STAT 636, Fall 2015 - Assignment 6
SOLUTIONS

1. For the stock price data in textbook Table 8.4 ($n = 103$, $p = 5$):

   (a) Using the sample covariance matrix $\mathbf{S}$, find the sample principal components $\mathbf{y}_1 = \mathbf{X}\mathbf{a}_1$, $\mathbf{y}_2 = \mathbf{X}\mathbf{a}_2$, ..., $\mathbf{y}_p = \mathbf{X}\mathbf{a}_p$, where $\mathbf{X}$ is the $n \times p$ matrix of stock prices. Print the first five rows of $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_p]$.
   SEE CODE BELOW. HERE ARE THE FIRST FIVE ROWS OF $\mathbf{Y}$:

	PC1	PC2	PC3	PC4	PC5
[1,]	0.025211170	0.011998822	-0.0384604565	-0.0014851773	0.003473245
[2,]	-0.022477337	0.005379563	0.0001338018	0.0025744885	0.013159477
[3,]	0.009091407	-0.010694384	0.0080312540	-0.0119552884	-0.011384079
[4,]	0.016363352	0.026824644	-0.0211235548	-0.0002609519	-0.014600282
[5,]	-0.047307028	-0.015030242	-0.0106928576	0.0108208650	0.008328983

   (b) Determine the proportion of the total sample variance explained by the first three principal components. Interpret these components.
   THE FIRST THREE PCS EXPLAIN 52.9%, 27.1%, AND 9.8% OF THE TOTAL SAMPLE VARIANCE, RESPECTIVELY. HERE ARE THE COEFFICIENTS OF THE LINEAR COMBINATIONS THAT DEFINE THESE PCS:

	PC1	PC2	PC3
JPM	-0.2228228	0.6252260	-0.32611218
C	-0.3072900	0.5703900	0.24959014
WF	-0.1548103	0.3445049	0.03763929
RDS	-0.6389680	-0.2479475	0.64249741
EM	-0.6509044	-0.3218478	-0.64586064

   THE FIRST PC IS BASICALLY A WEIGHTED AVERAGE OF THE STOCK PRICES, WITH HIGHEST WEIGHTS ASSIGNED TO RDS AND EM. WE MIGHT VIEW THIS PC AS AN OVERALL "INDEX" PRICE. THE SECOND PC IS A DIFFERENCE BETWEEN JPM, C, WF AND RDS, EM. THE FIRST THREE STOCKS ARE FOR BANK COMPANIES, WHILE THE LAST TWO ARE FOR OIL COMPANIES. THUS, WE MIGHT VIEW THIS PC AS AN "INDUSTRY" INDICATOR. THE THIRD PC IS ROUGHLY A DIFFERENCE BETWEEN C, RDS AND JPM, EM. THIS IS NOT AS EASY TO INTERPRET.

2. For the census-tract data in Table 8.5 ($n = 61$, $p = 5$), convert median home value to be measured in thousands rather than hundred thousands (multiply that column by 100).

   (a) Carry out two principal component analyses, one based on the sample covariance matrix and the other based on the sample correlation matrix, using the modified data. For each analysis, report the proportions of total sample variance explained by the first three PCs as well as the correlation between these PCs and the individual variables.
   THE COEFFICIENTS OF THE LINEAR COMBINATIONS THAT DEFINE THE FIRST THREE PCS FOR EACH METHOD ARE SHOWN IN TABLE 1. THE CORRESPONDING CORRELATIONS BETWEEN THE PCS AND THE INDIVIDUAL VARIABLES ARE SHOWN IN TABLE 2.

