Multivariate Resampling Techniques for Assessing Sample Sizes for Biplots

J. A. Hopkins
Department of Probability and Statistics. University of Sheffield.
GB-S10 2UN. Sheffield. UK.
E-mail: stp96JAH@shf.ac.uk

Introduction

Many different field studies, including archaeology, generate multivariate data, which are analysed by any or several multivariate techniques, such as biplots, principal components analysis, correspondence analysis, etc. [1]. Sometimes the data analysed are all the data, that were potentially available. However, in other cases, particularly in archaeology, there are more data, which could be collected, if necessary. These 'extra data', though, might be superfluous, and it may be possible to make worthwhile inferences on questions of interest, with a smaller sample. It is desirable to be economical in data collection, yet still be able to obtain conclusive results. For example, complete pots might be collected at a particular site, and measurements, such as rim circumference, height and base circumference, might be taken. The objective could be to look for similarities between pots, with obvious questions being: can it be shown graphically how various pots are related, and, are there any obvious groupings of similar pots? Are there any pots which are particularly unusual? Further questions raised are, how many pots need to be measured, and how many variables should be recorded (is it worth distinguishing between strongly correlated ones), bearing in mind that returning to a site, to supplement an inadequate sample, might be additionally expensive? The requirement is that there should be sufficient data, for inferences drawn from the analysis, to be adequately 'reliable.' Assessing whether sufficient or superfluous data have been collected, must be done in relation to the multivariate statistical technique, to be used in the analysis. We illustrate here, some methods for such assessment on the biplot, though similar methods can be used for any other multivariate technique.

Biplots, in common with principal components and correspondence analysis, are multivariate techniques, yielding graphical displays of a data matrix, where rows represent observations and columns represent variables. Often, more variables and more cases are measured than are strictly necessary, to make worthwhile statistical inferences on questions of interest, and similar results could also be obtained, by collecting smaller samples and measuring fewer variables. Modelling the data, by some appropriate distribution (Multivariate Normal or Multinomial, according to context), and then resampling, in the form of 'parametric bootstrapping', can be used to 'replicate' the data and to make inferences on the stability of the rows and columns of the data matrix. By randomly permuting the rows and selecting a subset of them, without replacement, it is possible to assess the effect of sample size and variables measured, in the analysis, and hence to make practical suggestions.

The biplot

The biplot is an exploratory multivariate technique that displays the rows (observations) of a data matrix as points, and the columns (variables) as lines (vectors), in low-dimensional space [2].
There are several types of biplot, the 3 main types being:

- Covariance
- Correlation
- Coefficient of Variation

All 3 types rely on a singular value decomposition (SVD) and are determined by different scalings of the data matrix X. Here, the Correlation Biplot is illustrated. This takes the data matrix X and standardises each column, by subtracting the column means, and dividing by the column standard deviations.

The data

We illustrate our methods on data, taken from Impey & Pollard [3]. They consist of 13 measurements, on each of 30 ceramic pots. A description of the measurements is given in Table 1, below, and the data, in Table 2. We can think of the data as a matrix X, with the dimensions, 30 x 13.

Table 1. Ceramic Pot Measurements

Measurement	Description
1	Internal height at centre
2	External diameter at lip
3	Internal diameter 2cm from base
4	External diameter 2cm from base
5	Internal diameter at lip
6	Overall height
7	Height from point of angle
8	Diameter at point of angle
9	External diameter of footring at base
10	Internal diameter of footring at base
11	Internal depth of footring at centre
12	Thickness of wall at 2cm from base
13	Thickness of lip

