

# Multivariate Data Analysis

## Exercises 1

**Released 10 October 2011, work submitted by ~Tuesday 24 October 2011 will be marked and returned. Work submitted after solutions are made available will not be marked**

1) Dataset `nfl2000.Rdata`\* gives performance statistics for 31 teams in the US National Football League for the year 2000. Twelve measures of performance were made, six include the syllable `home` in the variable name and six include the syllable `opp`. The measures of performance were

<code>homedrives50</code>	drives begun in opponents' territory
<code>homedrives20</code>	drives begun within 20 yards of the goal
<code>oppdrives50</code>	opponents drives begun in team's territory
<code>oppdrives20</code>	opponents drives begun within 20 yards of goal
<code>hometouch</code>	touchdowns scored by team
<code>opptouch</code>	touchdowns scored against team
<code>homeyards</code>	total yardage gained by offence
<code>oppyards</code>	total yardage allowed by defence
<code>hometop</code>	time of possession by offence (in minutes)
<code>opptop</code>	time of possession by opponents' offence
<code>homelsts</code>	first downs obtained by offence
<code>opplsts</code>	first downs allowed by defence

The dataset contains a three letter abbreviation for the team as a row name. The coding is

<b>initials</b>	<b>team</b>	<b>initials</b>	<b>team</b>
<b>ARI</b>	Arizona Cardinals	<b>BAL</b>	Baltimore Ravens
<b>ATL</b>	Atlanta Falcons	<b>BUF</b>	Buffalo Bills
<b>CAR</b>	Carolina Panthers	<b>CIN</b>	Cincinnati Bengals
<b>CHI</b>	Chicago Bears	<b>CLE</b>	Cleveland Browns
<b>DAL</b>	Dallas Cowboys	<b>DEN</b>	Denver Broncos
<b>DET</b>	Detroit Lions	<b>IND</b>	Indianapolis Colts
<b>GB</b>	Green Bay Packers	<b>JAX</b>	Jacksonville Jaguars
<b>MIN</b>	Minnesota Vikings	<b>KC</b>	Kansas City Chiefs
<b>NO</b>	New Orleans Saints	<b>MIA</b>	Miami Dolphins
<b>NYG</b>	New York Giants	<b>NE</b>	New England Patriots
<b>PHI</b>	Philadelphia Eagles	<b>NYJ</b>	New York Jets
<b>SF</b>	San Francisco 49ers	<b>OAK</b>	Oakland Raiders
<b>STL</b>	St. Louis Rams	<b>PIT</b>	Pittsburgh Steelers
<b>TB</b>	Tampa Bay Buccaneers	<b>SD</b>	San Diego Chargers
<b>WAS</b>	Washington Redskins	<b>SEA</b>	Seattle Seahawks
		<b>TEN</b>	Tennessee Titans



- i) Do the syllables *home* and *opp* most probably refer to when the team was playing at *home* and playing *away* or do they refer to events *by* the team and *against* the team?
- ii) Use principal component analysis to identify and describe the main sources of variation of the performances.
- iii) Produce a scatter plot of the teams referred to their principal component scores and comment on any features you think worthy of mention.

**(NB: You are strongly advised to work through Task Sheet 2, Q3 if you have not already done so).**

\*source: *Journal of Statistics Education Data Archive*



2) Measurements of various chemical properties were made on 43 samples of soil taken from areas close to motorway bridges suffering from corrosion. The corrosion can be of either of two types and the ultimate aim of the investigation was to see whether these measurements could be used to discriminate between the two types. Before such a full-scale analysis was undertaken some preliminary analyses were performed, using MINITAB. The record of the session (edited in places) is given below.

