MVA HW #5

8.4. Find the principal components and the proportion of the total population variance explained by each when the covariance matrix is

$$\Sigma = \begin{bmatrix} \sigma^2 & \sigma^2\rho & 0 \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho \\ 0 & \sigma^2\rho & \sigma^2 \end{bmatrix}, \quad -\frac{1}{\sqrt{2}} < \rho < \frac{1}{\sqrt{2}}$$

8.6. Data on x_1 = sales and x_2 = profits for the 10 largest companies in the world were listed in Exercise 1.4 of Chapter 1.
From Example 4.12

$$\bar{\mathbf{x}} = \begin{bmatrix} 155.60 \\ 14.70 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} 7476.45 & 303.62 \\ 303.62 & 26.19 \end{bmatrix}$$

(a) Determine the sample principal components and their variances for these data. (You may need the quadratic formula to solve for the eigenvalues of $\mathbf{S}$.)

(b) Find the proportion of the total sample variance explained by $\hat{y}_1$.

(c) Sketch the constant density ellipse $(\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) = 1.4$, and indicate the principal components $\hat{y}_1$ and $\hat{y}_2$ on your graph.

(d) Compute the correlation coefficients $r_{\hat{y}_1, x_k}, k = 1, 2$. What interpretation, if any, can you give to the first principal component?

8.7. Convert the covariance matrix **S** in Exercise 8.6 to a sample correlation matrix **R**.

(a) Find the sample principal components $\hat{y}_1$, $\hat{y}_2$ and their variances.

(b) Compute the proportion of the total sample variance explained by $\hat{y}_1$.

(c) Compute the correlation coefficients $r_{\hat{y}_1, z_k}$, $k = 1, 2$. Interpret $\hat{y}_1$.

(d) Compare the components obtained in Part a with those obtained in Exercise 8.6(a). Given the original data displayed in Exercise 1.4, do you feel that it is better to determine principal components from the sample covariance matrix or sample correlation matrix? Explain.

8.13. In the radiotherapy data listed in Table 1.7 (see also the radiotherapy data on the website www.prenhall.com/statistics), the $n = 98$ observations on $p = 6$ variables represent patients' reactions to radiotherapy.

(a) Obtain the covariance and correlation matrices **S** and **R** for these data.

(b) Pick one of the matrices **S** or **R** (justify your choice), and determine the eigenvalues and eigenvectors. Prepare a table showing, in decreasing order of size, the percent that each eigenvalue contributes to the total sample variance.

(c) Given the results in Part b, decide on the number of important sample principal components. Is it possible to summarize the radiotherapy data with a single reaction-index component? Explain.

(d) Prepare a table of the correlation coefficients between each principal component you decide to retain and the original variables. If possible, interpret the components.

Table 1.7 Radiotherapy Data

x_1 Symptoms	x_2 Activity	x_3 Sleep	x_4 Eat	x_5 Appetite	x_6 Skin reaction
.889	1.389	1.555	2.222	1.945	1.000
2.813	1.437	.999	2.312	2.312	2.000
1.454	1.091	2.364	2.455	2.909	3.000
.294	.941	1.059	2.000	1.000	1.000
2.727	2.545	2.819	2.727	4.091	.000
⋮	⋮	⋮	⋮	⋮	⋮
4.100	1.900	2.800	2.000	2.600	2.000
.125	1.062	1.437	1.875	1.563	.000
6.231	2.769	1.462	2.385	4.000	2.000
3.000	1.455	2.090	2.273	3.272	2.000
.889	1.000	1.000	2.000	1.000	2.000

Source: Data courtesy of Mrs. Annette Tealey, R.N. Values of x_2 and x_3 less than 1.0 are due to errors in the data-collection process. Rows containing values of x_2 and x_3 less than 1.0 may be omitted.

8.16. Over a period of five years in the 1990s, yearly samples of fishermen on 28 lakes in Wisconsin were asked to report the time they spent fishing and how many of each type of game fish they caught. Their responses were then converted to a catch rate per hour for

x_1 = Bluegill x_2 = Black crappie x_3 = Smallmouth bass

x_4 = Largemouth bass x_5 = Walleye x_6 = Northern pike

The estimated correlation matrix (courtesy of Jodi Barnet)

$$\mathbf{R} = \begin{bmatrix} 1 & .4919 & .2636 & .4653 & -.2277 & .0652 \\ .4919 & 1 & .3127 & .3506 & -.1917 & .2045 \\ .2635 & .3127 & 1 & .4108 & .0647 & .2493 \\ .4653 & .3506 & .4108 & 1 & -.2249 & .2293 \\ -.2277 & -.1917 & .0647 & -.2249 & 1 & -.2144 \\ .0652 & .2045 & .2493 & .2293 & -.2144 & 1 \end{bmatrix}$$

is based on a sample of about 120. (There were a few missing values.)

