
X4	0.78335	-0.02175	-0.10986	0.53676
X5	0.88222	-0.07646	-0.17931	0.16171
X6	0.09457	0.90675	0.37909	0.14412

8.14 S is given in Example 5.2.

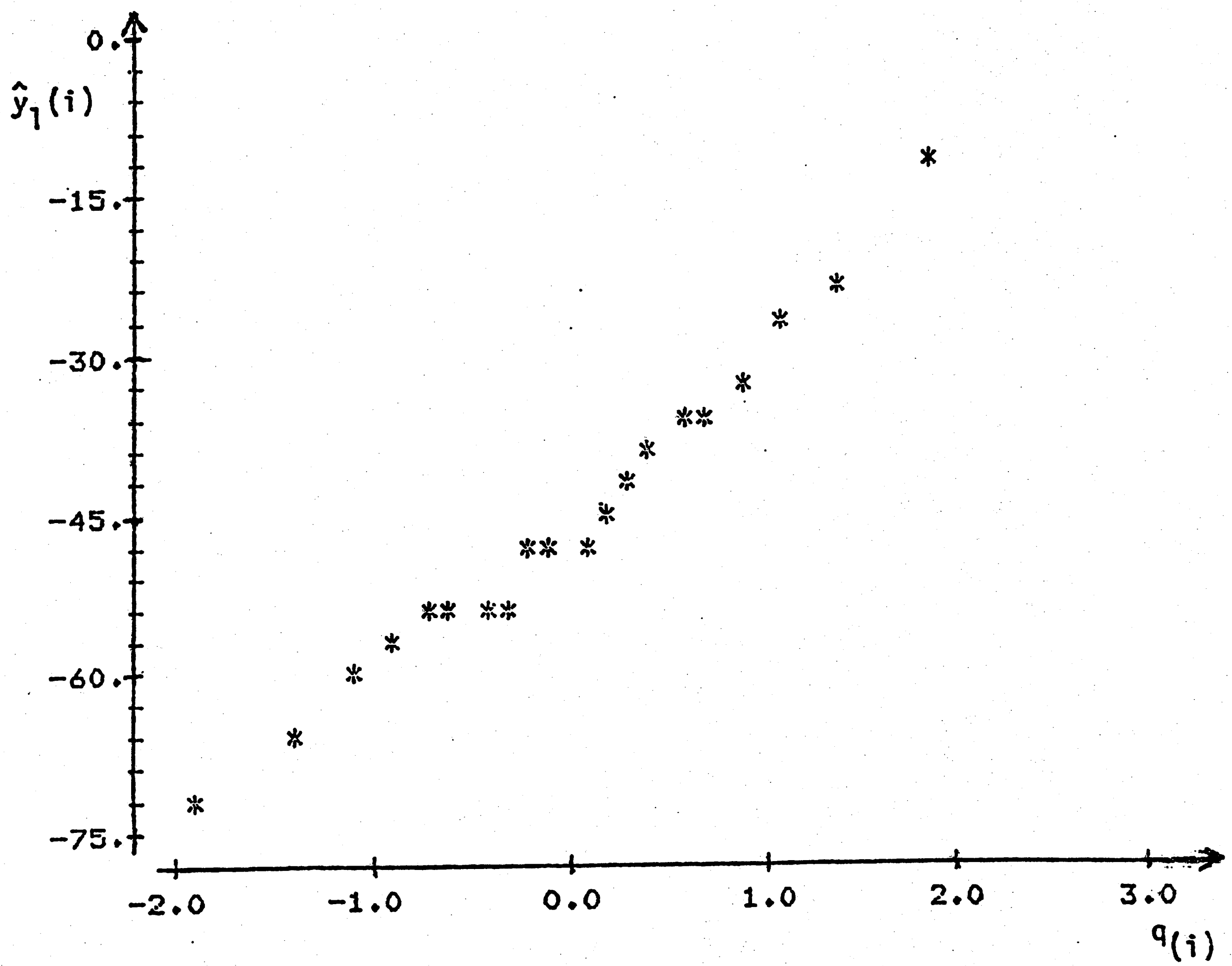
$$\hat{\lambda}_1 = 200.5, \quad \hat{\lambda}_2 = 4.5, \quad \hat{\lambda}_3 = 1.3$$

The first sample principal component explains a proportion $200.5/(200.5 + 4.5 + 1.3) = .97$ of the total sample variance. Also,

$$\hat{\underset{\sim}{e}}_1' = [-.051, -.998, .029]$$

Hence $\hat{y}_1 = -.051x_1 - .998x_2 + .029x_3$

The first principal component is essentially X_2 = sodium content. (Note the (relatively) large sample variance for sodium in S). A Q-Q plot of the $\hat{y}_1$ values is shown below. These data appear to be approximately normal with no suspect observations.



Q-Q plot for $\hat{y}_1$.

8.15

$$S = \begin{bmatrix} 1088.40 & 831.28 & 763.23 & 784.09 \\ & 1128.41 & 850.32 & 926.73 \\ & & 1336.15 & 904.53 \\ \text{(Symmetric)} & & & 1395.15 \end{bmatrix}$$

$$\hat{\lambda}_1 = 3779.01; \quad \hat{\lambda}_2 = 468.25; \quad \hat{\lambda}_3 = 452.13; \quad \hat{\lambda}_4 = 248.72$$

Consequently, the first sample principal component accounts for a proportion $3779.01/4948.11 = .76$ of the total sample variance. Also,

$$\hat{\underset{\sim}{e}}_1' = [.45, \ .49, \ .51, \ .53]$$

Consequently,

$$\hat{y}_1 = .45x_1 + .49x_2 + .51x_3 + .53x_4$$

The interpretation of the first component is the same as the interpretation of the first component, obtained from **R**, in Example 8.6. (Note the sample variances in S are nearly equal).

8.16. Principal component analysis of Wisconsin fish data

(a) All are positively correlated.

(b) Principal component analysis using $x1 - x4$

```
Eigenvalues of R
2.1539 0.7875 0.6157 0.4429

Eigenvectors of R
0.7032  0.4295  0.1886 -0.7071
0.6722  0.3871 -0.4652  0.4702
0.5914 -0.7126 -0.2787 -0.3216
0.6983 -0.2016  0.4938  0.5318

                      pc1    pc2    pc3    pc4
         St. Dev. 1.4676 0.8874 0.7846 0.6655
    Prop. of Var. 0.5385 0.1969 0.1539 0.1107
 Cumulative Prop. 0.5385 0.7354 0.8893 1.0000
```

The first principal component is essentially a total of all four. The second contrasts the Bluegill and Crappie with the two bass.

(c) Principal component analysis using $x1 - x6$

```
Eigenvalues of R
2.3549 1.0719 0.9842 0.6644 0.5004 0.4242

Eigenvectors of R
-0.6716  0.0114  0.5284 -0.0471  0.3765 -0.7293
-0.6668 -0.0100  0.2302 -0.7249 -0.1863  0.5172
-0.5555 -0.2927 -0.2911  0.1810 -0.6284 -0.3081
-0.7013 -0.0403  0.0355  0.6231  0.3407  0.5972
 0.3621 -0.4203  0.0143 -0.2250  0.5074  0.0872
-0.4111  0.0917 -0.8911 -0.2530  0.4021 -0.1731

                      pc1    pc2    pc3    pc4    pc5    pc6
         St. Dev. 1.5346 1.0353 0.9921 0.8151 0.7074 0.6513
    Prop. of Var. 0.3925 0.1786 0.1640 0.1107 0.0834 0.0707
 Cumulative Prop. 0.3925 0.5711 0.7352 0.8459 0.9293 1.0000
```

The Walleye is contrasted with all the others in the first principal component (look at the covariance pattern). The second principal component is essentially the Walleye and somewhat the largemouth bass. The third principal component is nearly a contrast between Northern pike and Bluegill.

8.17

COVARIANCE MATRIX

x1	.0130016					
x2	.0103784	.0114179				
x3	.0223500	.0185352	.0803572			
x4	.0200857	.0210995	.0667762	.0694845		
x5	.0912071	.0085298	.0168369	.0177355	.0115684	
x6	.0079578	.0089085	.0128470	.0167936	.0080712	.0105991

The eigenvalues are

0.164 0.018 0.008 0.003 0.002 0.001

and the first two principal components are

$$[\ .218\ ,\ .204\ ,\ .673\ ,\ .633\ ,\ .181\ ,\ .159\ ]\ \underset{\sim}{x}$$

$$[\ .337\ ,\ .432\ ,\ -.500\ ,\ .024\ ,\ .430\ ,\ .514\ ]\ \underset{\sim}{x}$$

8.18 (a) & (b) Principal component analysis of the correlation matrix follows.

