23

Principal Components Analysis of compositional data in archaeology

M. J. Baxter*

M. P. Heyworth†

23.1 Introduction

The aim of this paper is to compare the use of some multivariate techniques for analysing compositional data in archaeology, with particular reference to glass compositions. An attempt is made to assess the value of these methods for the archaeological scientist.

Principal component analysis (PCA) is widely used as a data reduction method for aiding the perception of pattern in multivariate data. Correspondence analysis (CA) has similar aims and in this paper is treated as a form of PCA after suitable transformation of the data. Five variants of PCA, involving different transformations of the data, will be examined.

Suppose n objects are available for analysis. The term 'compositional data' is used to imply that p commensurable variables are measured for each object and that the sum of the measurements is fixed, and the same, for each object. Such data are common in archaeology and present their own difficulties in statistical analysis. The paper deals with the chemical composition of glass specimens in which the percentage presence of a range of oxides or elements is measured and sums to 100%, apart from measurement errors.

The problem examined here is that of deriving a classification of the objects based on the compositional data only. A successful classification would normally be interpreted in the light of any archaeological information available, but this information is not used in deriving the classification.

In Section 23.2 the nature of glass compositions, as they commonly arise in archaeological contexts, is described. A general strategy for statistical analysis is outlined in Section 23.3. Section 23.4 and Section 23.5 discuss the methodology and its application. Section 23.6 looks at the methodology from an archaeological perspective in which the questions the archaeological scientist is trying to answer in analysing such compositions are discussed. An attempt is then made to assess the usefulness, and limitations of, PCA based approaches in this context. The paper concludes with a brief discussion of areas where further research would be useful.

The books by Jolliffe 1986 and Greenacre 1984 contain extensive accounts of PCA and CA respectively. Aitchison 1986 is the seminal work on the analysis of compositional data. More extensive applications of the methods discussed here are given in Baxter 1989a, and a paper using similar methodologies in the same spirit, though in a different archaeological context, is that of Ringrose 1988.

* Dept. of Mathematics, Statistics & Operational Research
Nottingham Polytechnic
Clifton
Nottingham NG11 8NS

† Ancient Monuments Laboratory,
English Heritage
Fortress House
23 Savile Row
London W1X 2HE

23.2 The composition of glass

The study of glass from archaeological contexts has frequently involved the use of analytical techniques to obtain compositional data which can be used to interpret the raw materials used in glass production. Techniques such as neutron activation analysis (NAA), X-ray fluorescence (XRF) and inductively coupled plasma spectrometry (ICPS) have been used to provide data on a range of oxides and elements in glass (Sanderson & Hunter 1982, Sanderson *et al.* 1984, Heyworth *et al.* 1988).

Glass is a complex material which is manufactured by the fusion of silica, usually in the form of sand, with the aid of an alkaline flux, normally either soda or potash. Other substances such as calcium are usually added to increase the durability of the finished glass. Specific colourants, opacifiers and decolourants can be added to the glass batch, often in the form of metallic oxides, to produce a desired visual appearance: see Henderson 1985 for more details.

The final appearance of a glass could also be affected by other factors such as the furnace temperature, furnace atmosphere, and the duration of the glass-making process. The recycling of scrap glass, known as cullet, was also practised in antiquity leading to intermediate compositions which do not necessarily directly reflect the raw materials used in the production of a particular batch of glass.

A typical glass composition can therefore be broken down into suites of oxides/elements in terms of (i) the major oxides necessary to produce a working glass, usually silica, soda or potash and lime; (ii) minor oxides present as impurities in the raw materials; (iii) trace elements which are associated with the major and minor elements; and (iv) those elements which were added deliberately to the glass batch.

It is therefore important in the study of glass compositions to obtain data for the full range of the major, minor and trace elements contained in the glass. It is then necessary to be able to manipulate the data effectively through statistical analysis to obtain the maximum information possible. This information can then be combined with any relevant archaeological or typological information to attempt to answer specific questions about the production of the glass.

23.3 Analytical strategy

The majority of analyses described in this paper were carried out using the MINITAB 6.1 release. The increasing availability of such user-friendly interactive packages greatly facilitates the rapid exploratory analysis of complex data sets.

Any proposed multivariate analysis is usefully preceded by standard descriptive, univariate and bivariate graphical approaches. For example, the examination of dotplots will often serve to identify multimodality with respect to variables that is subsequently reflected in a classification exercise. Similarly, unusual observations will be revealed and it is often sensible to remove these prior to further analysis. This kind of approach is illustrated in Section 23.5.

It is possible to carry out PCA with similar ease so that the method can be regarded as a rapid data exploratory tool. MINITAB also allows the use of macros so that alternative forms of PCA are easily programmed, essentially by transforming the data and then using the MINITAB PCA command. Thus many different multivariate analyses can be undertaken very rapidly.