Table 2. Ceramic Pot Data

Pot Number	Measurements (cm)												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	6.50	7.85	5.65	6.70	7.30	7.00	6.30	6.55	4.80	4.10	0.15	0.30	0.30
2	6.85	7.40	5.85	6.80	7.10	7.55	6.70	6.75	4.10	3.50	0.30	0.25	0.30
3	6.85	7.30	5.80	6.80	6.60	7.55	6.75	6.65	4.30	3.55	0.15	0.35	0.35
4	5.90	6.80	5.70	6.75	6.30	7.00	6.35	6.75	4.40	3.60	0.30	0.40	0.25
5	6.30	7.30	4.70	6.55	6.75	6.95	6.35	6.35	4.60	4.15	0.30	0.55	0.30
6	6.15	7.15	6.20	7.20	6.20	6.85	6.15	7.00	4.75	3.90	0.25	0.40	0.45
7	6.50	7.40	6.25	6.85	6.70	7.40	6.75	6.15	4.00	3.45	0.25	0.25	0.30
8	6.35	7.60	5.25	6.75	6.95	6.95	6.40	6.40	4.65	3.95	0.35	0.50	0.30
9	6.60	7.70	6.55	7.05	6.80	7.40	6.70	6.75	4.20	3.55	0.25	0.20	0.30
10	6.60	7.35	6.40	7.00	7.00	7.45	6.80	6.60	4.05	3.40	0.30	0.25	0.28
11	6.80	7.40	6.30	7.00	6.85	7.40	6.60	6.70	4.20	3.65	0.15	0.20	0.30
12	6.50	7.40	5.10	6.60	6.80	7.15	6.55	6.45	4.50	4.00	0.15	0.55	0.30
13	5.90	7.25	6.55	7.25	6.30	6.90	6.05	7.05	4.20	3.50	0.25	0.30	0.45
14	5.85	7.10	6.55	7.20	6.35	7.05	6.15	7.10	4.40	3.55	0.30	0.35	0.35
15	6.10	7.15	6.50	7.15	6.25	7.10	6.35	7.10	4.45	3.50	0.25	0.40	0.40
16	6.65	7.45	5.85	6.90	6.60	7.35	6.60	6.70	4.50	3.85	0.10	0.35	0.35
17	6.55	7.45	5.25	6.60	6.70	7.10	7.10	6.55	4.65	4.10	0.20	0.50	0.35
18	6.15	7.20	6.70	7.20	6.20	7.15	6.65	6.85	4.35	3.50	0.30	0.30	0.35
19	6.00	7.00	6.50	7.10	6.15	7.05	6.30	6.90	4.10	3.30	0.25	0.35	0.45
20	6.00	7.20	6.55	7.15	6.30	7.05	6.25	7.00	4.35	3.70	0.30	0.30	0.45
21	6.75	7.50	5.95	6.90	6.75	7.45	6.80	6.65	4.30	3.55	0.15	0.35	0.35
22	6.30	6.80	6.50	6.90	5.95	7.30	6.30	6.80	4.30	3.55	0.20	0.25	0.35
23	6.10	7.25	6.65	7.40	6.40	7.00	6.10	7.25	4.40	3.70	0.15	0.30	0.50
24	6.75	7.55	6.60	6.95	6.80	7.35	6.55	6.75	4.20	3.60	0.20	0.30	0.35
25	6.30	7.60	5.20	6.55	6.80	7.10	6.50	6.55	4.40	3.95	0.30	0.50	0.35
26	6.90	7.60	6.25	6.80	6.65	7.55	6.85	6.55	4.10	3.55	0.15	0.25	0.30
27	6.40	7.65	5.55	6.70	6.85	7.20	6.50	6.40	4.75	4.15	0.20	0.55	0.40
28	6.20	7.55	5.75	6.85	6.95	7.05	6.45	6.65	4.75	4.10	0.40	0.50	0.25
29	6.15	8.05	5.65	7.05	7.15	6.85	6.20	6.70	4.50	3.85	0.30	0.55	0.40
30	6.50	7.55	5.15	6.85	6.90	7.05	6.60	6.65	4.75	4.05	0.35	0.60	0.35

Interpretation

Carrying out a correlation biplot analysis and representing the results in 2 dimensions gives Figure 1. Each of the 30 pots is represented by a circle, and each of the 13 variables, by a line emanating from the origin.

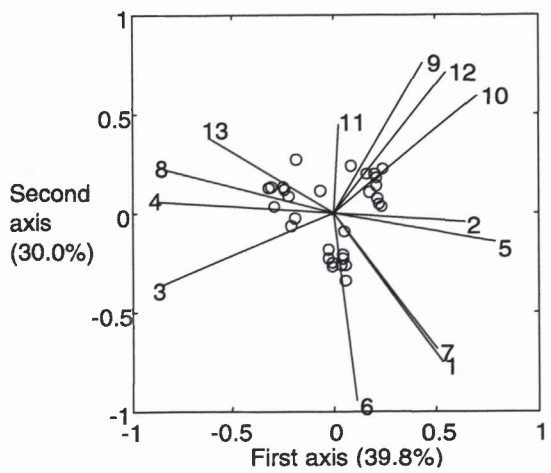


Figure 1. Correlation Biplot of the Ceramic Pot Data.