Correlations: 100m(s), 200m(s), 400m(s), 800m, 1500m, 3000m, Marathon

	100m(s)	200m(s)	400m(s)	800m	1500m	3000m
200m(s)	0.941					
400m(s)	0.871	0.909				
800m	0.809	0.820	0.806			
1500m	0.782	0.801	0.720	0.905		
3000m	0.728	0.732	0.674	0.867	0.973	
Marathon	0.669	0.680	0.677	0.854	0.791	0.799

Eigenanalysis of the Correlation Matrix

Eigenvalue	5.8076	0.6287	0.2793	0.1246	0.0910	0.0545	0.0143
Proportion	0.830	0.090	0.040	0.018	0.013	0.008	0.002
Cumulative	0.830	0.919	0.959	0.977	0.990	0.998	1.000

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7
100m(s)	0.378	-0.407	0.141	-0.587	0.167	-0.540	0.089
200m(s)	0.383	-0.414	0.101	-0.194	-0.094	0.745	-0.266
400m(s)	0.368	-0.459	-0.237	0.645	-0.327	-0.240	0.127
800m	0.395	0.161	-0.148	0.295	0.819	0.017	-0.195
1500m	0.389	0.309	0.422	0.067	-0.026	0.189	0.731
3000m	0.376	0.423	0.406	0.080	-0.352	-0.240	-0.572
Marathon	0.355	0.389	-0.741	-0.321	-0.247	0.048	0.082

$$\hat{y}_1 = .378z_1 + .383z_2 + .368z_3 + .395z_4 + .389z_5 + .376z_6 + .355z_7$$
$$\hat{y}_2 = -.407z_1 - .414z_2 - .459z_3 + .161z_4 + .309z_5 + .423z_6 + .389z_7$$

	z_1	z_2	z_3	z_4	z_5	z_6	z_7
$r_{\hat{y}_1,z_i}$	.911	.923	.887	.952	.937	.906	.856
$r_{\hat{y}_2,z_i}$	-.323	-.328	-.364	.128	.245	.335	.308

Cumulative proportion of total sample variance explained by the first two components is .919.

(c) All track events contribute about equally to the first component. This component might be called a track index or track excellence component. The second component contrasts the times for the shorter distances (100m, 200m 400m) with the times for the longer distances (800m, 1500m, 3000m, marathon) and might be called a distance component.

(d) The "track excellence" rankings for the first 10 and very last countries follow. These rankings appear to be consistent with intuitive notions of athletic excellence.

1. USA 2. Germany 3. Russia 4. China 5. France 6. Great Britain 7. Czech Republic 8. Poland 9. Romania 10. Australia .... 54. Somoa

8.19 Principal component analysis of the covariance matrix follows.

Covariances: 100m/s, 200m/s, 400m/s, 800m/s, 1500m/s, 3000m/s, Marm/s

	100m/s	200m/s	400m/s	800m/s	1500m/s	3000m/s
100m/s	0.0905383					
200m/s	0.0956063	0.1146714				
400m/s	0.0966724	0.1138699	0.1377889			
800m/s	0.0650640	0.0749249	0.0809409	0.0735228		
1500m/s	0.0822198	0.0960189	0.0954430	0.0864542	0.1238405	
3000m/s	0.0921422	0.1054364	0.1083164	0.0997547	0.1437148	0.1765843
Marm/s	0.0810999	0.0933103	0.1018807	0.0943056	0.1184578	0.1465604

	Marm/s
Marm/s	0.1667141

Eigenanalysis of the Covariance Matrix

Eigenvalue	0.73215	0.08607	0.03338	0.01498	0.00885	0.00617	0.00207
Proportion	0.829	0.097	0.038	0.017	0.010	0.007	0.002
Cumulative	0.829	0.926	0.964	0.981	0.991	0.998	1.000

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7
100m/s	0.310	-0.376	0.098	-0.585	-0.046	-0.624	0.138
200m/s	0.357	-0.434	0.089	-0.323	-0.030	0.689	-0.311
400m/s	0.379	-0.519	-0.274	0.667	-0.187	-0.124	0.132
800m/s	0.299	0.053	-0.053	0.128	0.894	-0.136	-0.265
1500m/s	0.391	0.211	0.435	0.055	0.127	0.236	0.734
3000m/s	0.460	0.396	0.427	0.184	-0.357	-0.199	-0.499
Marm/s	0.423	0.445	-0.730	-0.237	-0.136	0.081	0.095

$$\hat{y}_1 = .310x_1 + .357x_2 + .379x_3 + .299x_4 + .391x_5 + .460x_6 + .423x_7$$
$$\hat{y}_2 = -.376x_1 - .434x_2 - .519x_3 + .053x_4 + .211x_5 + .396x_6 + .445x_7$$

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
$r_{\hat{y}_1,x_i}$	.882	.902	.874	.944	.951	.937	.886
$r_{\hat{y}_2,x_i}$	-.367	-.376	-.410	.057	.176	.276	.320

Cumulative proportion of total sample variance explained by the first two components is .926.

The interpretation of the sample component is similar to the interpretation in Exercise 8.18. All track events contribute about equally to the first component. This component might be called a track index or track excellence component. The second component contrasts times in m/s for the shorter distances (100m, 200m 400m) with the times for the longer distances (800m, 1500m, 3000m, marathon) and might be called a distance component.

The "track excellence" rankings for the countries are very similar to the rankings for the countries obtained in Exercise 8.18.

8.20 (a) & (b) Principal component analysis of the correlation matrix follows.

Eigenanalysis of the Correlation Matrix

Eigenvalue	6.7033	0.6384	0.2275	0.2058	0.0976	0.0707	0.0469	0.0097
Proportion	0.838	0.080	0.028	0.026	0.012	0.009	0.006	0.001
Cumulative	0.838	0.918	0.946	0.972	0.984	0.993	0.999	1.000

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
100m	0.332	0.529	0.344	-0.381	0.300	-0.362	0.348	-0.066
200m	0.346	0.470	-0.004	-0.217	-0.541	0.349	-0.440	0.061
400m	0.339	0.345	-0.067	0.851	0.133	0.077	0.114	-0.003
800m	0.353	-0.089	-0.783	-0.134	-0.227	-0.341	0.259	-0.039
1500m	0.366	-0.154	-0.244	-0.233	0.652	0.530	-0.147	-0.040
5000m	0.370	-0.295	0.183	0.055	0.072	-0.359	-0.328	0.706
10,000m	0.366	-0.334	0.244	0.087	-0.061	-0.273	-0.351	-0.697
Marathon	0.354	-0.387	0.335	-0.018	-0.338	0.375	0.594	0.069

$$\hat{y}_1 = .332z_1 + .346z_2 + .339z_3 + .353z_4 + .366z_5 + .370z_6 + .366z_7 + .354z_8$$
$$\hat{y}_2 = .529z_1 + .470z_2 + .345z_3 - .089z_4 - .154z_5 - .295z_6 - .334z_7 - .387z_8$$

	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8
$r_{\hat{y}_1,z_i}$	.860	.896	.878	.914	.948	.958	.948	.917
$r_{\hat{y}_2,z_i}$	.423	.376	.276	-.071	-.123	-.236	-.267	-.309

Cumulative proportion of total sample variance explained by the first two components is .918.

(c) All track events contribute about equally to the first component. This component might be called a track index or track excellence component. The second component contrasts the times for the shorter distances (100m, 200m 400m) with the times for the longer distances (800m, 1500m, 5000m, 10,000m, marathon) and might be called a distance component.

(d) The male "track excellence" rankings for the first 10 and very last countries follow. These rankings appear to be consistent with intuitive notions of athletic excellence.

1. USA 2. Great Britain 3. Kenya 4. France 5. Australia 6. Italy
7. Brazil 8. Germany 9. Portugal 10. Canada .... 54. Cook Islands

The principal component analysis of the men's track data is consistent with that for the women.

