11
Graphical presentation of results from principal components analysis

M. J. Baxter and C. C. Beardah
Department of Mathematics, Statistics and O.R.,
The Nottingham Trent University, Clifton Campus, Nottingham, NG11 8NS, U.K.

11.1 Introduction

The main purpose of this paper is to discuss approaches to the informative display of output from a principal component analysis (PCA), in the context of common archaeological applications. Displays of the kind we have in mind are made possible, and more widely available, by the development of powerful and accessible software packages. We concentrate on use of the recently released Version 4 of the MATLAB package; other possibilities are noted in a later section.

No novelty is claimed for the statistical and graphical approaches used here, though application of them to PCA is uncommon in some cases, and we know of no applications to archaeological data of some of the methods. What we describe is 'work in progress', the aims of which are (a) to explore the capabilities of the MATLAB package, and (b) to develop what are hopefully useful, but uncommon, approaches to the display of PCA output that can be applied routinely.

There is a substantive problem that initially motivated this work that is outlined briefly in the next section and used for illustrative application in the text. The reasons for using MATLAB as the analytical package are discussed in Section 11.3, followed by a variety of applications in Section 11.4. Possible alternatives to MATLAB are noted in Section 11.5, and Section 11.6 concludes the paper.

11.2 The Substantive Problem

Barrera and Velde (1989) (referred to as BV in the paper) have published chemical analyses of 486 specimens of French Medieval glass. The percentage presence of ten oxides (based on the elements Ca, Na, K, Mg, P, Si, Al, Fe, Mn, Cl) was measured. Additional information was given on the period, site of origin, type and colour of the glass. On the basis of the level of Na the specimens divide clearly into two groups. The larger of these, with 361 specimens, is termed 'calco-potassic' glass with the percentage presence of CaO + K_2O being in excess of 22%.

On the basis of consideration of the levels of Na_2O, CaO, K_2O and MgO, BV sub-divide the calco-potassic group into three sub-groups. The largest of these groups (Type A in BV) corresponds to compositions that are typical of the Argonnes area in north-eastern France from period II (BV, 95), having relatively low values of Na_2O (<1%) and MgO (<4.5%) and a ratio of CaO/(CaO+K_2O) between 0.5 and 0.7. Type C in BV consists of specimens whose values for these three quantities lie outside the ranges given. These are rich in potassium and typical of northern, central or western traditions (BV, 94). Type B is intermediate in composition between these two groups. It should be noted that the sub-division proposed in BV is not entirely clear; however, interpreting their groupings differently than above does not affect the conclusions of this paper.

These data were drawn to our attention by Dr. Ian Freestone of the British Museum's Department of Scientific Research, who was interested in whether or not this typology would be supported if all the data were used. Of particular interest was the relationship between the composition of the French glass and that of specimens analysed by Dr. Freestone from the retable at Westminster Abbey (Freestone n.d.). It is intended to report separately on this (Baxter *et al.* in preparation)

11.3 Initial Data Analysis and MATLAB

Initial bi- and tri-variate plotting of the data suggested two main concentrations of points, and a third more dispersed group. An initial PCA of the data, based on the correlation matrix of seven of the oxides (for reasons given in Baxter *et al.*, in preparation), was undertaken using STATGRAPHICS and the resultant plots for the first four components are shown in Figure 11.1.

There are perhaps three concentrations of points evident on the plot of the first and second components, with the third component separating out two smaller subgroups within the larger concentrations.

Further investigation, involving labelling points on the plots is problematic within STATGRAPHICS because of the large number of points involved. A more general difficulty with large data sets is that the density of points on plots such as Figure 11.1 can make the perception of structure difficult. Facilities in the MATLAB package allow these problems to be overcome in ways that are explored in the rest of the paper. MATLAB, Version 4 of which has recently become available, is a sophisticated software package with a strong user base in Further and Higher Education institutions, particularly in Departments of Mathematics. The package is particularly useful in applications involving numerical matrices and the graphical representation thereof.