	$\mathbf{S}$			$\boldsymbol{\rho}$		
% explained	0.9551	0.0311	0.0116	0.3984	0.2735	0.1728
`tot_pop`	0.0008	-0.0411	0.0706	0.2626	-0.4630	0.7839
`prof_deg`	0.0378	0.0696	0.0739	-0.5934	-0.3256	-0.1641
`emp`	-0.0017	-0.5131	0.8540	0.3257	-0.6051	-0.2249
`gov_emp`	0.0309	0.8540	0.5098	-0.4792	0.2525	0.5507
`med_home`	0.9988	-0.0299	-0.0172	-0.4932	-0.4996	-0.0688

Table 1: Coefficients that define the first three PCs for each method, as well as the proportions of variance explained.

	$\mathbf{S}$			$\boldsymbol{\rho}$		
`tot_pop`	0.0250	0.2277	-0.2381	0.3706	0.5414	0.7287
`prof_deg`	0.6877	-0.2284	-0.1478	-0.8374	0.3808	-0.1525
`emp`	-0.0128	0.7018	-0.7120	0.4597	0.7077	-0.2090
`gov_emp`	0.1853	-0.9232	-0.3359	-0.6763	-0.2953	0.5119
`med_home`	1.0000	0.0054	0.0019	-0.6961	0.5843	-0.0640

Table 2: Correlations between the first three PCs and the individual variables.

In the analysis based on $\mathbf{S}$, the first PC explains about 96% of the total variance and is basically equal to median home value. In the analysis based on $\boldsymbol{\rho}$, the first PC explains about 40% of the total variance, represents a difference between `tot_pop`, `emp` and `prof_deg`, `gov_emp`, `med_home`, and is moderately correlated with most of the variables. When we based PCA on $\mathbf{S}$, the results were heavily driven by home values, because this variable was the dominant source of variability on all its own. When the variables are not on similar scales, it is arguably best to first standardize them (equivalent to "PCA on the correlation matrix").

(b) Interpret the PCs. Comment on how the interpretations differ between the two analyses. Which of the two would you recommend using, and why?

See comments on previous problem.

3. For the bull data in Table 1.10 ($n = 76$, $p = 8$):

(a) Perform PCA on the standardized variables (i.e., based on the sample correlation matrix), excluding the variable `Breed`. What proportions of variability do the PCs explain?

The first five PCs explain 95% of the variance, with PC 1 explaining 53% and PC 2 explaining 21%.

(b) Report a scree plot and comment on how many PCs are minimally necessary to adequately represent $\mathbf{X}$.

The scree plot is in Figure 1. Based on the picture, I would say 4 PCs looks like the best choice. The remaining PCs do not contribute much.

(c) Interpret the first two PCs.

The first PC is roughly a weighted average of all the variables. The second PC is roughly a difference between `PrctFFB` and `SalePr`, `BkFat`, `SaleWt`. Roughly speaking, then, the dominant pattern in these data is due to differences in their differences on all 8 variables. Similarly, the second largest source of variability is due to differences between fatter and leaner bulls.