8.21 Principal component analysis of the covariance matrix follows.

Covariances: 100m/s, 200m/s, 400m/s, 800m/s, 1500m/s, 5000m/s, 10,000m/s, Marathonm/s

	100m/s	200m/s	400m/s	800m/s	1500m/s
100m/s	0.0434979				
200m/s	0.0482772	0.0648452			
400m/s	0.0434632	0.0558678	0.0688217		
800m/s	0.0314951	0.0432334	0.0428221	0.0468840	
1500m/s	0.0425034	0.0535265	0.0537207	0.0523058	0.0729140
5000m/s	0.0469252	0.0587731	0.0617664	0.0571560	0.0766388
10,000m/s	0.0448325	0.0572512	0.0599354	0.0553945	0.0745719
Marathonm/s	0.0431256	0.0562945	0.0567342	0.0541911	0.0736518

	5000m/s	10,000m/s	Marathonm/s
5000m/s	0.0959398		
10,000m/s	0.0937357	0.0942894	
Marathonm/s	0.0905819	0.0909952	0.0979276

Eigenanalysis of the Covariance Matrix

Eigenvalue	0.49405	0.04622	0.01391	0.01332	0.00752	0.00575	0.00322
Proportion	0.844	0.079	0.024	0.023	0.013	0.010	0.006
Cumulative	0.844	0.923	0.947	0.970	0.983	0.993	0.998

Eigenvalue	0.00112
Proportion	0.002
Cumulative	1.000

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
100m/s	0.244	-0.432	0.173	-0.450	-0.390	0.119	0.584	-0.119
200m/s	0.311	-0.523	0.235	-0.318	0.341	-0.247	-0.535	0.096
400m/s	0.317	-0.469	-0.684	0.420	0.046	0.177	0.039	-0.008
800m/s	0.278	-0.033	0.436	0.543	0.332	-0.368	0.432	-0.070
1500m/s	0.364	0.063	0.439	0.317	-0.303	0.608	-0.327	-0.044
5000m/s	0.428	0.261	-0.111	-0.016	-0.374	-0.334	-0.006	0.696
10,000m/s	0.421	0.310	-0.187	-0.100	-0.215	-0.352	-0.180	-0.693
Marathonm/s	0.416	0.387	-0.128	-0.339	0.584	0.391	0.215	0.074

$$\hat{y}_1 = .244x_1 + .311x_2 + .317x_3 + .278x_4 + .364x_5 + .428x_6 + .421x_7 + .416x_8$$
$$\hat{y}_2 = -.432x_1 - .523x_2 - .469x_3 - .033x_4 + .063x_5 + .261x_6 + .310x_7 + .387x_8$$

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
$r_{\hat{y}_1, x_i}$	.822	.858	.849	.902	.948	.971	.964	.934
$r_{\hat{y}_2, x_i}$	-.445	-.442	-.384	-.033	.050	.181	.217	.266

Cumulative proportion of total sample variance explained by the first two components is .923.

The interpretation of the sample component is similar to the interpretation in Exercise 8.20. All track events contribute about equally to the first component. This component might be called a track index or track excellence component. The second component contrasts times in m/s for the shorter distances (100m, 200m 400m, 800m) with the times for the longer distances (1500m, 5000m, 10,000m, marathon) and might be called a distance component.

The "track excellence" rankings for the countries are very similar to the rankings for the countries obtained in Exercise 8.20.

8.22 Using S

Eigenvalues of the Covariance Matrix

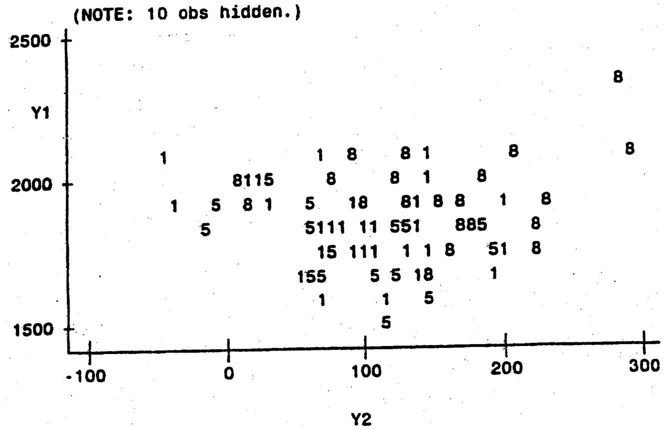
	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	20579.6	15704.9	0.808198	0.80820
PRIN2	4874.7	4869.2	0.191437	0.99964
PRIN3	5.4	2.1	0.000213	0.99985
PRIN4	3.3	2.8	0.000130	0.99998
PRIN5	0.5	0.4	0.000018	1.00000
PRIN6	0.1	0.1	0.000003	1.00000
PRIN7	0.0	.	0.000000	1.00000

✓

Eigenvectors

	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5	PRIN6	PRIN7	
X3	0.005887	0.009680	0.286337	0.608787	0.535569	-.509727	0.024592	yrhgt
X4	0.487047	0.872697	-.034277	-.003227	0.000444	-.000457	-.000253	ftfrbody
X5	0.008526	0.029196	0.904389	-.425175	0.008388	0.010389	0.014293	prctffb
X6	0.003112	0.004886	0.133267	0.311194	0.390573	0.855204	-.037984	frame
X7	0.000069	-.000493	-.018864	-.005278	0.011906	0.043786	0.998778	bkfat
X8	0.009330	0.008577	0.284215	0.593037	-.748598	0.082331	0.013820	saleht
X9	0.873259	-.487193	0.004847	-.005597	0.002665	-.000341	-.000256	salewt

Plot of Y1*Y2. Symbol is value of X1.
(NOTE: 10 obs hidden.)



8.22 (Continued)

Using *R*

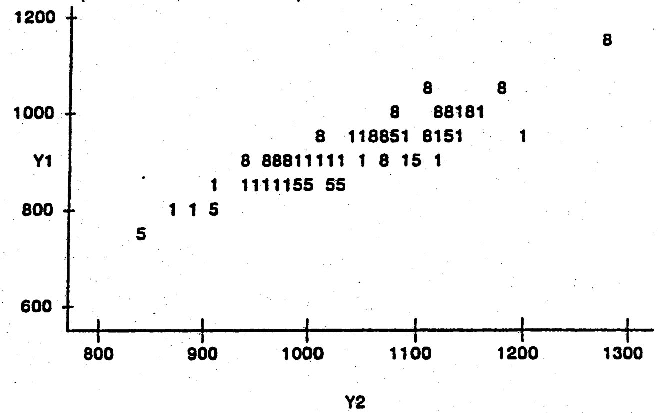
Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	4.12070	2.78357	0.588671	0.58867
PRIN2	1.33713	0.59575	0.191018	0.77969
PRIN3	0.74138	0.31996	0.105912	0.88560
PRIN4	0.42143	0.23562	0.060204	0.94580
PRIN5	0.18581	0.03930	0.026544	0.97235
PRIN6	0.14650	0.09945	0.020929	0.99328
PRIN7	0.04706	.	0.006722	1.00000

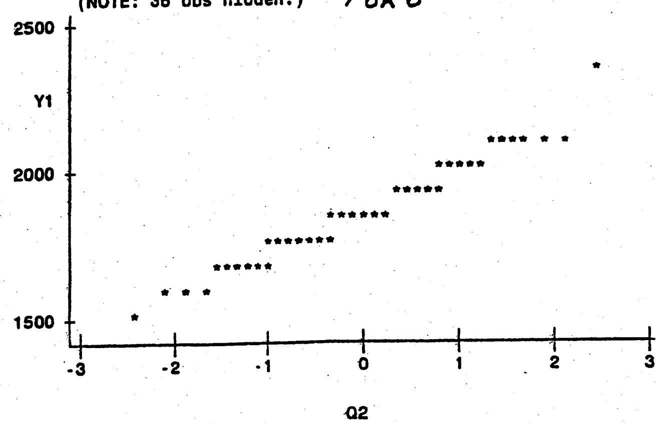
Eigenvectors

	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5	PRIN6	PRIN7	
X3	0.449931	-.042790	-.415709	0.113356	0.065871	-.072234	0.774926	yrhgt
X4	0.412326	0.129837	0.450292	0.247479	-.719343	-.177061	0.017768	ftfrbody
X5	0.355562	-.315508	0.568273	0.314787	0.579367	0.127800	-.002397	prctffb
X6	0.433957	0.007728	-.452345	0.242818	0.142995	-.434144	-.582337	frame
X7	-.186705	0.714719	-.038732	0.618117	0.160238	0.208017	0.042442	bkfat
X8	0.452854	0.101315	-.176650	-.215769	-.109535	0.799288	-.236723	saleht
X9	0.269947	0.600515	0.253312	-.582433	0.290547	-.276561	0.047036	salewt

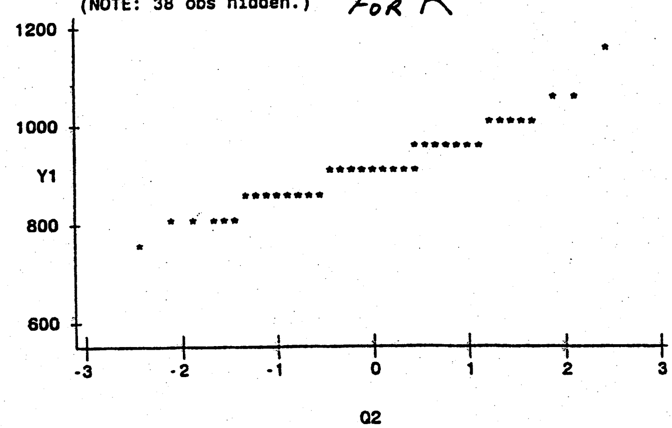
Plot of Y1*Y2. Symbol is value of X1.
(NOTE: 27 obs hidden.)