We see that 69.8%, of the variation in the data, is represented in these 2 dimensions. We interpret the biplot as follows: pairs of variables with small angles between them, such as 1 and 7, are highly positively correlated; pairs of variables with an angle of approximately 90° between them, such as 1 and 10, are uncorrelated; pairs of variables with an angle of approximately 180° between them, such as 2 and 4, are highly negatively correlated; pots which are similar, as regards measurements, are located close together; there appear to be three groups of pots.

Sample size

A: Do we have more data than necessary?

Our data consist of 13 measurements on 30 pots. With limited time available, it may not be possible to measure all 13 variables, and in particular, it may not be necessary to measure all variables, which are highly correlated. From examining Figure 1, we choose variables 2, 6, 8 and 10, as measuring different quantities in the data, and confine our analysis to these four. In addition to this, it may not be necessary to measure all 30 pots. Here, we choose a random subset of 15 pots from the original 30.

A correlation biplot analysis of 15 pots, on these four variables in the first two dimensions, is given in Figure 2. It shows three groups of pots, as in Figure 1, with the four variables, all reasonably spaced apart, and hence, measuring different aspects of the data.

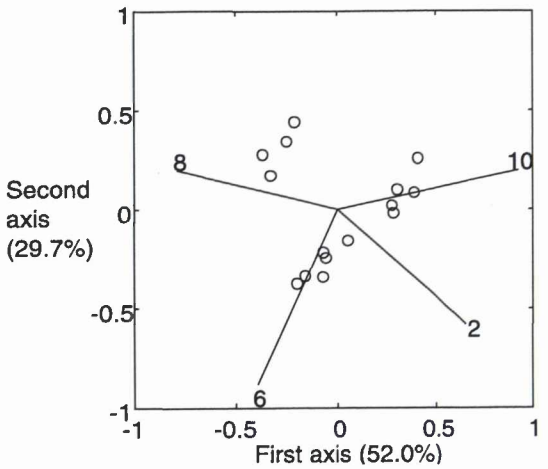


Figure 2. Correlation Biplot of Ceramic Pot Data - 15 pots, 4 variables

B: Do we have enough data?

It may be that 30 pots are not enough to indicate the true structure of the data, and that more should be collected. To investigate this possibility, we model the measurements on the original 30 pots as a multivariate normal distribution, with vector of column means $\overline{X}$, and variance-covariance matrix S. We then generate, say, 60 pots from this distribution. A biplot analysis of these 60 pots, on the original 13 variables, gives Figure 3. The variables are located in almost identical positions to Figure 1, though, of course, there no longer appears to be three groups of pots, as we did not model the data as such.

Figure 3. Correlation Biplot of Ceramic Pot Data - 60 pots, 13 variables

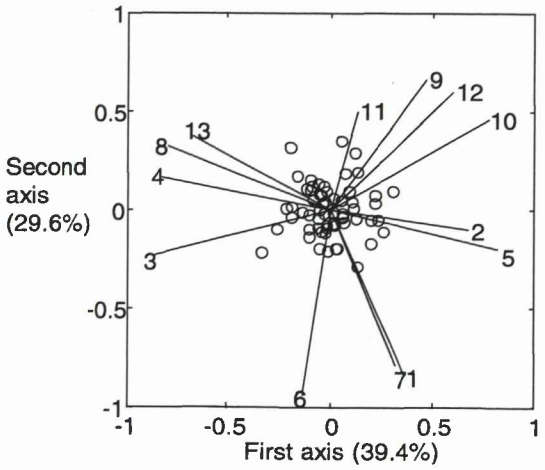
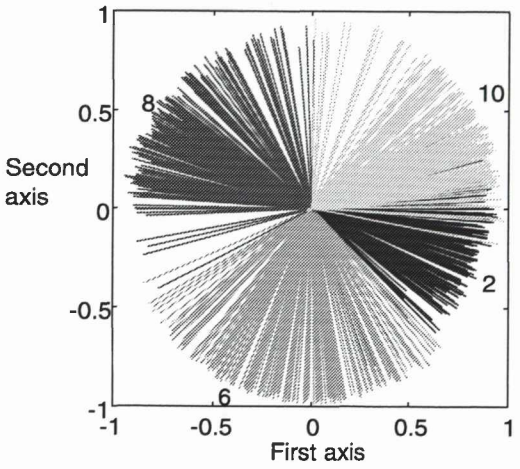