Since such ease of application is likely to promote increased use of the available methodologies an understanding of their relationships and properties in practice is important. The view adopted here is that it will often be informative to examine the data in several different ways. Several possibilities arise.

1. All the methods used lead to similar conclusions in which one may have a high degree of confidence.

2. Different methods give different and equally valid results. For example, one approach may be dominated by differences in the major elements, reflecting differences in the 'recipes' used to make the glass. A second approach might be dominated by the minor elements, reflecting differences in the colour of the glass or the origin of the raw materials used in making it.

3. Different methods give rise to different results some of which are archaeologically more valid than others. For example some of the methods to be discussed will sometimes produce results dominated by a single variable for essentially mathematical reasons.

23.4 Methodology

Assume that there are n specimens whose composition is measured with respect to p variables. Let X_{ij} be the percentage of oxide j present in specimen i and assume, for the purposes of exposition, that the sum of oxides for each specimen is 100%. The n by p raw data matrix so defined will be denoted by X. The PCA is to be based on a transformation, Y, of X and the following transformations will be considered where $\overline{X}_j$ is the mean of the j'th variable and S_j the standard deviation.

$$Y_{ij} = (X_{ij} - \overline{X}_j) \tag{23.1}$$

$$Y_{ij} = (X_{ij} - \overline{X}_j)/S_j \tag{23.2}$$

$$Y_{ij} = \ln(X_{ij}) - p^{-1}\sum_{j=1}^{p} \ln(X_{ij}) \tag{23.3}$$

$$Y_{ij} = (X_{ij} - \overline{X}_j)/\sqrt{\overline{X}_j} \tag{23.4}$$

$$Y_{ij} = (W_{ij} - \overline{W}_i)/\sqrt{\overline{W}_i} \tag{23.5}$$

where

$$W_{ij} = X_{ij}/\overline{X}_j \tag{23.6}$$

and

$$\overline{W}_i = p^{-1}\sum_{j=1}^{p} W_{ij} \tag{23.7}$$

In the case of (23.5)–(23.7) the elements of the leading components are rescaled before plotting.

The aim of a PCA as used here is to reduce the p dimensional data set defined by Y to a 2 dimensional set that can be plotted and inspected for structure. A geometric interpretation is that, on the scale of measurement used, the distance between two points, i and k, is defined by

$$d_{ik}^2 = \sum_{j=1}^{p} (Y_{ij} - Y_{kj})^2 \tag{23.8}$$

The two dimensional configuration arising from a PCA attempts to reproduce these distances as closely as possible in two dimensions. Methodological details and reasons for using the transformations (23.1)–(23.5) are discussed in the references previously cited.

Use of (23.1) and (23.2) involves a PCA of the covariance and correlation matrix of the raw data. The latter approach is that often adopted. Both approaches are

subject to the criticisms of Aitchison 1986 concerning the lack of interpretability of the correlations induced by the row-sum constraint on the data. The problem arises because the compositional nature of the data means it is constrained to lie within a $(p-1)$ dimensional simplex rather than p dimensional Euclidean space.

Aitchison's own approach, embodied in the use of (23.3), is to transform the data from the simplex to Euclidean space. Standard methods can then be applied to the transformed data.

Correspondence analysis provides a method for the simultaneous graphical representation of the rows and columns of a matrix of non-negative elements. If the elements in each row sum to 100% exactly then CA of the raw data matrix, for the representation of rows only, is equivalent to a PCA based on (23.4). If, prior to applying CA, the data is 'preprocessed' by dividing by the column means as suggested by Underhill & Peisach 1985 then a PCA based on (23.5) results.

Previous work (Baxter 1989b, Baxter 1989a) suggests the following points.

1. Using (23.1) it is evident from the structure of (23.8) that only those oxides with a reasonably high (greater than 5%) presence will contribute significantly to the analysis. These are typically the oxides SiO_2, CaO, Na_2O and K_2O, or a subset thereof. Thus the effective dimensionality of the data is reduced to 3 or 4 for which simple graphical methods, such as the use of ternary diagrams, will suffice. Aitchison's critique of the use of the raw data applies particularly strongly here. Nevertheless a PCA will identify groups differing in their composition relative to the oxides named above and this will often correspond to well established archaeological classifications such as the distinction between soda and potash glass.

2. Use of (23.2) will typically involve more of the oxides in an analysis often resulting in less clear-cut results. Although Aitchison's (1986) critique applies, if a majority of the oxides have low absolute presence a reasonable analysis may still be obtained (see contributions to the discussion of Aitchison 1982).