(d) Report a scatterplot of the first two PCs, with the points color-coded by `Breed`. Can you distinguish groups representing the three breeds of cattle? Are there any outliers?
See Figure 2. There is some separation between the three breeds, though it is not particularly strong. I don't see any obvious outliers.

```
####
#### (1)
####

## Load data.
X <- as.matrix(read.table("T8-4.DAT", header = FALSE))
colnames(X) <- c("JPM", "C", "WF", "RDS", "EM")

## Summary statistics.
x_bar <- colMeans(X)
S <- var(X)

##
## PCA.
##

## Can be done with 'prcomp' function. Matches the output of 'eigen'; the standard
## deviations returned by 'prcomp' are the square roots of the eigenvalues of S.
pca <- prcomp(X)

ee <- eigen(S)
lambda <- ee$values
ee <- ee$vectors

## Sample principal components.
Y <- X %*% pca$rotation

## The first three PCs account for 52.9%, 27.1%, and 9.8% of the total variation,
## respectively.
pca$sdev ^ 2 / sum(pca$sdev ^ 2)

####
#### (2)
####

## Load data.
X <- as.matrix(read.delim("T8-5.DAT", header = FALSE))
colnames(X) <- c("tot_pop", "prof_deg", "emp", "gov_emp", "med_home")

## Convert median home value to be in thousands, rather than hundred thousands.
X_new <- X
X_new[, "med_home"] <- X_new[, "med_home"] * 100

## Summary statistics.
x_bar_orig <- colMeans(X)
```

```
S_orig <- var(X)
x_bar_new <- colMeans(X_new)
S_new <- var(X_new)
R_new <- cor(X_new)

##
## PCA.
##

## On original variables, using the sample covariance matrix.
pca_orig <- prcomp(X); ee_orig <- eigen(S_orig)
pca_orig$sdev ^ 2 / sum(pca_orig$sdev ^ 2)

corr_orig <- cbind(ee_orig$vectors[, 1] * sqrt(ee_orig$values[1] / diag(S_orig)),
  ee_orig$vectors[, 2] * sqrt(ee_orig$values[2] / diag(S_orig)),
  ee_orig$vectors[, 3] * sqrt(ee_orig$values[3] / diag(S_orig)))

## Now with the modified home value variable, using the sample covariance matrix.
pca_new_S <- prcomp(X_new); ee_new_S <- eigen(S_new)
pca_new_S$sdev ^ 2 / sum(pca_new_S$sdev ^ 2)

corr_new_S <- cbind(ee_new_S$vectors[, 1] * sqrt(ee_new_S$values[1] / diag(S_new)),
  ee_new_S$vectors[, 2] * sqrt(ee_new_S$values[2] / diag(S_new)),
  ee_new_S$vectors[, 3] * sqrt(ee_new_S$values[3] / diag(S_new)))

## With modified home value variable, using the sample correlation matrix.
pca_new_R <- prcomp(X_new, center = TRUE, scale = TRUE); ee_new_R <- eigen(R_new)
pca_new_R$sdev ^ 2 / sum(pca_new_R$sdev ^ 2)

corr_new_R <- cbind(ee_new_R$vectors[, 1] * sqrt(ee_new_R$values[1] / diag(R_new)),
  ee_new_R$vectors[, 2] * sqrt(ee_new_R$values[2] / diag(R_new)),
  ee_new_R$vectors[, 3] * sqrt(ee_new_R$values[3] / diag(R_new)))

out_1 <- cbind(rbind(pca_new_S$sdev ^ 2 / sum(pca_new_S$sdev ^ 2), pca_new_S$rotation),
  rbind(pca_new_R$sdev ^ 2 / sum(pca_new_R$sdev ^ 2), pca_new_R$rotation))
out_1 <- round(out_1[, c(1:3, 6:8)], 4)

out_2 <- round(cbind(corr_new_S, corr_new_R), 4)

####
#### (3)
####

## Load data.
X <- as.matrix(read.table("T1-10.DAT", header = FALSE))
```

```
colnames(X) <- c("Breed", "SalePr", "YrHgt", "FtFrBody", "PrctFFB", "Frame", "BkFat",
  "SaleHt", "SaleWt")

Z <- factor(X[, "Breed"])
X <- X[, -1]

n <- nrow(X)
p <- ncol(X)

## Summary statistics.
x_bar <- colMeans(X)
S <- var(X)
R <- cor(X)

##
## PCA on the standardized variables.
##

pca <- prcomp(X, center = TRUE, scale = TRUE)
summary(pca)

PCs <- scale(X, center = TRUE, scale = TRUE) %*% pca$rotation

## Same as eigen-analysis of R.
ee <- eigen(R)
lambda <- ee$values
ee <- ee$vectors

## Scree plot.
pdf("figures/scree.pdf")
plot(1:p, eigen(R)$values, type = "b", xlab = "i", ylab = expression(hat(lambda)[i]))
dev.off()

## Scatterplot of first two PCs.
pdf("figures/scatter.pdf")
plot(PCs[, 1], PCs[, 2], xlab = expression(PC[1]), ylab = expression(PC[2]),
  type = "n")
points(PCs[Z == 1, 1], PCs[Z == 1, 2], pch = 20, col = "blue")
points(PCs[Z == 5, 1], PCs[Z == 5, 2], pch = 20, col = "green")
points(PCs[Z == 8, 1], PCs[Z == 8, 2], pch = 20, col = "red")
legend(-6, -2.5, legend = c(1, 5, 8), lwd = rep(3, 3),
  col = c("blue", "green", "red"), bty = "n")
dev.off()
```