Plot of Y1*Q2. Symbol used is '*'.
(NOTE: 36 obs hidden.) *For S*



Plot of Y1*Q2. Symbol used is '*'.
(NOTE: 38 obs hidden.) *For R*



8.23 **a)** <u>Using *S*</u>

Eigenvalues of S

4478.87 152.47 32.32 8.12 1.52 0.54

Eigenvectors of S (in columns)

-0.849339	0.470832	-0.226606	0.074260	-0.008692	-0.000202
-0.368552	-0.846078	-0.368132	0.012754	-0.110784	-0.019105
-0.194132	-0.058127	0.303143	-0.928388	-0.012289	-0.070597
-0.314678	-0.216748	0.848576	0.355060	-0.082353	0.032666
-0.043918	-0.060354	0.001815	-0.060162	0.440119	0.892805
-0.064458	-0.092026	0.033880	0.052267	0.887138	-0.443264

The first component might be identified as a "size" component. It is dominated by Weight, Body length and Girth, those variables with the largest sample variances. The first component explains 4478.87/4673.84 = .958 or 95.8% of the total sample variance. The second component essentially contrasts Weight with the remaining body size variables, Body length, Neck, Girth, Head length, and Head width, although the sample correlation between the second component and Neck is small (–.05). The first two components explain 99.1% of the total sample variance.

These body measurement data can be effectively summarized in one dimension.

b) <u>Using *R*</u>

R

1.0000	0.8752	0.9559	0.9437	0.9025	0.9045
0.8752	1.0000	0.9013	0.9177	0.9461	0.9503
0.9559	0.9013	1.0000	0.9635	0.9270	0.9200
0.9437	0.9177	0.9635	1.0000	0.9271	0.9439
0.9025	0.9461	0.9270	0.9271	1.0000	0.9544
0.9045	0.9503	0.9200	0.9439	0.9544	1.0000

Eigenvalues of R

5.6447 0.1758 0.0565 0.0492 0.0473 0.0266

Eigenvectors of R (in columns)

-0.403672	-0.558334	0.286817	0.261937	-0.598371	0.128024
-0.404313	0.532348	-0.186741	0.719785	0.004276	0.012490
-0.409938	-0.389366	0.035396	0.073950	0.561034	-0.599053
-0.411999	-0.222694	-0.581252	-0.228969	0.231095	0.580499
-0.409162	0.318718	0.695916	-0.291938	0.251473	0.313431
-0.410333	0.319513	-0.243840	-0.519785	-0.458838	-0.435168

8.23 (Continued)

Again, the first principal component is a "size" component. All variables contribute equally to the first component. This component explains 5.6447/6 = .941 or 94.1% of the total sample variance. The second principal component contrasts Weight, Neck and Girth with Body length, Head length and Head width. The first two components explain 97% of the total sample variance.

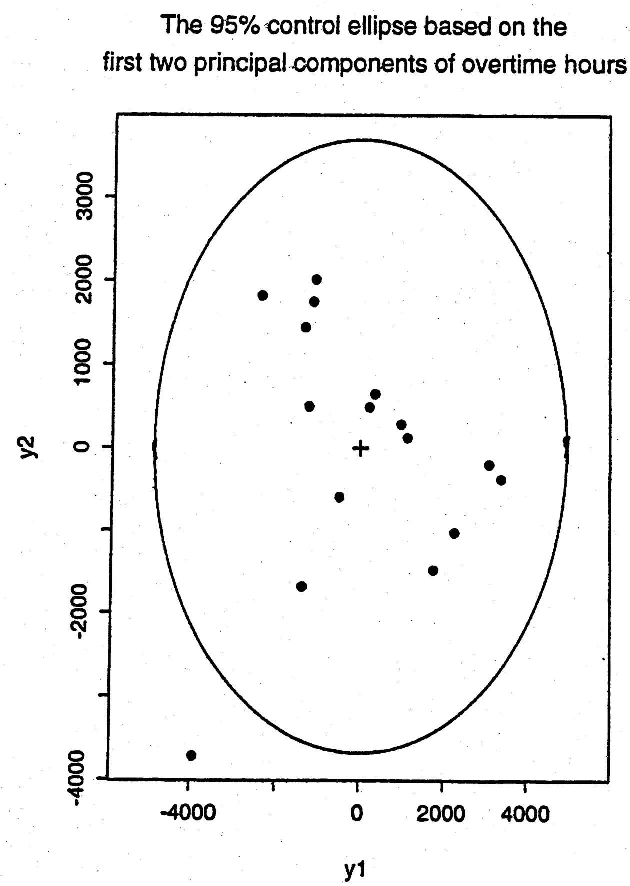
These data can be effectively summarized in one dimension.

c) The results are similar for both the covariance matrix S and the correlation matrix R. The first component in each analysis is a "size" component and almost all of the variation in the data. The analyses differ a bit with respect to the second and remaining components, but these latter components explain very little of the total sample variance.

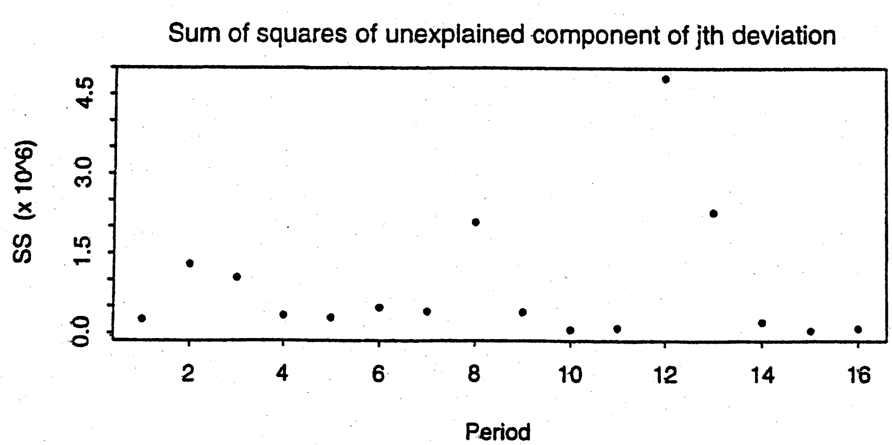
8.24 An ellipse format chart based on the first two principal components of the Madison, Wisconsin, Police Department data

```
XBAR
3557.8 1478.4 2676.9 13563.6 800 7141

S
367884.7   -72093.8    85714.8    222491.4  -44908.3   101312.9
-72093.8  1399053.1    43399.9    139692.2  110517.1  1161018.3
 85714.8    43399.9  1458543.0  -1113809.8  330923.8  1079573.3
222491.4   139692.2 -1113809.8  1698324.4 -244785.9  -462615.6
-44908.3   110517.1   330923.8  -244785.9  224718.0   427767.5
101312.9  1161018.3  1079573.3  -462615.6  427767.5  2488728.4

Eigenvalues of S
4045921.9 2265078.9  761592.1  288919.3  181437.0   94302.6

Eigenvectors of S
-0.0008 -0.0567 -0.5157  0.6122  0.4311 -0.4126
-0.3092 -0.5541  0.5615  0.4932 -0.1796 -0.0810
-0.4821  0.3862 -0.3270  0.3404 -0.5696  0.2667
 0.3675 -0.6415 -0.4898 -0.0642 -0.4308  0.1543
-0.1544  0.0359 -0.0316 -0.3071 -0.4062 -0.8453
-0.7163 -0.3575 -0.2662 -0.4094  0.3269  0.1173

Principal components
       y1      y2      y3     y4     y5     y6
1  1745.4 -1479.3   618.7  222.6    7.2  178.1
2 -1096.6  2011.8   652.5  -69.5  636.9  560.2
3   210.6   490.6   365.8 -899.8 -293.5  -15.2
4 -1360.1  1448.1   420.1  523.5 -972.2   88.5
5 -1255.9   502.1  -422.4 -893.8  359.9 -273.7
6   971.6   284.7  -316.9 -942.8  -83.5  -70.1
7  1118.5   123.7   572.9  319.9  -60.8 -598.5
8 -1151.6  1752.0 -1322.1  700.2 -242.2 -158.8
9  -497.3  -593.0   209.5 -149.2  101.6 -586.2
10 -2397.1 1819.6    -9.5 -147.6 -109.9  207.8
11 -3931.9 -3715.7  924.1   35.1 -274.2  152.9
12 -1392.4 -1688.0 -2285.1  372.1  444.0   85.2
13   326.8   650.8  1251.6  728.8  809.5 -140.0
14  3371.4  -379.1  -499.9 -114.6 -324.3  286.9
15  3076.6  -199.1  -105.7  419.8 -122.3    3.4
16  2261.9 -1029.3   -53.7 -104.5  123.8  279.6
```

$$2.5\times 10^{-7}y_1^2 + 4.4\times 10^{-7}y_2^2 = 5.99$$



8.25 A control chart based on the sum of squares d_{Uj}^2. Period 12 looks unusual.



8.26 (a)-(c) Principal component analysis of the correlation matrix **R**.