3. Empirically, with p of the order of 9 to 12, it has been observed that (23.2) and (23.4) often produce similar results. This appears to happen because $\sqrt{\overline{X}_j}$ and S_j are often highly correlated ($r > .9$) so that (23.2) and (23.4) are similar. In these cases analysis of the raw correlation matrix approximates a correspondence analysis (and is perhaps justified on this basis).

4. The structure of (23.3) in conjunction with (23.8) suggests that an analysis will be dominated by those oxides with a high relative variation. These are typically oxides with a low absolute presence (where the variation may be a function of experimental error, particularly when nearing detection limits). Experience confirms that this is often the case.

   In extreme cases three or fewer of the oxides will dominate, resulting in a large reduction in the effective dimensionality of the data. Although the classification obtained may make archaeological sense this is not inevitable. An example is given in the next section.

   Essentially this problem arises when some observations in a column are small. In extreme cases oxides may be recorded as traces or as zero, for which (23.3) is undefined. This problem may be resolved in various ways and an illustration is given in the next section.

5. Empirically it has been observed that the results of using (23.3) are often very similar to the results of using (23.5). It can be shown that use of (23.5) is similar to a generalized PCA based on distances, (23.8), that are a first order approximation to the distances used in (23.3).

6. In summary, for both theoretical and practical reasons, none of the methods described are ideal or uniquely 'correct'. Different methods emphasise different aspects of the data so the use of more than one method is often revealing. Where apparently interesting results are not wholly a consequence of an interaction between unusual data and the method used interpretation is a matter for the subject specialist.

All the analyses are readily carried out using MINITAB 6.1 commands or macros. With earlier versions macros need to be written for the basic PCA analysis. Details are available from the first author.

23.5 An application

As an example to illustrate points made in the previous section data given in Christie *et al.* 1979 will be used. They presented information on 19 specimens of Roman glass found in Norway and analysed with respect to 10 oxides SiO_2, TiO_2, Al_2O_3, Fe_2O_3, Mn_3O_4, MgO, CaO, Na_2O, K_2O and CuO. Their statistical analyses included the use of PCA based on (23.2). Baxter 1989a has examined the use of (23.2)–(23.5). Here a more detailed and slightly different treatment is adopted.

The original publication may be referred to for the raw data. The 19 specimens will be labelled A B ... S in subsequent diagrams and discussion. Fig. 23.1 shows MINITAB dotplots for the raw data. This is the same, apart from the scales, as the plots for (23.2) and (23.4).

Some multi-modality is evident, with respect to SiO_2 and TiO_2 for example, suggesting the likelihood of detecting some structure in the data. The sum of the compositions for specimens A to L was less than 97.5% and as low as 95.43% in the extreme case. It is inevitable that some measurement error will be present, particularly as the data is obtained from two different analytical techniques(atomic absorption spectrometry and X-ray fluorescence analysis), but the rather limited range of oxides analysed certainly means that the full glass composition is not available. Thus the data are not ideal for use with (23.3)–(23.5) which assume the composition sums to 100%. This was investigated by (rather arbitrarily) recalculating the SiO_2 percentage so that the compositional sum was 100% and reanalysing the data. This had little effect on the results and will not be presented here.

Inspection of the dotplots suggests the presence of outliers for Mn_3O_4 and CaO. These are associated with different specimens. As omission of the specimens has little effect on the results obtained using (23.1), (23.2) or (23.4) the full data set will be used.

For the data transformed as in (23.3) the main feature of the corresponding dotplots in Fig. 23.2 is the clear outlier for Mn_3O_4 corresponding to a raw value of .02%. The effect of this is that Mn_3O_4 is an important determinant of the leading two components and the outlier is clearly isolated on the component plot (Baxter 1989a). In contrast with analyses previously reported the analysis based on (23.3) will omit the offending specimen I. The effect for the transformation based on (23.5) was not so dramatic but, to assist comparison, specimen I will also be omitted from this analysis.

Specimens N, O, P and Q had zero values for CuO. In using (23.3) these were recoded as 0.0099, just slightly less than the next smallest value recorded for CuO. The effect of this is to ensure that the four specimens do not unduly influence the results obtained.

The plots based on the leading two components for each of methods (23.1) to (23.5) are shown in Figs. 23.3–23.7. Figs. 23.6 and 23.7 represent approximate correspondence analyses (for rows only), because of the measurement error, but are similar to the exact analyses in Baxter 1989a.