- (a) The principal component analysis has been performed on the correlation matrix rather than the covariance matrix. Why is this to be preferred for these data?
- (b) By using some suitable informal graphical technique, how many components would you recommend using in subsequent analyses?
- (c) What features of the samples do the first three components reflect?
- (d) What, approximately, is the values of the sample correlation between the scores of PC-1 and PC-2?
- (e) After looking at the various scatter plots of the principal component scores, what recommendation would you give to the investigator regarding the advisability of continuing with a discriminant analysis?



```
Worksheet size: 100000 cells
MTB > Retrieve "C:\soil.MTW".
```

```
MTB > desc c2-c9;
SUBC> by c1.
```

## Descriptive Statistics

Variable	Type	N	Mean	StDev
pH	Type 1	25	8.416	0.962
	Type 2	18	8.0722	0.3102
Water	Type 1	25	1.693	0.716
	Type 2	18	2.831	1.812
Acid	Type 1	25	0.5672	0.3937
	Type 2	18	0.4322	0.2603
Pyrite	Type 1	25	0.4628	0.2563
	Type 2	18	1.019	0.500
Carbon	Type 1	25	11.251	4.230
	Type 2	18	9.783	1.862
Moisture	Type 1	25	23.712	4.975
	Type 2	18	21.922	2.647
Organic	Type 1	25	2.556	0.720
	Type 2	18	2.272	0.530
MassLos	Type 1	25	5.536	1.575
	Type 2	18	6.833	0.807

```
MTB > PCA 'pH'-'MassLos';
SUBC> Coefficients c31-c38;
SUBC> Scores'PC-1'-'PC-8'.
```

## Principal Component Analysis

## Eigenanalysis of the Correlation Matrix

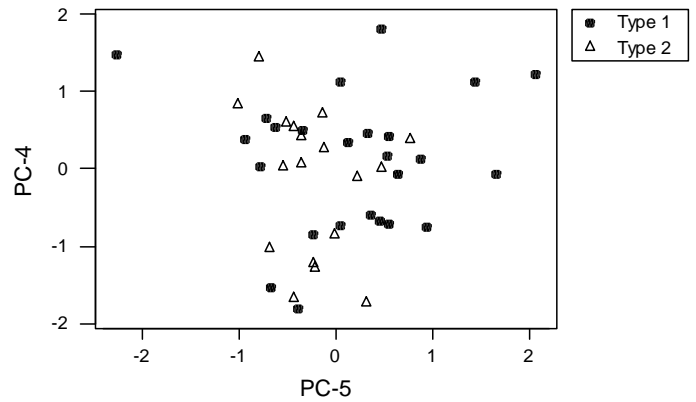
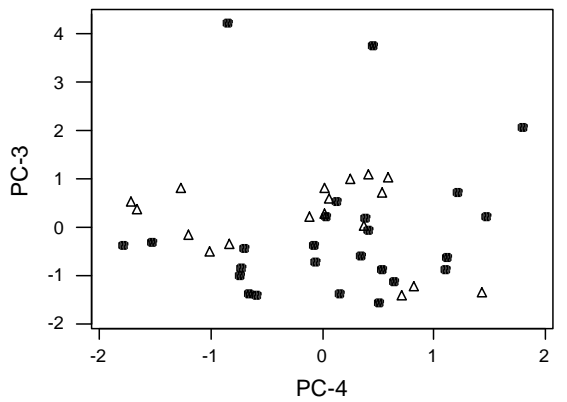
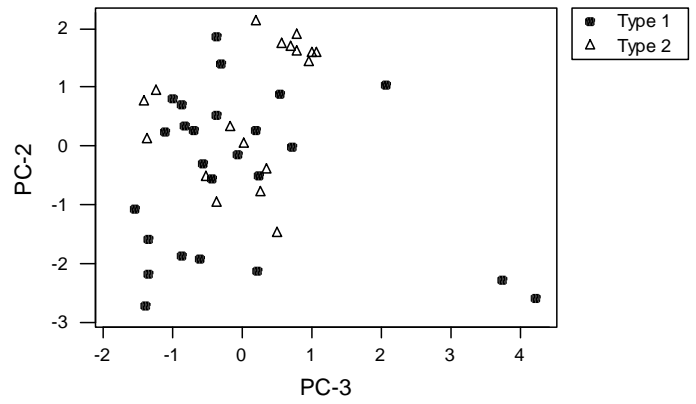
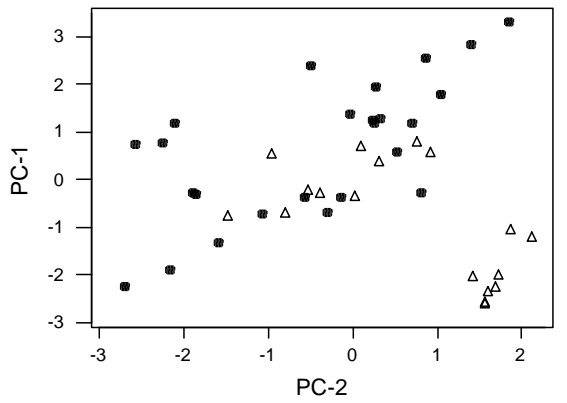
Eigenvalue	2.351	1.862	1.504	0.827	0.612	0.412	0.230	0.197
Proportion	0.294	0.233	0.188	0.103	0.077	0.052	0.029	0.025
Cumulative	0.294	0.527	0.715	0.818	0.895	0.947	0.975	1.000