Figure 4. Bootstrap Fans - 200 samples, 15 pots, 4 variables



Bootstrap fans

A: Four variables

Our inferences, so far, are based on one particular dataset. But what happens if there is a slight change in the data collected on site - would this lead to a change in our conclusions? We can investigate how stable our particular sample is, by generating many samples of, say, 15 pots on 4 variables ('bootstrap sampling'), assuming a multivariate normal distribution. In this example, 200 samples were generated, and the variables 2, 6, 8 and 10 were used. Carrying out a biplot analysis on each of the 200 samples separately, we obtain 200 lines, representing each variable, one from each sample, and we can make inferences on the stability of the original sample.

A biplot analysis produces Figure 4. Each variable is represented in a different shade of grey (Table 3), with 200 lines for each variable, one from each generated sample. Looking at the angles between the widest lines of the 'fans' for each variable, we see that variable 10 is the most stable, as it subtends the smallest angle (100°) between the widest lines of its fan. Thus, we can be more confident that our inferences, regarding this variable, are reliable. Considering overlap between the variables, we see that there is little, suggesting that the variables are, in fact, measuring different aspects of the data and giving separate information. The choice of variables 2, 6, 8 and 10 is, therefore, not unreasonable.

Table 3. Variable shadings for Figure 4

	Variable 2
	Variable 6
	Variable 8
	Variable 10

B: Five variables

We can investigate the stability of the data when we choose five, not four variables. This time we consider 100 samples of 15 pots, on variables 2, 5, 6, 8 and 10. Having carried out a biplot analysis on each sample, Figure 5 shows the bootstrap fans of variables 2 and 5 (variables 6, 8 and 10 are omitted from the figure for clarity). Variable 2 is shown in dark grey, and variable 5 in light grey (Table 4). It is evident that there is much overlap between these two variables, suggesting that the variables are measuring similar aspects of the data, and it may not be necessary to measure both, to be able to detect the features of the data (i.e. division into subgroups). If we refer back to Figure 1, this is not surprising, since the angle subtended between the lines, representing these two variables, is small, indicating that they are highly correlated.

Figure 5. Bootstrap Fans - 100 samples, variables 2 and 5

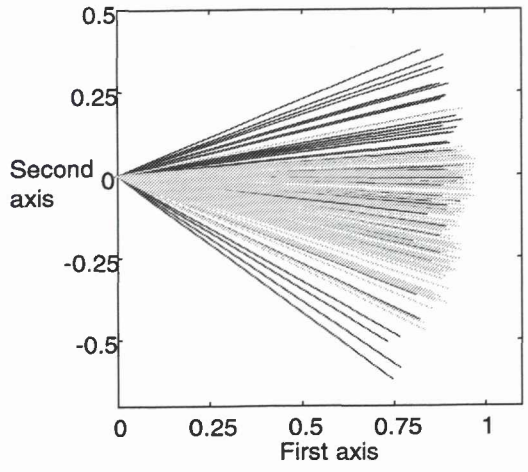


Table 4. Variable shadings for Figure 5

	Variable 2
	Variable 5

Conclusions

We can make four comments regarding the above analysis:

1. Measuring half or double the original number of pots does not alter the assessment of the relationship between any of the 13 variables.
2. Variables 2, 6, 8 and 10 appear to be measuring different quantities in the data and are relatively stable.
3. Measuring half or double the number of original pots, and only these four variables, does not alter the assessment of the relationship between these variables.
4. Future studies can be based on measuring fewer variables on more pots.

References

[1] BAXTER, M.J. (1994), *Multivariate Exploratory Analysis in Archaeology*, Edinburgh University Press Ltd., Edinburgh.

[2] GOWER, J.C. & HAND, D.J. (1996), *Biplots*, Chapman and Hall, London.

[3] IMPEY, O.R. & POLLARD, M. (1985), "A multivariate metrical study of ceramics made by three potters", *Oxford Journal of Archaeology*, 4, pp. 157-164.