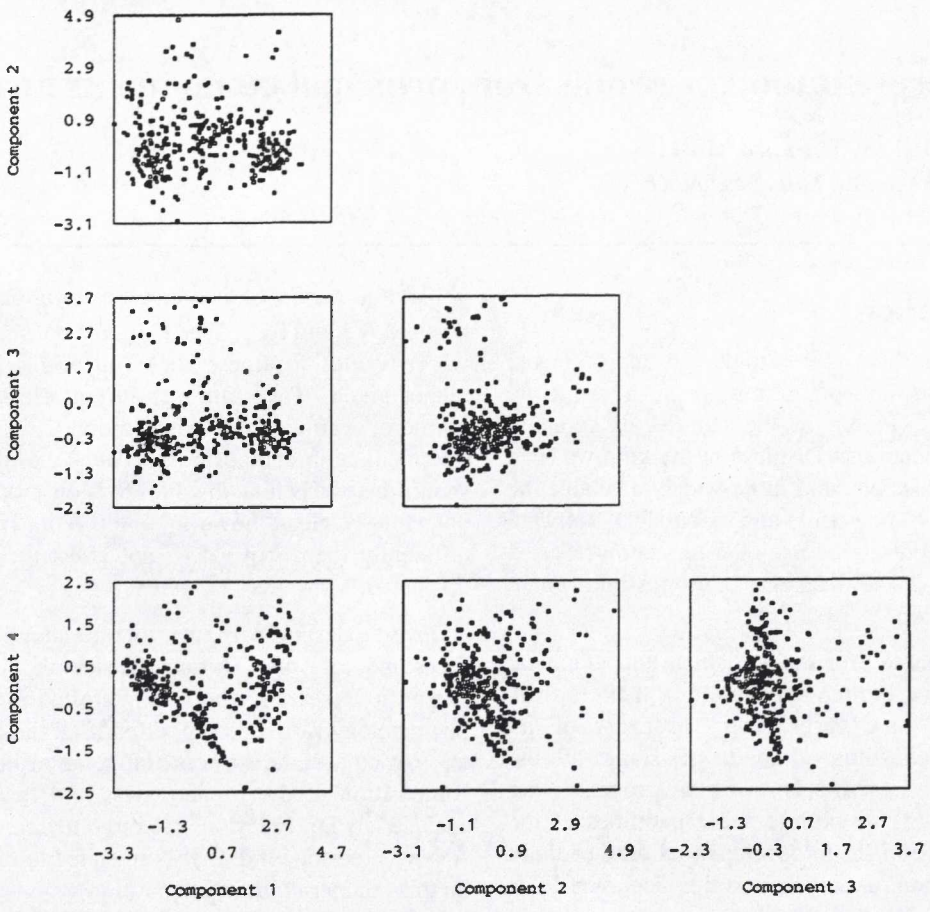

Figure 11.1: A 'draftsman' plot, obtained using STATGRAPHICS, showing all possible pairwise component plots based on the first four principal components from a PCA of the correlation matrix of the French medieval glass data.

Multivariate data can be represented as a matrix of values, where columns indicate different variables. This matrix representation of data, when coupled with MATLAB's graphics, matrix manipulation and programming capabilities, makes high quality, relatively low cost, graphical depiction of data accessible to all.

11.4 Applications

This section illustrates some possible uses of MATLAB for presenting results from PCA. The first two sub-sections, on colour and three-dimensional plotting, could be emulated in many packages – given the PCA scores – and are not new ideas. The other sub-sections on the use of density estimation methods are more novel, as applied to PCA, and may require more specialised software.

11.4.1 Use of colour plotting

For interactive inspection of large data sets the use of colour has obvious, and well known, attractions (e.g. Scott, 1992). In the present case colouring points on a plot of the first two components according to BV's typology made it immediately clear that their typology did not match that suggested by the plot. Their groups A and C tend to concentrate centrally and to the right on a plot of the first and second components; group B is dispersed over the entire plot with concentrations to the right and left. These features can be seen in the upper part of Figure 11.2 which is in black and white and uses different symbols for the groups. This is less satisfactory than the use of colour, primarily because of the large number of points involved.

11.4.2 Three-dimensional plotting

Three dimensional plots, based on the first three components, are increasingly used but still uncommon in publications. (e.g. Baxter 1994, 131; Hughes 1991; Rauret *et al.* 1987). With many data points, and for black and white printing, such plots can be difficult to read and presentation as in Figure 11.1 may be preferable. For exploratory use at a terminal the ability to colour points is a distinct advantage, as is the ability to rotate plots to obtain informative points of view and this is readily programmed within MATLAB.

It is possible to represent a third dimension on a two-dimensional plot by varying the type or size of plotting symbol (e.g. Hodson 1969), but this has been exploited infrequently in archaeological applications. In three dimensions different colours can be used to represent a fourth dimension, though we have yet to produce examples where this has been convincingly useful.

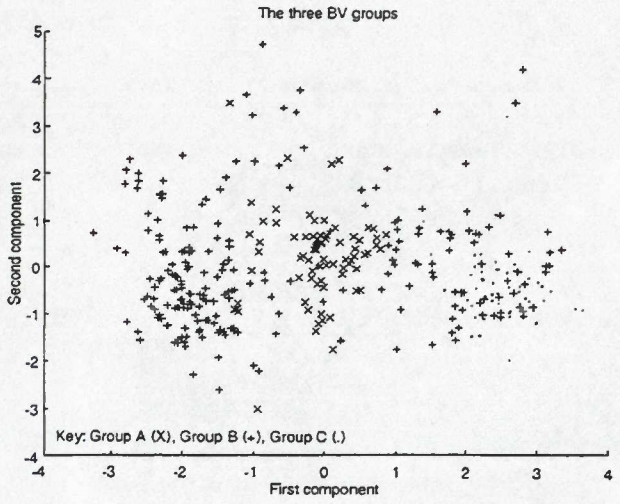

Figure 11.2: Two-dimensional component plots labelled according to the grouping suggested in Barrera and Velde (1989). The plot suggests three main clusters of points, different from BV's grouping.