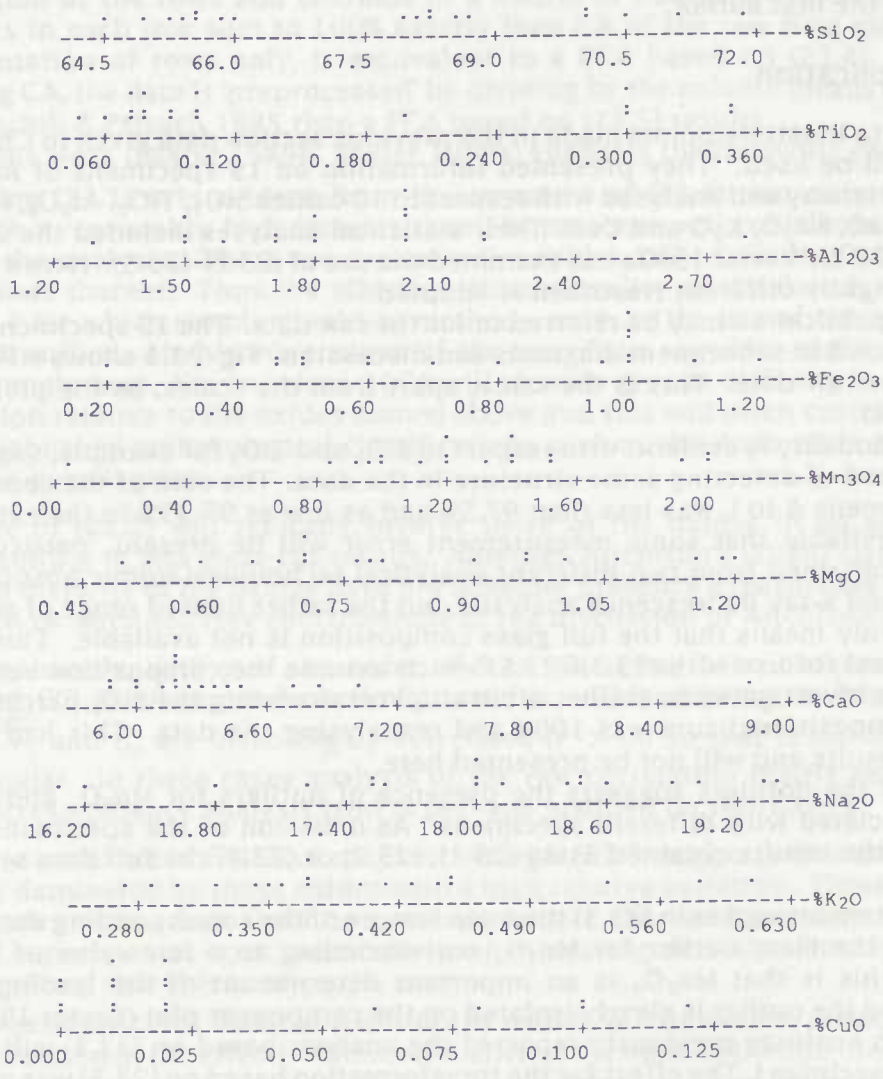

Figure 23.1: Minitab dotplots for raw data

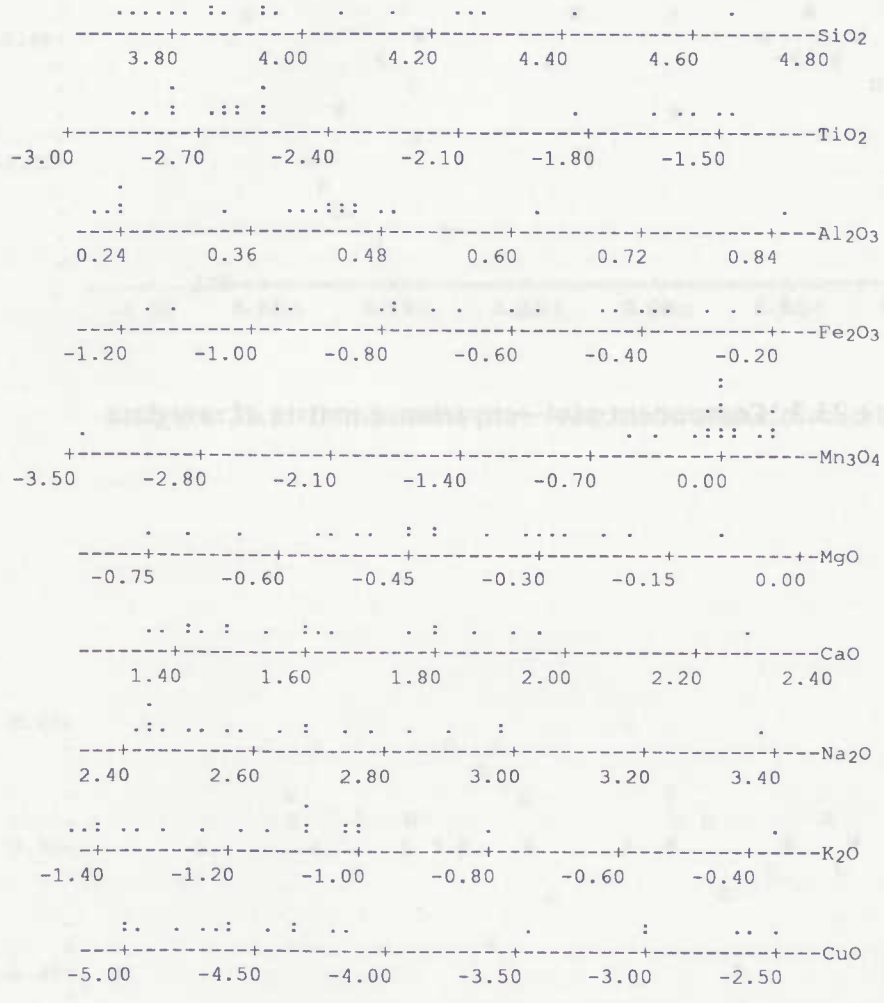

Figure 23.2: Minitab dotplots for Aitchison transformed data

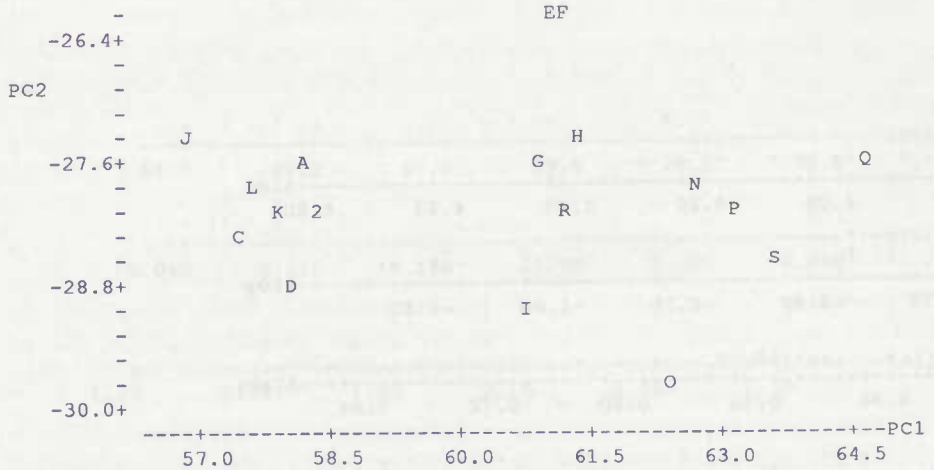

Figure 23.3: Component plot—covariance matrix of raw data

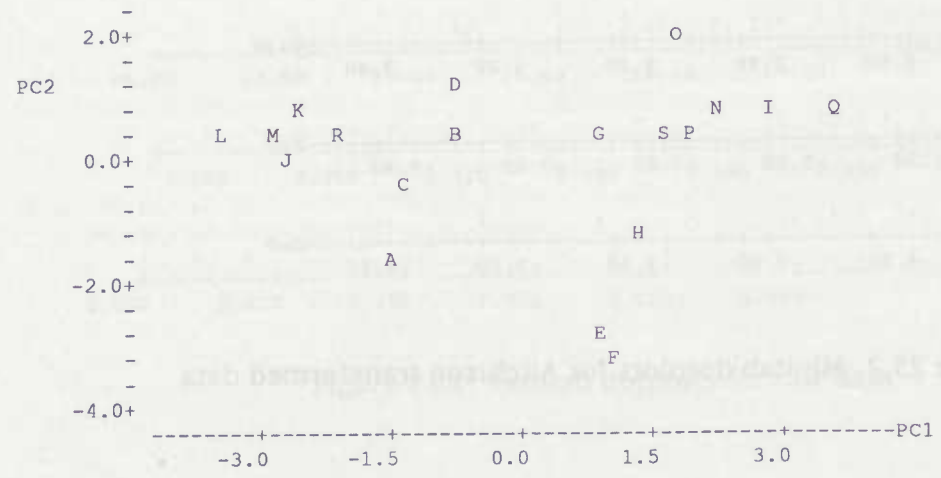