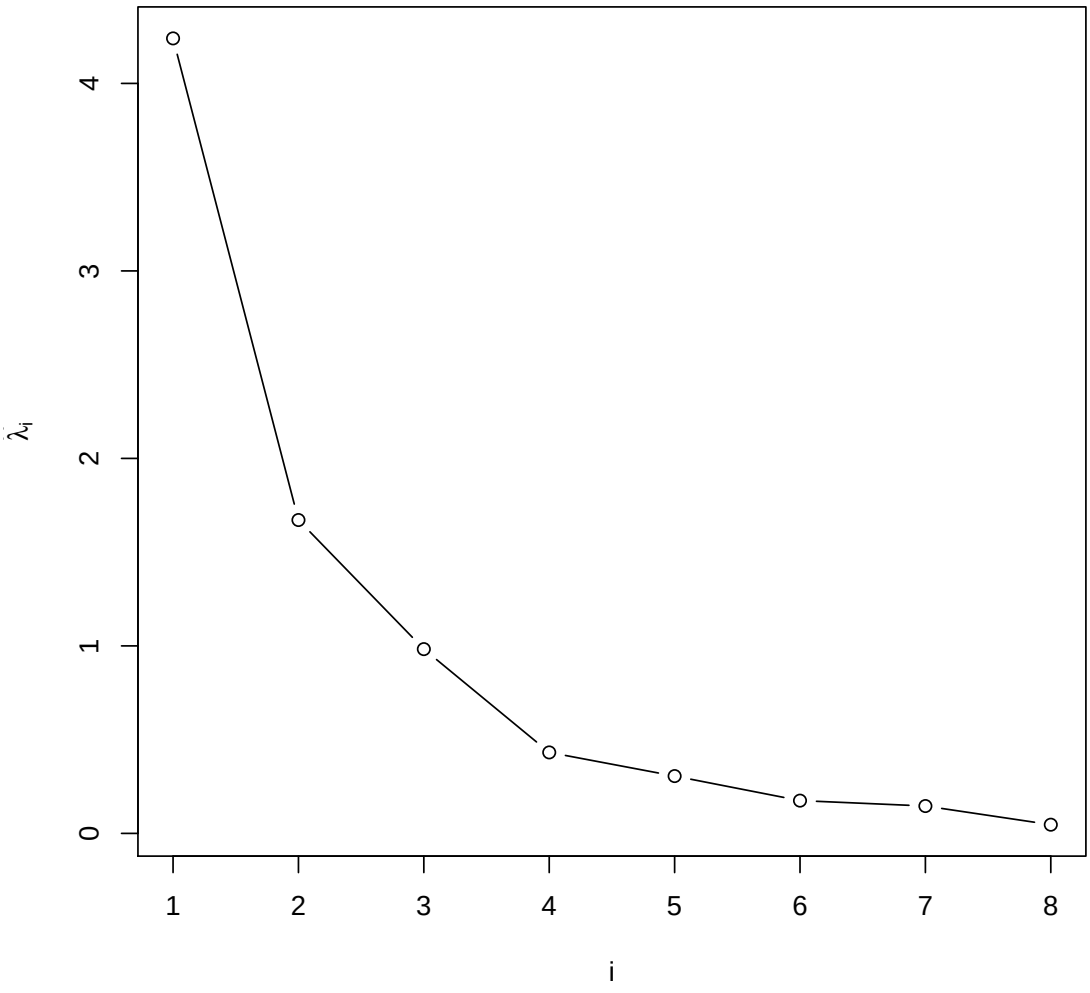

Figure 1: Scree plot for the bull data.