Correlations: Indep, Supp, Benev, Conform, Leader

	Indep	Supp	Benev	Conform
Supp	-0.173			
Benev	-0.561	0.018		
Conform	-0.471	-0.327	0.298	
Leader	0.187	-0.401	-0.492	-0.333

Cell Contents: Pearson correlation

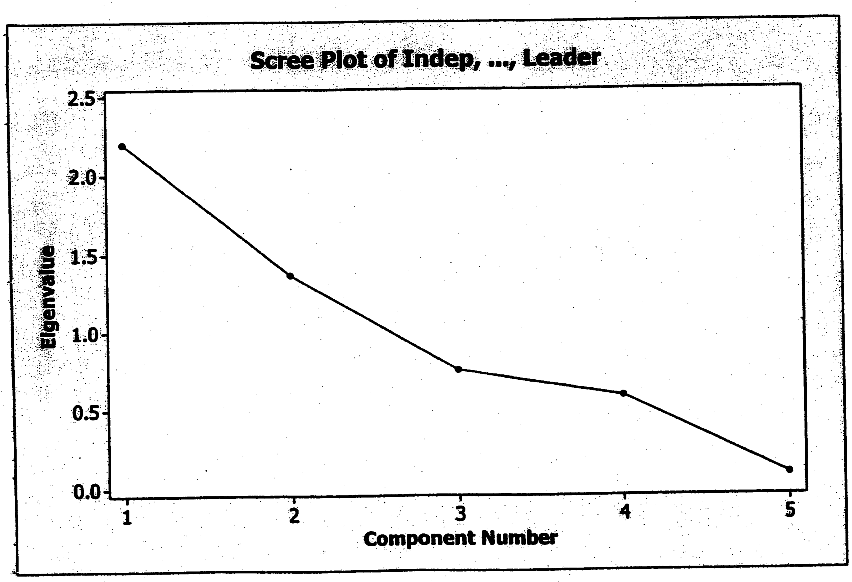
Principal Component Analysis: Indep, Supp, Benev, Conform, Leader

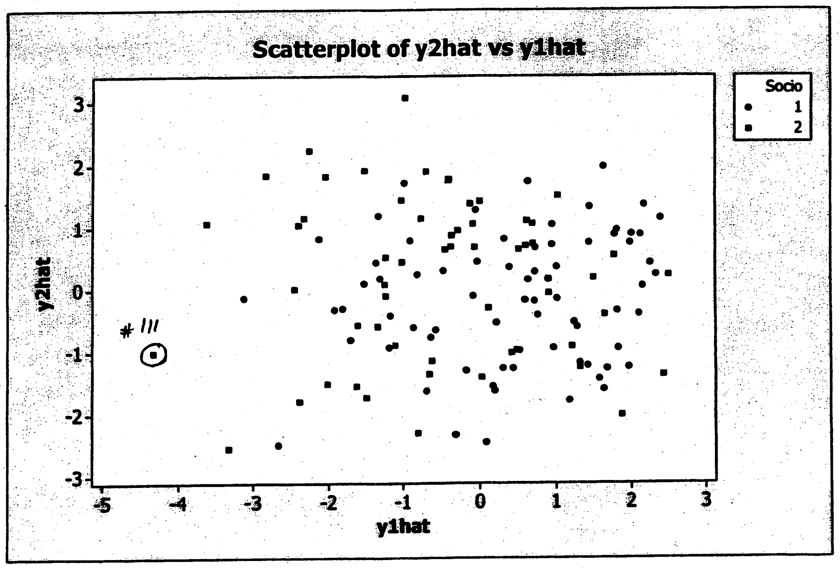
Eigenanalysis of the Correlation Matrix

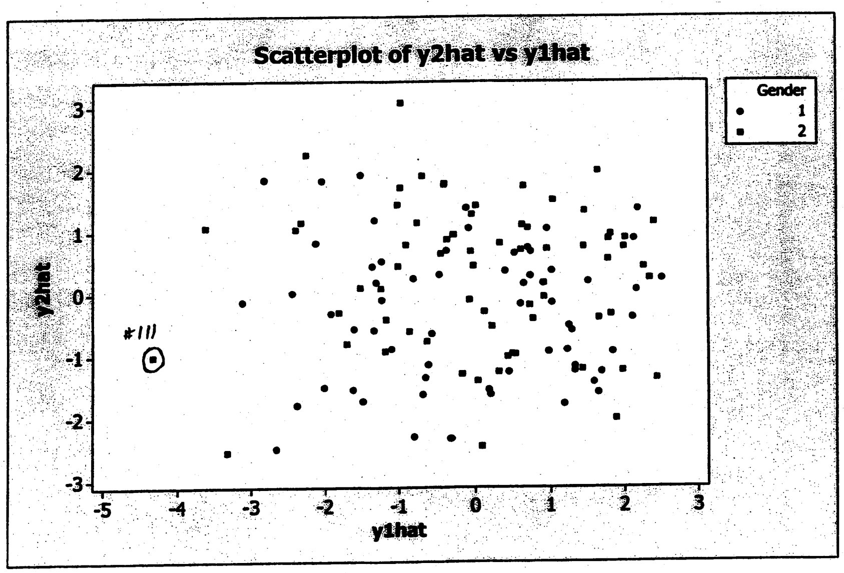
Eigenvalue	2.1966	1.3682	0.7559	0.5888	0.0905
Proportion	0.439	0.274	0.151	0.118	0.018
Cumulative	0.439	0.713	0.864	0.982	1.000

Variable	PC1	PC2	PC3	PC4	PC5
Indep	-0.521	0.087	-0.667	-0.253	-0.460
Supp	0.121	0.788	0.187	0.351	-0.454
Benev	0.548	-0.008	0.115	-0.733	-0.386
Conform	0.439	-0.491	-0.295	0.525	-0.451
Leader	-0.469	-0.361	0.648	0.007	-0.480

Using the scree plot and the proportion of variance explained, it appears as if 4 components should be retained. These components explain almost all (98%) of the variability. It is difficult to provide an interpretation of the components without knowing more about the subject matter. All four of the components represent contrasts of some form. The first component contrasts independence and leadership with benevolence and conformity. The second component contrasts support with conformity and leadership and so on.







The two dimensional plot of the scores on the first two components suggests that the two socioeconomic levels cannot be distinguished from one another nor can the two genders be distinguished. Observation #111 is a bit removed from the rest and might be called an outlier.

(a)-(d) Principal component analysis of the covariance matrix **S**.

Covariances: Indep, Supp, Benev, Conform, Leader

	Indep	Supp	Benev	Conform	Leader
Indep	34.7502				
Supp	-4.2767	17.5134			
Benev	-18.0718	0.4198	29.8447		
Conform	-15.9729	-7.8682	9.3488	33.0426	
Leader	5.7165	-8.7233	-13.9422	-9.9419	26.9580

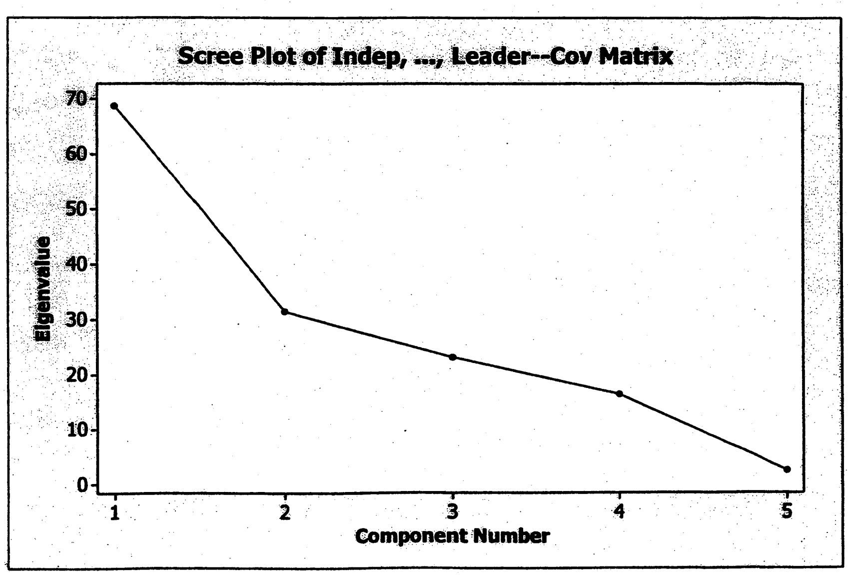
Principal Component Analysis: Indep, Supp, Benev, Conform, Leader