Figure 23.4: Component plot—correlation matrix of raw data

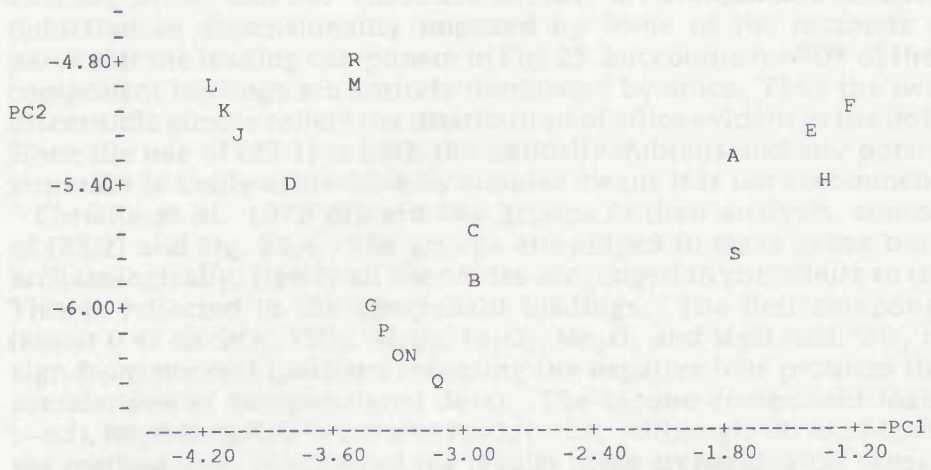

Figure 23.5: Component plot—Aitchison transformed data

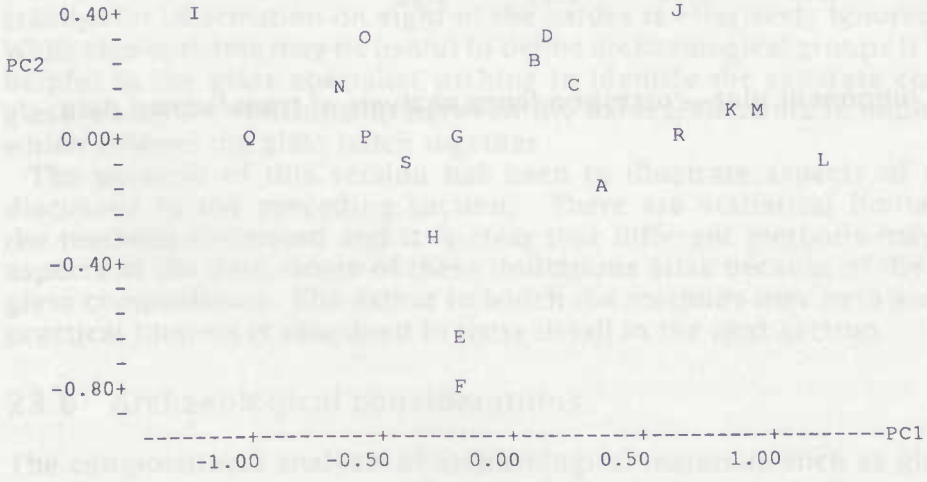

Figure 23.6: Component plot—correspondence analysis of raw data

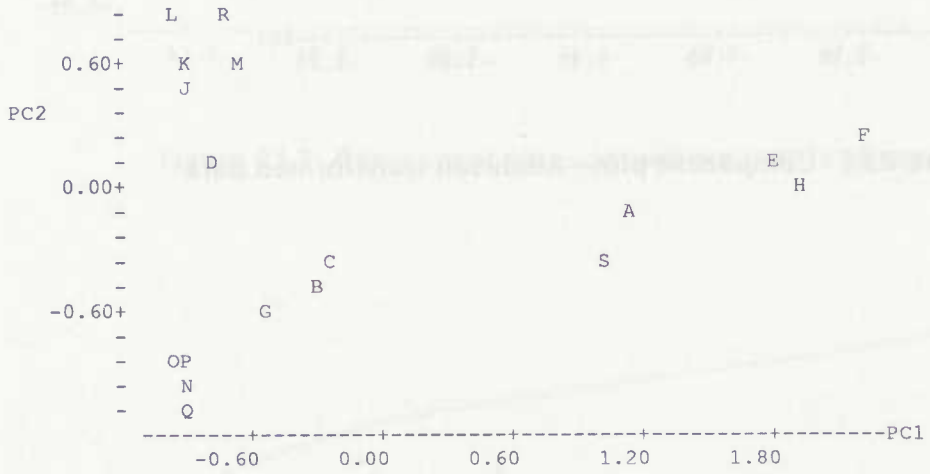

Figure 23.7: Component plot—correspondence analysis of transformed data

As expected, apart from orientation, Fig. 23.4 is similar to 23.6 and Fig. 23.5 is similar to 23.7. Figs. 23.3–23.5, which will be discussed in more detail, are clearly different and illustrate that different methods can indeed give rise to different results.

