

Part I. (60 points) Do all calculations in SAS. Use a word processor of your choice to write a report. Insert computer text output and graphics to support what you are saying, but you need to write something that looks like an academic paper — not a pile of computer output. Turn in a hard copy of your HW in class (i.e., don't email me your HW).

(60^{pts}) **1. Prehistoric goblets of Thailand:**

The following data can be found in Manly's text, "Multivariate statistical methods — A primer". The data consist of 6 measurements (heights and widths: MouthW, TotalW, TotalH, BaseW, StemW, and StemH, which I'll simply call x_1 – x_6) in cm. on each of 25 prehistoric goblets found in Thailand.

60 pts

x1	x2	x3	x4	x5	x6
13	21	24	14	7	8
14	14	24	19	5	9
19	23	24	20	6	12
17	18	16	16	11	8
19	20	16	16	10	7
12	20	24	17	6	9
12	19	22	16	6	10
12	22	25	15	7	7
11	15	17	11	6	5
11	13	14	11	7	4
12	20	25	18	5	12
13	21	23	15	9	8
12	15	19	12	5	6
13	22	26	17	7	10
14	22	26	15	7	9
14	19	20	17	5	10
15	16	15	15	9	7
19	21	20	16	9	10
12	20	26	16	7	10
17	20	27	18	6	14
13	20	27	17	6	9
9	9	10	7	4	3
8	8	7	5	2	2
9	9	8	4	2	2
12	19	27	18	5	12



(a) (20 pts) A PCA of the goblets shows that the major source of variation among the goblets is due to differences in size, as measured by a weighted linear combination of the 6 features. If I were primarily interested in the variation in shapes across goblets, a sensible strategy might be to transform the data by dividing each measurement on a goblet by the sum of the measurements on a goblet (there are other strategies). Create the 6 size adjusted variables x_1s – x_6s , defined by

$$\begin{aligned}
 x_1s &= x_1 / (x_1 + x_2 + x_3 + x_4 + x_5 + x_6) \\
 x_2s &= x_2 / (x_1 + x_2 + x_3 + x_4 + x_5 + x_6) \\
 x_3s &= x_3 / (x_1 + x_2 + x_3 + x_4 + x_5 + x_6) \\
 x_4s &= x_4 / (x_1 + x_2 + x_3 + x_4 + x_5 + x_6) \\
 x_5s &= x_5 / (x_1 + x_2 + x_3 + x_4 + x_5 + x_6) \\
 x_6s &= x_6 / (x_1 + x_2 + x_3 + x_4 + x_5 + x_6)
 \end{aligned}$$

Make a few (3) two-dimensional plots of the size adjusted measurements and qualitatively describe what you see.

- (b) (15 pts) Do a PCA on the standardized size adjusted measurements (i.e., using the correlation matrix for the size adjusted measurements). Interpret all the principal components (treating relatively small loadings as zeroes).
- (c) (5 pts) How many principal components appear to be sufficient to explain most of the variation in the size adjusted measurements? Discuss.
- (d) (15 pts) Make two-dimensional plots of PRIN1 against PRIN2, PRIN1 against PRIN3, and PRIN2 against PRIN3). Is anything interesting suggested by these plots?
- (e) (5 pts) One principal component has sample variance exactly zero. Explain why.

60 pts