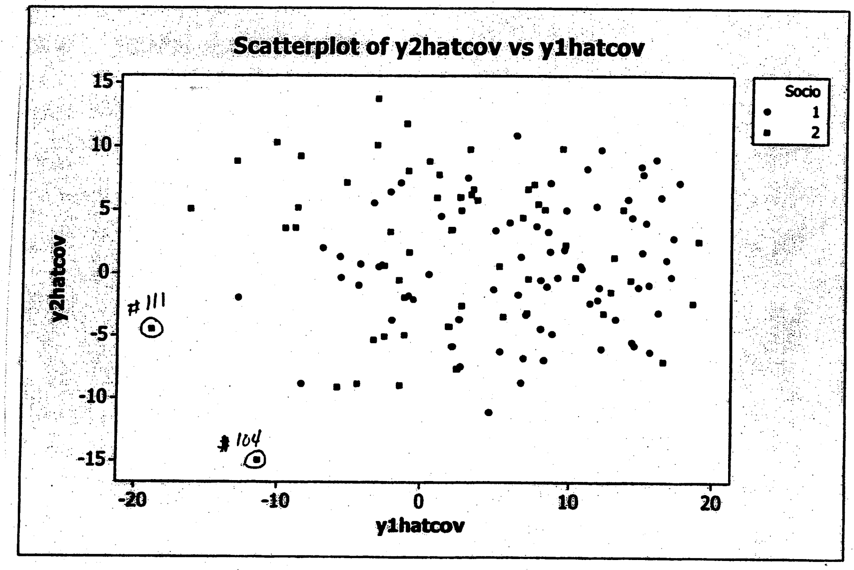
Eigenanalysis of the Covariance Matrix

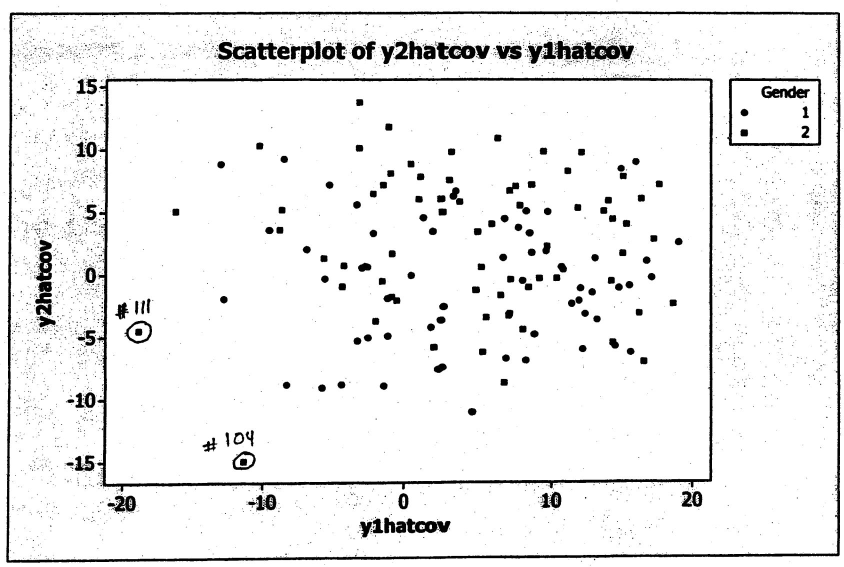
Eigenvalue	68.752	31.509	23.101	16.354	2.392
Proportion	0.484	0.222	0.163	0.115	0.017
Cumulative	0.484	0.706	0.868	0.983	1.000

Variable	PC1	PC2	PC3	PC4	PC5
Indep	-0.579	0.079	-0.643	0.309	0.386
Supp	0.042	0.612	0.140	-0.515	0.583
Benev	0.524	0.219	0.119	0.734	0.352
Conform	0.493	-0.572	-0.422	-0.304	0.398
Leader	-0.380	-0.494	0.612	0.090	0.478

Using the scree plot and the proportion of variance explained, it appears as if 4 components should be retained. These components explain almost all (98%) of the variability. The components are very similar to those obtained from the correlation matrix **R**. All four of the components represent contrasts of some form. The first component contrasts independence and leadership with benevolence and conformity. The second component contrasts support with conformity and leadership and so on. In this case, it makes little difference whether the components are obtained from the sample correlation matrix or the sample covariance matrix.







The two dimensional plot of the scores on the first two components suggests that the two socioeconomic levels cannot be distinguished from one another nor can the two genders be distinguished. Observations #111 and #104 are a bit removed from the rest and might be labeled outliers.

Large sample 95% confidence interval for λ_1:

$$\left(\frac{68.752}{(1+1.96\sqrt{2/130}}, \frac{68.752}{(1-1.96\sqrt{2/130}}\right) = (55.31,\ 90.83)$$

8.27 (a)-(d) Principal component analysis of the correlation matrix **R**.

Correlations: BL, EM, SF, BS

	BL	EM	SF
EM	0.914		
SF	0.984	0.942	
BS	0.988	0.875	0.975

Cell Contents: Pearson correlation

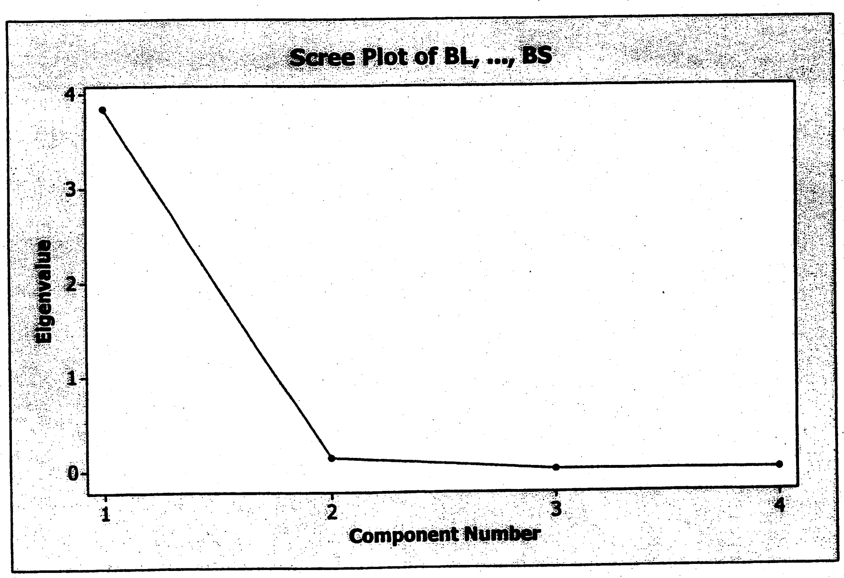
Principal Component Analysis: BL, EM, SF, BS

Eigenanalysis of the Correlation Matrix

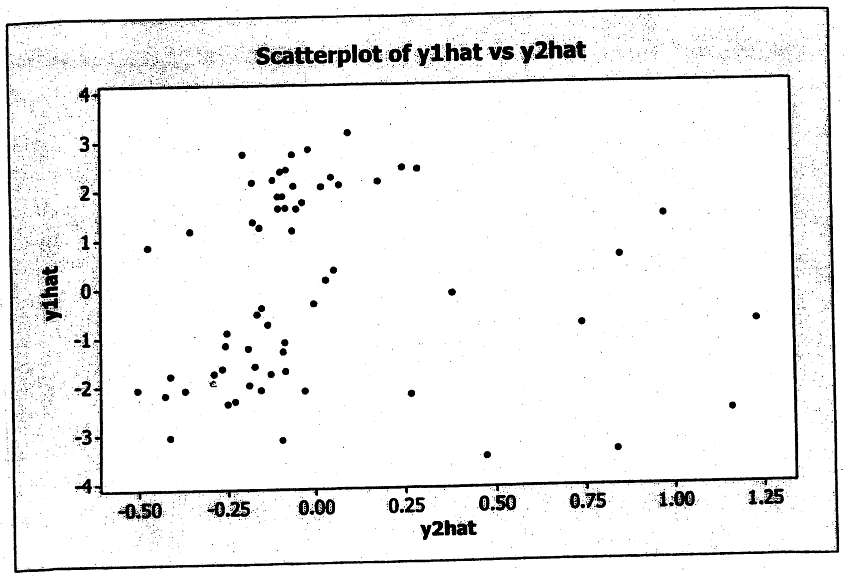
Eigenvalue	3.8395	0.1403	0.0126	0.0076
Proportion	0.960	0.035	0.003	0.002
Cumulative	0.960	0.995	0.998	1.000

Variable	PC1	PC2	PC3	PC4
BL	0.506	-0.261	-0.565	0.597
EM	0.485	0.819	-0.194	-0.237
SF	0.508	-0.020	0.800	0.318
BS	0.500	-0.510	-0.053	-0.698

The proportion of variance explained and the scree plot below suggest that one principal component effectively summarizes the paper properties data. All the variables load about equally on this component so it might be labeled an index of paper strength.