Fish caught by the same fisherman live alongside of each other, so the data should provide some evidence on how the fish group. The first four fish belong to the centrarchids, the most plentiful family. The walleye is the most popular fish to eat.

(a) Comment on the pattern of correlation within the centrarchid family x_1 through x_4. Does the walleye appear to group with the other fish?

(b) Perform a principal component analysis using only x_1 through x_4. Interpret your results.

(c) Perform a principal component analysis using all six variables. Interpret your results.

9.1. Show that the covariance matrix

$$\boldsymbol{\rho} = \begin{bmatrix} 1.0 & .63 & .45 \\ .63 & 1.0 & .35 \\ .45 & .35 & 1.0 \end{bmatrix}$$

for the $p = 3$ standardized random variables Z_1, Z_2, and Z_3 can be generated by the $m = 1$ factor model

$$\begin{aligned} Z_1 &= .9F_1 + \varepsilon_1 \\ Z_2 &= .7F_1 + \varepsilon_2 \\ Z_3 &= .5F_1 + \varepsilon_3 \end{aligned}$$

where $\text{Var}(F_1) = 1$, $\text{Cov}(\boldsymbol{\varepsilon}, F_1) = \mathbf{0}$, and

$$\boldsymbol{\Psi} = \text{Cov}(\boldsymbol{\varepsilon}) = \begin{bmatrix} .19 & 0 & 0 \\ 0 & .51 & 0 \\ 0 & 0 & .75 \end{bmatrix}$$

That is, write $\boldsymbol{\rho}$ in the form $\boldsymbol{\rho} = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}$.

9.18. Refer to Exercise 8.16 concerning the numbers of fish caught.

(a) Using only the measurements $x_1 - x_4$, obtain the principal component solution for factor models with $m = 1$ and $m = 2$.

(b) Using only the measurements $x_1 - x_4$, obtain the maximum likelihood solution for factor models with $m = 1$ and $m = 2$.

(c) Rotate your solutions in Parts (a) and (b). Compare the solutions and comment on them. Interpret each factor.

(d) Perform a factor analysis using the measurements $x_1 - x_6$. Determine a reasonable number of factors m, and compare the principal component and maximum likelihood solutions after rotation. Interpret the factors.

11.1. Consider the two data sets

$$\mathbf{X}_1 = \begin{bmatrix} 3 & 7 \\ 2 & 4 \\ 4 & 7 \end{bmatrix} \quad \text{and} \quad \mathbf{X}_2 = \begin{bmatrix} 6 & 9 \\ 5 & 7 \\ 4 & 8 \end{bmatrix}$$

for which

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}, \quad \bar{\mathbf{x}}_2 = \begin{bmatrix} 5 \\ 8 \end{bmatrix}$$

and

$$\mathbf{S}_{\text{pooled}} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

(a) Calculate the linear discriminant function in (11-19).

(b) Classify the observation $\mathbf{x}_0' = [2 \quad 7]$ as population π_1 or population π_2, using Rule (11-18) with equal priors and equal costs.

11.3. Prove Result 11.1.

Hint: Substituting the integral expressions for $P(2|1)$ and $P(1|2)$ given by (11-1) and (11-2), respectively, into (11-5) yields

$$\text{ECM} = c(2|1)p_1 \int_{R_2} f_1(\mathbf{x})\, d\mathbf{x} + c(1|2)p_2 \int_{R_1} f_2(\mathbf{x}) d\mathbf{x}$$

Noting that $\Omega = R_1 \cup R_2$, so that the total probability

$$1 = \int_{\Omega} f_1(\mathbf{x})\, d\mathbf{x} = \int_{R_1} f_1(\mathbf{x})\, d\mathbf{x} + \int_{R_2} f_1(\mathbf{x})\, d\mathbf{x}$$

we can write

$$\text{ECM} = c(2|1)p_1\left[1 - \int_{R_1} f_1(\mathbf{x})\, d\mathbf{x}\right] + c(1|2)p_2 \int_{R_1} f_2(\mathbf{x})\, d\mathbf{x}$$

By the additive property of integrals (volumes),

$$\text{ECM} = \int_{R_1} [c(1|2)p_2 f_2(\mathbf{x}) - c(2|1)p_1 f_1(\mathbf{x})]\, d\mathbf{x} + c(2|1)p_1$$

Now, $p_1, p_2, c(1|2)$, and $c(2|1)$ are nonnegative. In addition, $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are nonnegative for all $\mathbf{x}$ and are the only quantities in ECM that depend on $\mathbf{x}$. Thus, ECM is minimized if R_1 includes those values $\mathbf{x}$ for which the integrand

$$[c(1|2)p_2 f_2(\mathbf{x}) - c(2|1)p_1 f_1(\mathbf{x})] \leq 0$$

and excludes those $\mathbf{x}$ for which this quantity is positive.