The percentages of variance explained by the leading two components in Figs. 23.3–23.5 are 91, 65 and 84. These are not directly comparable because of the effective reduction in dimensionality imposed by some of the methods on the data. In particular the leading component in Fig. 23.3 accounts for 80% of the variance and the component loadings are entirely dominated by silica. Thus the two or three groups discernible simply reflect the distribution of silica evident in the dotplots of Fig. 23.1. Since the use of (23.1) is both theoretically dubious and any potential revelation of structure is easily achievable by simpler means it is not recommended.

Christie *et al.* 1979 discern two groups in their analysis, comparable to the use of (23.2) and Fig. 23.4. The groups are judged to make sense both chemically and archaeologically. Nearly all the oxides are judged to contribute to the discrimination. This is reflected in the component loadings. The first component loads heavily (about 0.4) on SiO_2, TiO_2, Al_2O_3, Fe_2O_3, Mn_3O_4 and MgO with SiO_2 having a different sign from the rest (perhaps reflecting the negative bias problem that arises in using correlations of compositional data). The second component loads highly on CuO (-0.7), Na_2O (0.6), K_2O (-0.3) and Fe_2O_3 (-0.3). Although, on Aitchison's (1986) critique, the methodology is unsound the results make archaeological sense in the opinion of Christie *et al.* 1979.

Turning to the application of Aitchison's (1986) approach in Fig. 23.5 three clear groups can be seen. Inspection of the component loadings shows that these groups are almost entirely defined by the oxides based on TiO_2 and CuO which have the lowest absolute levels and highest relative variation of the oxides recorded. The group consisting of specimens A, E, F, H and S includes all values of CuO greater than 0.05%. The group D, J, K, L, M and R includes all values of TiO_2 of 0.20% or more. The remaining specimens have low values on both oxides.

Given the nature of glass analyses as often reported this kind of outcome is common and sometimes predictable though not inevitable. From a purely statistical standpoint the three groups identified make perfect sense and would be identified by a simple bivariate plot of the oxides involved. From the chemical or archaeological standpoint information on eight of the oxides is effectively ignored in this analysis. While this outcome may be useful to define archaeological groups it is not necessarily helpful to the glass specialist wishing to identify the separate constituents of the glass where the relationship between the oxides/elements is important in deciding which entered the glass batch together.

The purpose of this section has been to illustrate aspects of the methodology discussed in the preceding section. There are statistical limitations to each of the methods discussed and it is clear that different methods may reveal different aspects of the data. Some of these limitations arise because of the special nature of glass compositions. The extent to which the methods may help answer problems of practical interest is examined in more detail in the next section.

23.6 Archaeological considerations

The compositional analysis of archaeological materials such as glass is undertaken in an attempt to answer specific questions which relate to the mode of production, the raw materials and technology involved in the manufacturing process and the production location. In attempting to answer these questions it is necessary to use all the available data, not just the compositional data. It is particlarly important to include the typological and archaeological evidence that relates to the objects under

study. However, it is usually necessary to summarise the analytical information before it can be used in conjunction with the other available data.

Any statistical method which reduces the amount of information provided by a multi-element analysis of a group of objects to a manageable form which can be more readily interpreted is helpful to the archaeological scientist.It is particularly helpful if it not only provides information on any structure that exists in the data with regard to the objects but also shows which oxides/elements are responsible for the pattern.

A common method of analysing compositional data in archaeology is to undertake a hierarchical cluster analysis to identify structure in the data. A stepwise discriminant analysis usually follows to obtain information on the variables that are producing the data structure. The procedures described in this paper seem to produce much more useful information and in a much more flexible manner. Often a publication of archaeological analyses will show a dendrogram from a single cluster analysis and all the discussion will stem from this. The use of an array of multivariate methods can now be achieved quickly and easily with modern computing facilities. This process will greatly aid the archaeological scientist since it allows more flexibility in approaching the analysis of the data. It is important to try different statistical techniques as each may provide different, but still significant, information.

Both CA and PCA (through the use of a biplot) can be used for a joint display of rows and columns. This has not been exploited in this paper but is useful in identifying suites of oxides/elements that can be used in more detailed investigations of data structure. A practical limitation is on the number of objects and/or variables that can be used before the picture becomes unintelligible.

The value of a multi-element compositional analysis is that the full range of major, minor and trace elements in the glass are quantified. The use of correlations to link oxides/elements together into suites that may have entered the glass melt in the same raw material is well known. Often it is the trace elements, present at parts per million level in the glass, that are the best discriminators between raw material types. Although the methods based on (23.3) and (23.5) can highlight such elements this may be at the expense of ignoring information from a majority of the oxides/elements. By contrast (23.2) and (23.4) appear to use more of the available information at the possible expense of 'smearing' real differences based on the minor and trace elements.