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
pH	0.348	-0.032	0.559	0.267	-0.126	0.599	0.334	0.095
Water	-0.455	0.270	0.339	0.219	0.042	-0.460	0.520	0.272
Acid	-0.002	-0.367	0.622	-0.053	0.347	-0.238	-0.520	0.168
Pyrite	-0.351	0.446	0.157	0.417	-0.344	0.157	-0.539	-0.214
Carbon	0.520	0.291	-0.077	0.022	-0.355	-0.285	-0.206	0.624
Moisture	-0.001	-0.582	0.068	0.148	-0.687	-0.318	0.090	-0.231
Organic	-0.204	-0.392	-0.387	0.616	0.181	0.188	-0.067	0.450
MassLos	-0.487	-0.118	0.048	-0.549	-0.336	0.363	-0.049	0.445

```
MTB > Plot 'PC-1'*'PC-2' 'PC-2'*'PC-3' 'PC-3'*'PC-4' 'PC-4'*'PC-5';
SUBC> Symbol 'Type';
SUBC> Type 6 19;
SUBC> Size 1.0 1.5;
SUBC> ScFrame;
SUBC> ScAnnotation.
```



MTB > STOP



(This question is taken from the PAS370 1999/2000 examination)



3) \*\*\* (Not for submission) Suppose  $X = \{x_{ij} ; i=1, \dots, p, j=1, \dots, n\}$  is a set of  $n$  observations in  $p$  dimensions with  $\sum_{j=1}^n x_{ij} = 0$  all  $i=1, \dots, p$  (i.e. each

of the  $p$  variables has zero mean, so  $\bar{x} = 0$ ) and  $S = XX'/(n-1)$  is the sample variance of the data. Let  $u_j = x_j' S^{-1} x_j$  ( $j=1, \dots, n$ ) (so  $u_j$  is the squared Mahalanobis distance of  $x_j$  from the sample mean 0). Suppose the data are projected into one dimension by  $Y = \beta' X$  ( $\beta$  a  $p \times 1$  vector). Let  $y_j = \beta' x_j$  and define  $U_j(\beta) = (n-1) y_j' (YY')^{-1} y_j$ .

- i) Shew that  $U_j(\beta)$  is maximized with respect to  $\beta$  by the (right) eigenvector of  $S^{-1} x_j x_j'$  corresponding to its only non-zero eigenvalue.
- ii) If this eigenvector is  $\beta_j$ , shew that this maximum value  $U_j(\beta_j)$  is equal to this non-zero eigenvalue.
- iii) Shew that  $u_j = U_j(\beta_j)$ .
- iv) Shew that the non-zero eigenvalue of  $S^{-1} x_j x_j'$  is  $x_j' S^{-1} x_j$  and the corresponding eigenvector is proportional to  $S^{-1} x_j$

(Note that  $YY' = \beta' XX' \beta$  is  $1 \times 1$ , i.e. a scalar, so  $U_j(\beta) = (n-1) y_j' y_j / \beta' XX' \beta = (n-1) x_j' \beta \beta' x_j / \beta' XX' \beta = (n-1) \beta' x_j x_j' \beta / \beta' XX' \beta$  since  $\beta' x_j$  &  $x_j' \beta$  are  $1 \times 1$  and so commute. Further note that multiplying  $\beta$  by a scalar constant does not alter the value of  $U_j(\beta)$  so the problem is not altered if you impose the constraint that the denominator of the expression is 1.)