Result 11.1. The regions R_1 and R_2 that minimize the ECM are defined by the values $\mathbf{x}$ for which the following inequalities hold:

$$R_1: \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$\begin{pmatrix}\text{density}\\ \text{ratio}\end{pmatrix} \geq \begin{pmatrix}\text{cost}\\ \text{ratio}\end{pmatrix}\begin{pmatrix}\text{prior}\\ \text{probability}\\ \text{ratio}\end{pmatrix} \qquad (11\text{-}6)$$

$$R_2: \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$\begin{pmatrix}\text{density}\\ \text{ratio}\end{pmatrix} < \begin{pmatrix}\text{cost}\\ \text{ratio}\end{pmatrix}\begin{pmatrix}\text{prior}\\ \text{probability}\\ \text{ratio}\end{pmatrix}$$

11.4. A researcher wants to determine a procedure for discriminating between two multivariate populations. The researcher has enough data available to estimate the density functions $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ associated with populations π_1 and π_2, respectively. Let $c(2|1) = 50$ (this is the cost of assigning items as π_2, given that π_1 is true) and $c(1|2) = 100$.

In addition, it is known that about 20% of all possible items (for which the measurements $\mathbf{x}$ can be recorded) belong to π_2.

(a) Give the minimum ECM rule (in general form) for assigning a new item to one of the two populations.

(b) Measurements recorded on a new item yield the density values $f_1(\mathbf{x}) = .3$ and $f_2(\mathbf{x}) = .5$. Given the preceding information, assign this item to population π_1 or population π_2.

11.23. Consider the data given in Exercise 1.14.

(a) Check the marginal distributions of the x_i's in both the multiple-sclerosis (MS) group and non-multiple-sclerosis (NMS) group for normality by graphing the corresponding observations as normal probability plots. Suggest appropriate data transformations if the normality assumption is suspect.

(b) Assume that $\Sigma_1 = \Sigma_2 = \Sigma$. Construct Fisher's linear discriminant function. Do all the variables in the discriminant function appear to be important? Discuss your answer. Develop a classification rule assuming equal prior probabilities and equal costs of misclassification.

(c) Using the results in (b), calculate the apparent error rate. If computing resources allow, calculate an estimate of the expected actual error rate using Lachenbruch's holdout procedure. Compare the two error rates.

Table 1.6 Multiple-Sclerosis Data

Non-Multiple-Sclerosis Group Data

Subject number	x_1 (Age)	x_2 ($S1L + S1R$)	x_3 $\lvert S1L - S1R \rvert$	x_4 ($S2L + S2R$)	x_5 $\lvert S2L - S2R \rvert$
1	18	152.0	1.6	198.4	.0
2	19	138.0	.4	180.8	1.6
3	20	144.0	.0	186.4	.8
4	20	143.6	3.2	194.8	.0
5	20	148.8	.0	217.6	.0
⋮	⋮	⋮	⋮	⋮	⋮
65	67	154.4	2.4	205.2	6.0
66	69	171.2	1.6	210.4	.8
67	73	157.2	.4	204.8	.0
68	74	175.2	5.6	235.6	.4
69	79	155.0	1.4	204.4	.0

Multiple-Sclerosis Group Data

Subject number	x_1	x_2	x_3	x_4	x_5
1	23	148.0	.8	205.4	.6
2	25	195.2	3.2	262.8	.4
3	25	158.0	8.0	209.8	12.2
4	28	134.4	.0	198.4	3.2
5	29	190.2	14.2	243.8	10.6
⋮	⋮	⋮	⋮	⋮	⋮
25	57	165.6	16.8	229.2	15.6
26	58	238.4	8.0	304.4	6.0
27	58	164.0	.8	216.8	.8
28	58	169.8	.0	219.2	1.6
29	59	199.8	4.6	250.2	1.0

Source: Data courtesy of Dr. G. G. Celesia.