The plot below of the scores on the first two sample principal components does not indicate any obvious outliers.



(a)-(d) Principal component analysis of the covariance matrix **S**.

Covariances: BL, EM, SF, BS

	BL	EM	SF	BS
BL	8.302871			
EM	1.886636	0.513359		
SF	4.147318	0.987585	2.140046	
BS	1.972056	0.434307	0.987966	0.480272

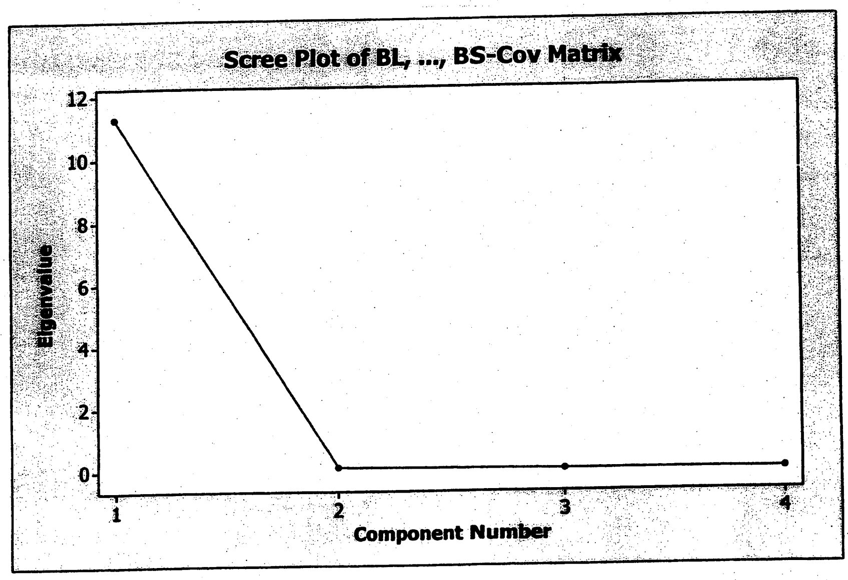
Principal Component Analysis: BL, EM, SF, BS

Eigenanalysis of the Covariance Matrix

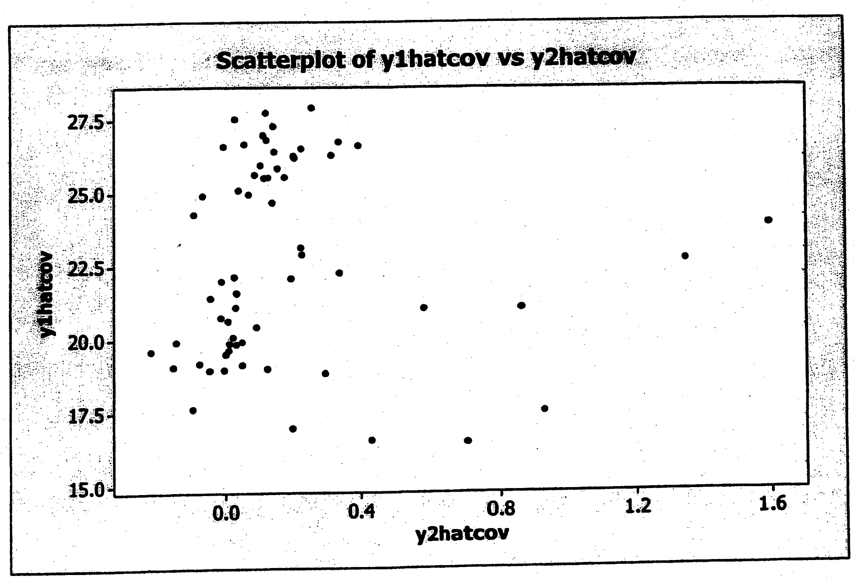
Eigenvalue	11.295	0.104	0.032	0.006
Proportion	0.988	0.009	0.003	0.001
Cumulative	0.988	0.997	0.999	1.000

Variable	PC1	PC2	PC3	PC4
BL	0.856	-0.364	-0.332	0.155
EM	0.198	0.786	-0.497	-0.310
SF	0.431	0.458	0.733	0.259
BS	0.204	-0.201	0.325	-0.901

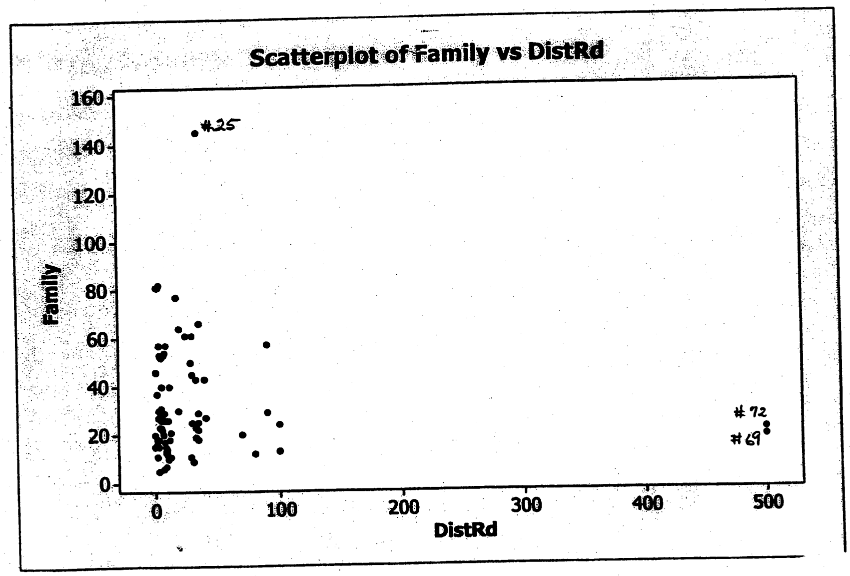
The proportion of variance explained and the scree plot that follows suggest that one principal component effectively summarizes the paper properties data. The loadings of the variables on the first component are all positive, but there are some differences in magnitudes. However, the correlations of the variables with the first component are .998, .928, .990 and .989 for BL, EM, SF and BS respectively. Again, this component might be labeled an index of paper strength.

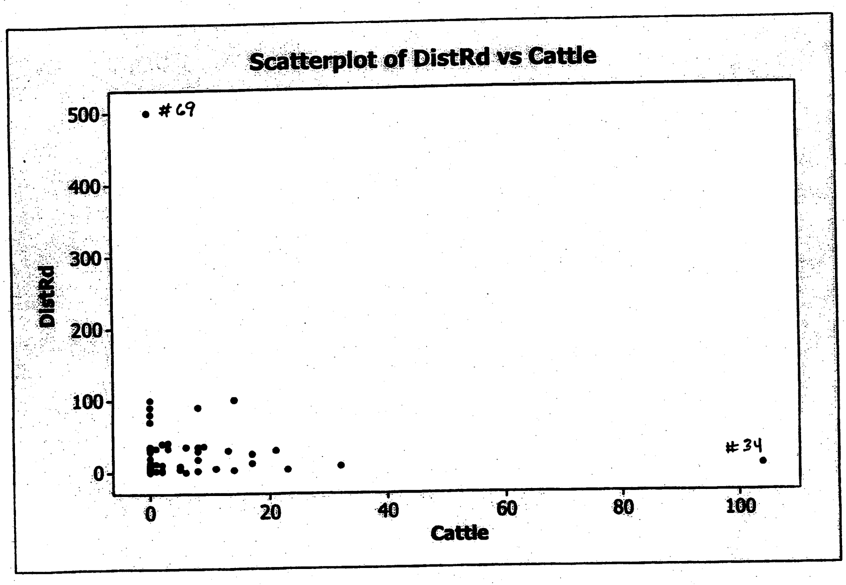


The plot below of the scores on the first two sample principal components does not indicate any obvious outliers.

