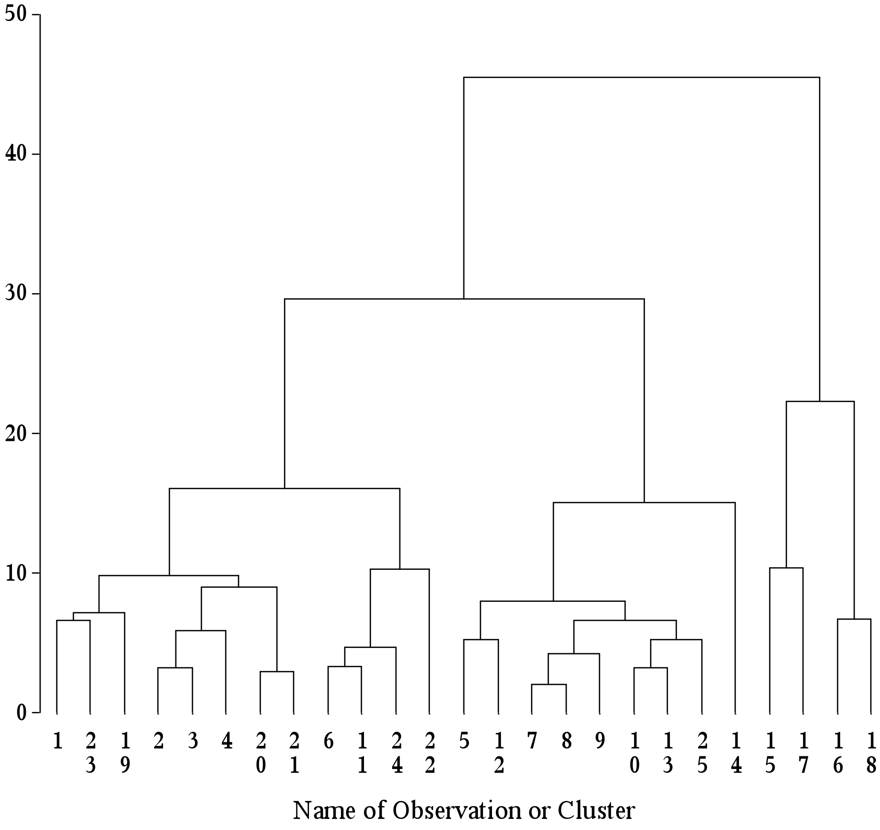
Problem 1 (*data from Applied Multivariate Techniques, by Subhash Sarma*) The file "fin.dat" includes seven financial performance variables for 25 companies of three different types (pharmaceutical, supermarket, and textile). Columns correspond to type, ID, return on revenue, debt to revenue ratio, sales, earnings per share, network performance management, profit to earning ratio, and profitability.

(a) Use cluster analysis with a hierarchical approach (you may choose the linkage method you wish) on the seven performance variables. Include a dendrogram.

Results using the complete linkage method are shown below.



(b) Comment on the agreement/disagreement between your results and the original companies' groupings by type.

These cluster classifications are reasonably consistent with the actual company types. Most (8/14) of the pharmaceutical companies were combined together in the first cluster. All but one of the supermarket companies were combined to form the second cluster, and all but one of the textile companies were correctly combined together as well. The table below summarizes these results.

	1	2	3	Total
Pharmaceutical	8	6	0	14
Supermarket	1	5	0	6
Textile	0	1	4	5

Problem 2 Data are collected for $n = 17$ years on price and consumption of pork and beef. The variables are: 1) Price of beef (cents/lb), 2) Price of pork (cents/lb), 3) Consumption of beef per capita (lb) and 4) Consumption of pork per capita (lb). Perform a canonical correlation analysis,

describing the relationships between the price and consumption variables. The data can be found in "meat.dat" with columns corresponding to the variables described above.

(a) Report the canonical correlations and the squared canonical correlations. Provide an interpretation of each of them.

The canonical correlations are .9728 and .7038. The first is largest correlation between any linear combination of price variables and any linear combination of consumption variables. The second is the largest such correlation among combinations uncorrelated with the first two. When these correlations are squared, they may be interpreted as the percent of variability in one set explained by the other set. For example, $.9728^2 = 94.63\%$ of the variability in the first price variable combination, U_1, is explained by the first consumption variable combination, V_1, and $.7038^2 = 49.47\%$ of the variability in the second price variable combination, U_2, is explained by the second consumption variable combination, V_2.

(b) Are any of the canonical correlations significantly different from zero? If so, which ones? Give the appropriate test statistic(s), degrees of freedom, and p-value(s).

The test statistic for whether both (population) canonical correlations are zero is $f = 32.96$ with 4 and 26 degrees of freedom. The p-value is less than .0001, which is highly significant evidence that at least one of these correlations is nonzero. Since the first is necessarily as large as the second, the follow-up test considers only the second. The test statistic for whether it is zero is $f = 13.71$ with 1 and 14 degrees of freedom. With a p-value of .0024, it is also significant. The conclusion is that both canonical correlations are significantly different from zero.