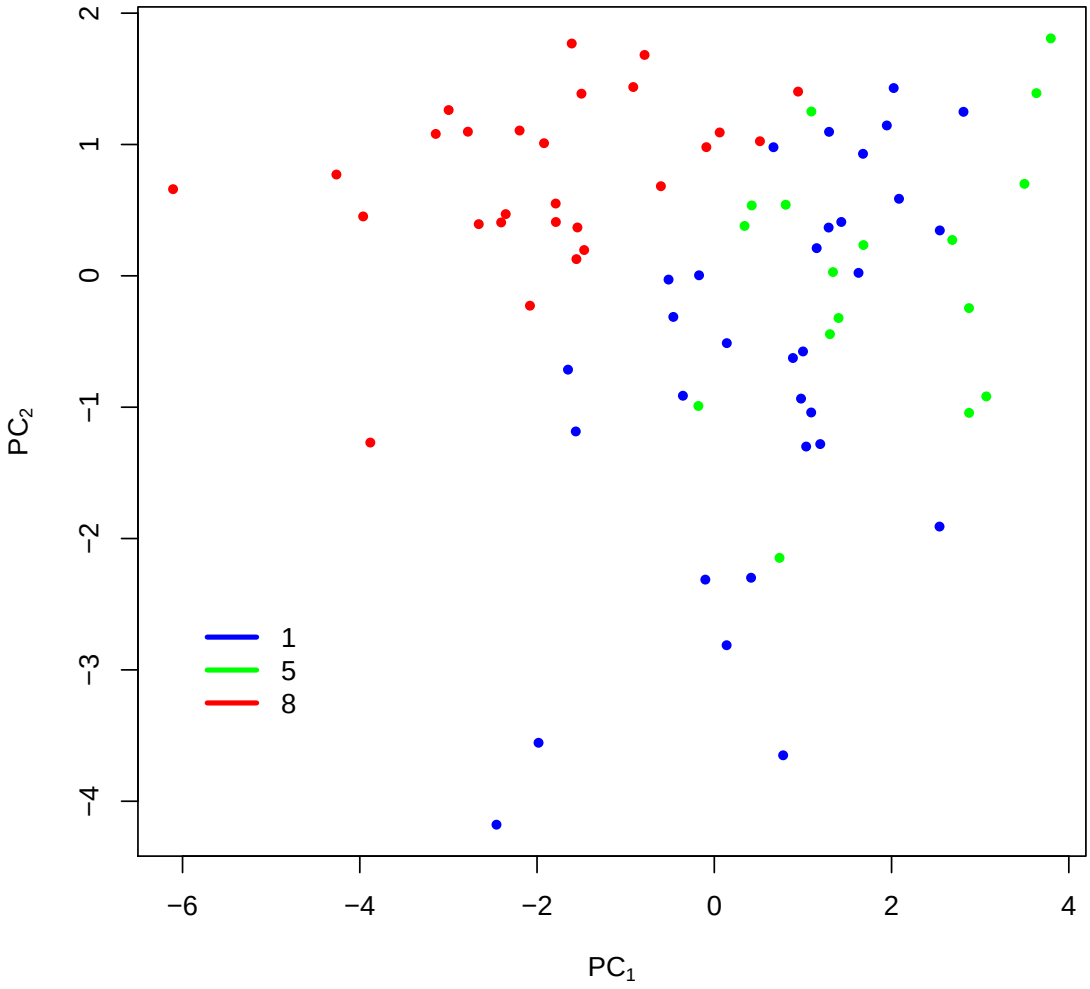

Figure 2: Scatterplot of first two PCs for the bull data, color coded by `Breed`.