UC DAVIS

STA 135

HW 5 SOLUTIONS

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

8.6 (a)

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

$\hat{y}_2 = -.041x_1 + .999x_2$ 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)

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

$\hat{y}_2 = .707z_1 - .707z_2$ 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.

8.18 (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

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

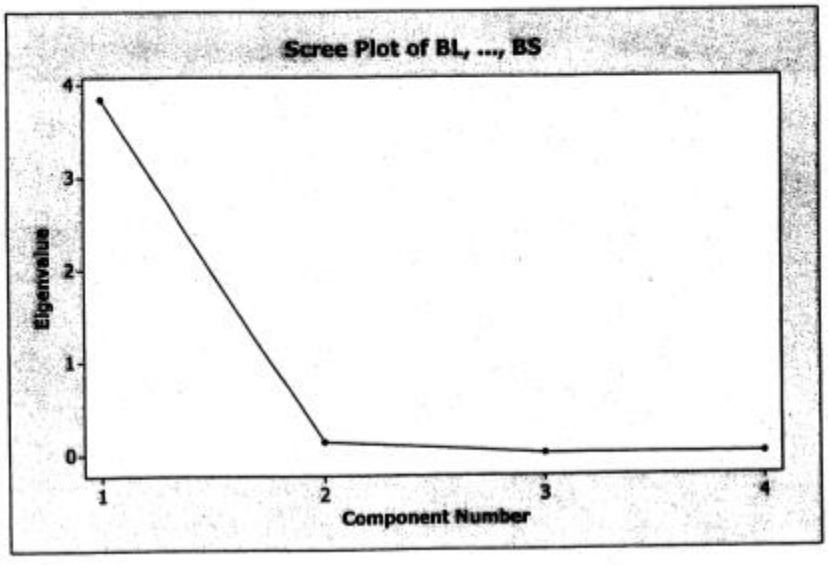
Correlations: BL, EM, SF, BS

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

Cell Contents: Pearson correlation
```

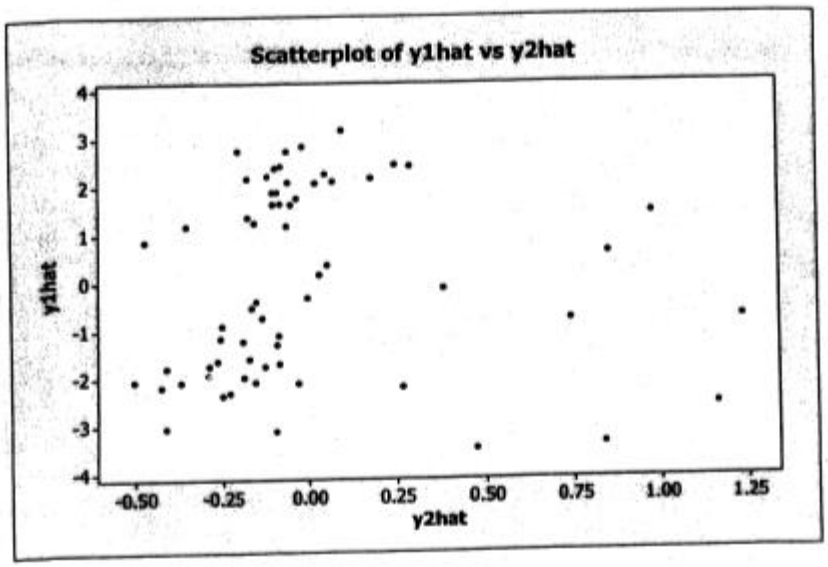
Principal Component Analysis: BL, EM, SF, BS

```
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.



9.1

$$L' = [.9 \quad .7 \quad .5]; \qquad LL' = \begin{bmatrix} .81 & .63 & .45 \\ .63 & .49 & .35 \\ .45 & .35 & .25 \end{bmatrix}$$

so $\rho = LL' + \Psi$

9.2 a) For $m = 1$ $\quad h_1^2 = \ell_{11}^2 = .81$

$h_2^2 = \ell_{21}^2 = .49$

$h_3^2 = \ell_{31}^2 = .25$

The communalities are those parts of the variances of the variables explained by the single factor.

b) $\text{Corr}(Z_i, F_1) = \text{Cov}(Z_i, F_1)$, $i = 1,2,3$. By (9-5) $\text{Cov}(Z_i, F_1) = \ell_{i1}$. Thus $\text{Corr}(Z_1, F_1) = \ell_{11} = .9$; $\text{Corr}(Z_2, F_1) = \ell_{21} = .7$; $\text{Corr}(Z_3, F_1) = \ell_{31} = .5$. The first variable, Z_1, has the largest correlation with the factor and therefore will probably carry the most weight in naming the factor.