(c) For the first pair of canonical variables, say U_1 and V_1, give the standardized coefficients for computing them from the price and consumption variables. Give the formulas for computing U_1 and V_1, or explain in words how they would be computed. Let X_1 and X_2 represent the price of beef and pork, respectively, standardized to have mean 0 and variance 1. Let Y_1 and Y_2 represent the consumption of beef and port, respectively, standardized to have mean 0 and variance 1. Then,

$$U_1 = -.8754X_1 - .2757X_2 \qquad U_2 = -.5944X_1 + 1.0215X_2$$
$$V_1 = .8841Y_1 + .7913Y_2 \qquad V_2 = .5578Y_1 - .6831Y_2$$

Problem 3 The following page gives results for a principal components analysis of college test scores variables. The data set contains scores for 87 college students on college level examination program tests of X_1 = social science and history (called social in the output), X_2= verbal and X_3 = science. Use the output to help you answer the parts of this question.

(a) Give the variances of the three variables: social, verbal, and science.

For social, verbal, and science, the variances are 5808.059, 126.054, and 23.112, respectively.

(b) What is the total variance for the three variables?

The total variance is $5808.059 + 126.054 + 23.112 = 5957.225$.

(c) Let Y_1 represent the first principal component in the analysis of the covariance matrix. Express Y_1 as a linear combination of the original three variables (social, verbal, science).

$$Y_1 = .994X_1 + .103X_2 + .039X_3$$

(d) One student's scores are social = 488, verbal = 62, science = 18. For this individual, calculate the value of Y_1 (as defined in part c), the first principal component for the covariance matrix.

From part c), $y_1 = .994(488) + .103(62) + .039(18) = 492.16$ is the observed principal component score for this student.

(e) Now consider the analysis of the correlation matrix. What is the total "variance" in this situation?

Since each standardized variable now has variance one, the total variance is three.

(f) What is the variance of the first principal component for the correlation matrix? What proportion of the total variance is explained by this principal component?

The variance of the first principal component is the first eigenvalue, 2.165, which explains 72.2% of the total variance.

(g) Write a general interpretation of the first two principal components of the correlation matrix. That is, roughly, what is the first component measuring? Roughly, what is the second component measuring?

The first and second principal components, Y_1 and Y_2, are defined as

$$Y_1 = .621Z_1 + .571Z_2 + .535Z_3 \quad \text{and} \quad Y_2 = -.114Z_1 - .610Z_2 + .784Z_3$$

where Z_1, Z_2, and Z_3 are the standardized values for social, verbal, and science, respectively. Y_1 appears to be a near equal combination of all three variables, perhaps a measure of general intelligence, whereas Y_2 contrasts the science score with the other two. This reflects the more quantitative nature of the science assessment.

SAS code:

```
proc cluster data=fin method=complete nosquare nonorm outtree=clust;
  var v1-v7;
  id id;  run;
proc tree data=clust nclusters=3 out=tree_out;  run;
proc freq data=combine;
  tables type*cluster / nopercent norow nocol;
  run; quit;
proc cancorr data=meat out=canout vprefix=price vname="Price Variables"
                       wprefix=consume wname="Consumption Variables";
  var pricebe pricepo;
  with conbe conpo;  run;
```

Descriptive Statistics for the Variables

```
                      Simple Statistics

                  social            verbal         science
  Mean          526.5862           54.690           25.126
  StD            76.211            11.227            4.807

                      Covariance Matrix

                   social            verbal       science
social            5808.059         597.835       222.030
verbal             597.835         126.054         23.389
science            222.030          23.389         23.112

                Total Variance    5957.225
```

Principal Component Analysis of Covariance Matrix (Variables aren’t standardized)

```
  Eigenvalues of the Covariance Matrix

         Eigenvalue    Difference    Proportion    Cumulative

    1    5878.792       5814.957        0.987         0.987
    2      63.835         49.237        0.011         0.998
    3      14.598                       0.002         1.000

                        Eigenvectors

                      Prin1         Prin2         Prin3
      social         0.994         –.104         –.037
      verbal         0.103         0.995         –.0108
      science        0.039         0.006          0.999
```

Principal Component Analysis of Correlation Matrix (Standardized Variables)

```
                  Eigenvalues of the Correlation Matrix

         Eigenvalue    Difference    Proportion    Cumulative
    1    2.165         1.591          0.722          0.722
    2    0.574          .314          0.191          0.913
    3    0.261                        0.087          1.000

                        Eigenvectors

                   Prin1         Prin2         Prin3
      social       0.621         –.114         –.775
      verbal       0.571         –.610         0.548
      science      0.535         0.784         0.314
```