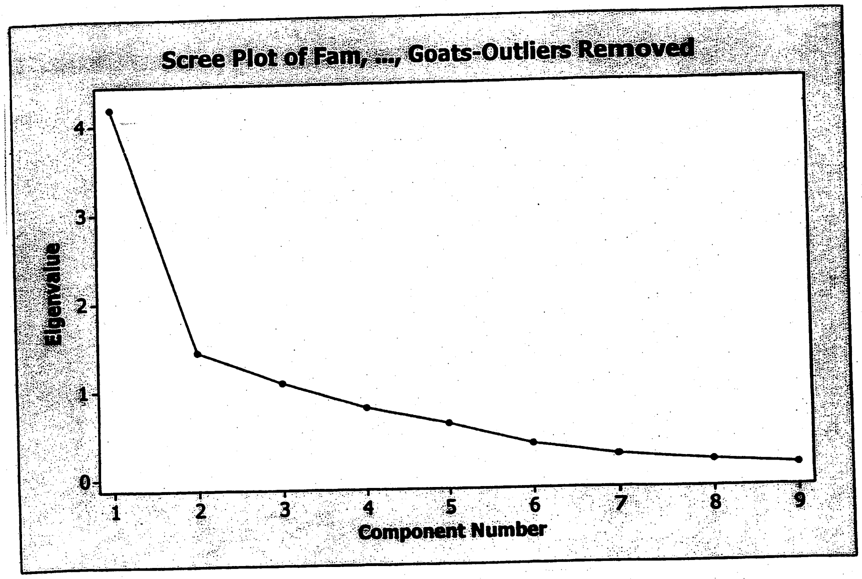


8.28 (a) See scatter plots below. Observations 25, 34, 69 and 72 are outliers.





(b) Principal component analysis of **R** follows. Removing the outliers has some but relatively little effect on the analysis. Five components explain about 90% of the total variability in the data set and seems a reasonable number given the scree plot.



Principal Component Analysis: AdjFam, AdjDistRd, AdjCotton, AdjMaize, AdjSorg,..(Outliers 25,34,69,72 removed)

Eigenanalysis of the Correlation Matrix

Eigenvalue	4.1851	1.4381	1.0845	0.7918	0.6043	0.3661	0.2400	0.1718
Proportion	0.465	0.160	0.121	0.088	0.067	0.041	0.027	0.019
Cumulative	0.465	0.625	0.745	0.833	0.900	0.941	0.968	0.987

Eigenvalue	0.1182
Proportion	0.013
Cumulative	1.000

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
AdjFam	0.434	0.065	0.098	0.171	0.011	-0.040	-0.797	-0.263	-0.249
AdjDistRd	0.008	-0.497	-0.569	0.496	-0.378	0.187	0.021	-0.048	-0.065
AdjCotton	0.446	-0.009	0.132	-0.027	-0.219	-0.200	0.361	0.329	-0.675
AdjMaize	0.352	-0.353	0.388	0.240	-0.079	-0.273	-0.024	0.363	0.574
AdjSorg	0.204	0.604	-0.111	-0.059	-0.645	0.246	-0.021	0.126	0.293
AdjMillet	0.240	0.415	-0.116	0.616	0.527	0.181	0.241	0.077	0.048
AdjBull	0.445	-0.068	-0.030	-0.146	-0.028	-0.134	0.396	-0.751	0.190
AdjCattle	0.355	-0.284	0.014	-0.373	0.218	0.759	-0.011	0.169	0.038
AdjGoats	0.255	0.049	-0.687	-0.351	0.249	-0.402	-0.131	0.274	0.149

Principal Component Analysis: Family, DistRd, Cotton, Maze, Sorg, Millet, Bull, ...

Eigenanalysis of the Correlation Matrix

Eigenvalue	4.1443	1.2364	1.0581	0.9205	0.6058	0.5044	0.2720	0.1470
Proportion	0.460	0.137	0.118	0.102	0.067	0.056	0.030	0.016
Cumulative	0.460	0.598	0.715	0.818	0.885	0.941	0.971	0.988

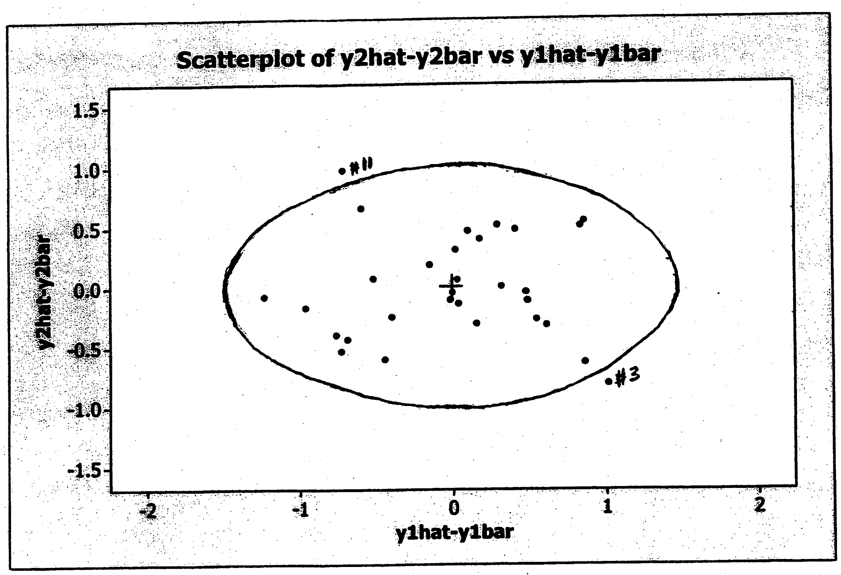
Eigenvalue	0.1114
Proportion	0.012
Cumulative	1.000

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Family	0.444	-0.100	-0.002	-0.123	-0.089	-0.127	-0.579	0.454	-0.461
DistRd	-0.033	-0.072	-0.831	0.502	-0.194	-0.051	-0.045	0.082	0.041
Cotton	0.411	-0.342	-0.068	0.030	0.100	-0.216	0.509	-0.372	-0.504
Maze	0.337	-0.554	0.170	0.164	-0.134	0.053	-0.352	-0.360	0.499
Sorg	0.311	0.452	-0.069	-0.229	-0.361	-0.632	0.055	-0.139	0.300
Millet	0.269	0.043	-0.385	-0.606	-0.182	0.594	0.089	-0.097	0.077
Bull	0.440	-0.029	0.122	0.197	0.129	0.110	0.458	0.621	0.357
Cattle	0.247	0.458	0.278	0.486	-0.392	0.407	-0.012	-0.215	-0.225
Goats	0.309	0.379	-0.173	0.100	0.770	0.043	-0.242	-0.242	0.095

(c) All the variables (all crops, all livestock, family) except for distance to road (DistRd) load about equally on the first component. This component might be called a farm size component. Millet and sorghum load positively and distance to road and maize load negatively on the second component. Without additional subject matter knowledge, this component is difficult to interpret. The third component is essentially a distance to the road and goats component. This component might represent subsistence farms. The fourth component appears to be a contrast between distance to road and millet versus cattle and goats. Again, this component is difficult to interpret. The fifth component appears to contrast sorghum with millet.

8.29 (a) The 95% ellipse format chart using the first two principal components from the covariance matrix **S** (for the first 30 cases of the car body assembly data) is shown below. The ellipse consists of all $\hat{y}_1, \hat{y}_2$ such that $\frac{\hat{y}_1^2}{\hat{\lambda}_1} + \frac{\hat{y}_2^2}{\hat{\lambda}_2} \le \chi_2^2(.05) = 5.99$

where $\hat{\lambda}_1 = .354, \hat{\lambda}_2 = .186$. Observations 3 and 11 lie outside the ellipse.



(b) To construct the alternative control chart based upon unexplained components of the observations we note that $\bar{d}_U^2 = .4137$, $s_{d^2}^2 = .0782$ so

$c = \frac{.0782}{2(.4137)} = .0946$, $\upsilon = 2\frac{(.4137)^2}{.0782} = 4.4$. Conservatively, we set the chi-squared degrees of freedom to $\upsilon = 5$ and the UCL becomes $c\chi_5^2(.05) = .0946(11.07) = 1.05$ or approximately 1.0. The alternative control chart is plotted on the next page and it appears as if multivariate observation 18 is out of control. For observation 18, $\hat{y}_4^2$ makes the largest contribution to d_{U18}^2 and the variables getting the most weight in $\hat{y}_4$ are the thickness measurements x_1 and x_2. Car body #18 could be examined at locations 1 and 2 to determine the cause of the unusual deviations in thickness from the nominal levels.