There is no right or wrong method in these circumstances but each method can give important information. It is necessary to combine the information gained from the various methods to maximise the potential of compositional analysis for providing answers to archaeological questions posed by the material.

23.7 Conclusions

Statistical packages such as MINITAB 6.1 make it extremely easy to apply PCA interactively and use it as a tool for data exploration. Five variants of PCA have been examined, including two based on correspondence analysis approaches.

It is clear that different methods may produce different results so that the regular use of more than one method has potential advantages. The reasons for the emergence of any structure in the data need to be investigated and the structure interpreted in substantive terms if possible. Reporting the results of multiple analyses can be complicated but the temptation to present the 'best' or 'good' results only should be resisted.

Despite theoretical objections to the use of PCA on the raw correlation matrix for compositional data useable results may still be obtained. This is possibly because many of the components of a typical glass composition are reasonably small. If there is a near linear relationship (through the origin) between the square roots of the

variable means and their standard deviations an approximation to a correspondence analysis of the raw data results.

For typical glass compositions the theoretically sounder approach of Aitchison 1986 will often produce results determined by a small number of variables. Although the results are often interpretable information contained in the major elements may well be ignored. Correspondence analysis after dividing observations by their column means produces an approximation to Aitchison's approach that is unaffected by zeroes in the data. It would be useful to have conditions under which the approximation is guaranteed to be good.

The foregoing observations are largely based on experience with data sets containing 9 to 12 variables. Analytical techniques such as ICPS now routinely provide information on over 30 variables. These variables range from major elements present at levels over 60% to trace elements present at the parts per million level. It would be useful to have a range of sensible and easily applied methods for such data sets. Work in progress suggests that the methods discussed in this paper can be informative. Their empirical similarities are less clear cut and the methods based on (23.3) to (23.5) may ignore much of the information in the data.

Separate analyses based on subsets of the variables are a possibility, but these would have to be undertaken using sensible groups of variables that could be defined in terms of the raw material components of the glass. A simple separation into major, minor and trace elements would not be useful in this context as correlations between different oxides/elements are of diagnostic importance in identifying additions to the glass batch. Cluster analysis is a widely used method and a comparative study of different methods using different scales of measurement would be useful. Finally, it may be noted that PCA attempts to explain the variance in a set of data, on whatever scale of measurement is used. Methods that focus on the covariance structure are also of interest and it is hoped to report on these at a later date.

Bibliography

AITCHISON, J. A. 1982. "The statistical analysis of compositional data (with discussion)", *Journal of the Royal Statistical Society B*, 44: 139–177.

AITCHISON, J. A. 1986. *The Statistical Analysis of Compositional Data*. Chapman and Hall, London.

BAXTER, M. J. 1989a. "An empirical study of principal component and correspondence analyses of glass compositions". Submitted for publication.

BAXTER, M. J. 1989b. "The multivariate analysis of compositional data in archaeology: a methodological note", *Archaeometry*, 31: 45–53.

CHRISTIE, O. H. J., J. A. BRENNA, & E. STRAUME 1979. "Multivariate classification of Roman glasses found in Norway", *Archaeometry*, 21: 233–241.

GREENACRE, M. J. 1984. *Theory and applications of correspondence analysis*. Academic Press, London.

HENDERSON, J. 1985. "The raw materials of early glass production", *Oxford Journal of Archaeology*, 4: 267–291.

HEYWORTH, M. P., J. R. HUNTER, S. E. WARREN, & J. N. WALSH 1988. "The analysis of archaeological materials using inductively coupled plasma spectrometry". *in* Slater, E. A. & Tate, J. O., (eds.), *Science and Archaeology, Glasgow 1987*, British Series 196. British Archaeological Reports, Oxford.

JOLLIFFE, I. T. 1986. *Principal component analysis.* Springer-Verlag, New York.

RINGROSE, T. J. 1988. "Exploratory multivariate analysis of stratigraphic data: Armstrong's data from Pin Hole Cave reexamined". *in* Slater, E. A. & Tate, J. O., (eds.), *Science and Archaeology Glasgow 1987*, British Series 196. British Archaeological Reports, Oxford.

SANDERSON, D. C. W. & J. R. HUNTER 1982. "The neutron activation analysis of archaeological glasses from Scandinavia and Britain", *PACT*, 7: 401–411.

SANDERSON, D. C. W., HUNTER. J. R., & S. E. WARREN 1984. "Energy dispersive X-ray fluorescence analysis of 1st millennium glasses from Britain", *Journal of Archaeological Science*, 11: 53–69.

UNDERHILL, L. G. & M. PEISACH 1985. "Correspondence analysis and its application to multielemental trace analysis", *Journal of Trace and Microprobe Techniques*, 3: 41–65.