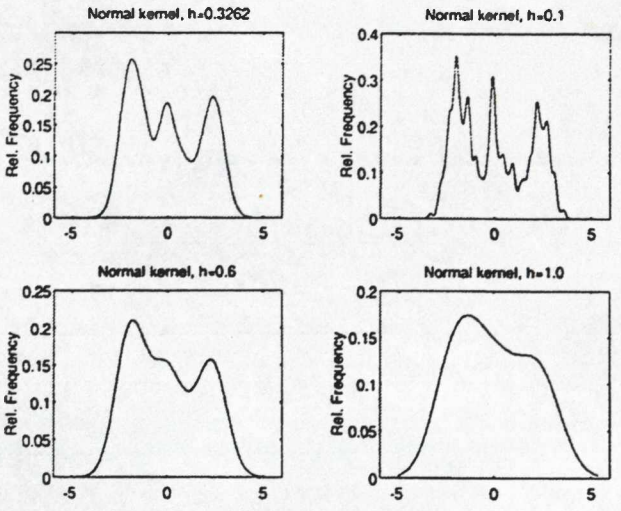

Figure 11.3: Four univariate kernel density estimates for the first principal component, using the normal kernel with different window widths (h).

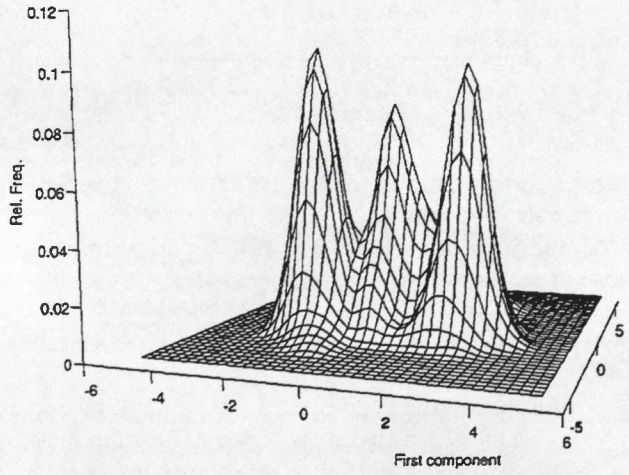

Figure 11.4: Bivariate kernel density estimates based on the first two principal components, using a normal kernel.

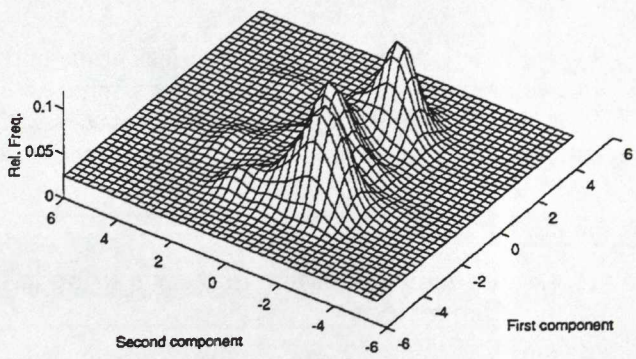

Figure 11.5: Bivariate kernel density estimate, as in Figure 11.4, showing a different angle of view.

11.4.3 Density estimation in one dimension

The example being used is typical, in that if there is structure present then it will often be captured by the first component. For display purposes with many data points the use of kernel density estimates is an option (Silverman 1986; Scott 1992). These do a similar job to the perhaps more familiar histogram but have a number of potential mathematical and aesthetic advantages. A histogram is defined by its origin and bin width (assumed fixed); its appearance can depend on the choice of both and is typically 'blocky'. Kernel density estimates depend on a 'window width' but not on choice of origin. They have a smoother appearance because of averaging, dependent on the choice of kernel, that takes place at each point of the curve. Figure 11.3 shows four density estimates using the normal kernel (Silverman 1986, 43) for different window widths.

The upper left figure uses an 'optimal' choice of window width (Silverman 1986, 45). The choice is optimal if the true density is normal, and tends to oversmooth multimodal densities (Silverman 1986, 46), so that the upper right figure, with a smaller window width, presents a better – if slightly more ragged – picture of the data. The oversmoothed figures in the lower part of Figure 11.3 smooth away much of the structure apparent in the upper figures. Using other kernels among those listed by Silverman makes little difference to the appearance. The three clear modes make evident the existence of three groups. For presentation purposes often only one figure is used. Scott (1992, 47) suggests that the 'examination of both undersmoothed and oversmoothed histograms should be routine' and gives further justification for this view on page 161 of his book.

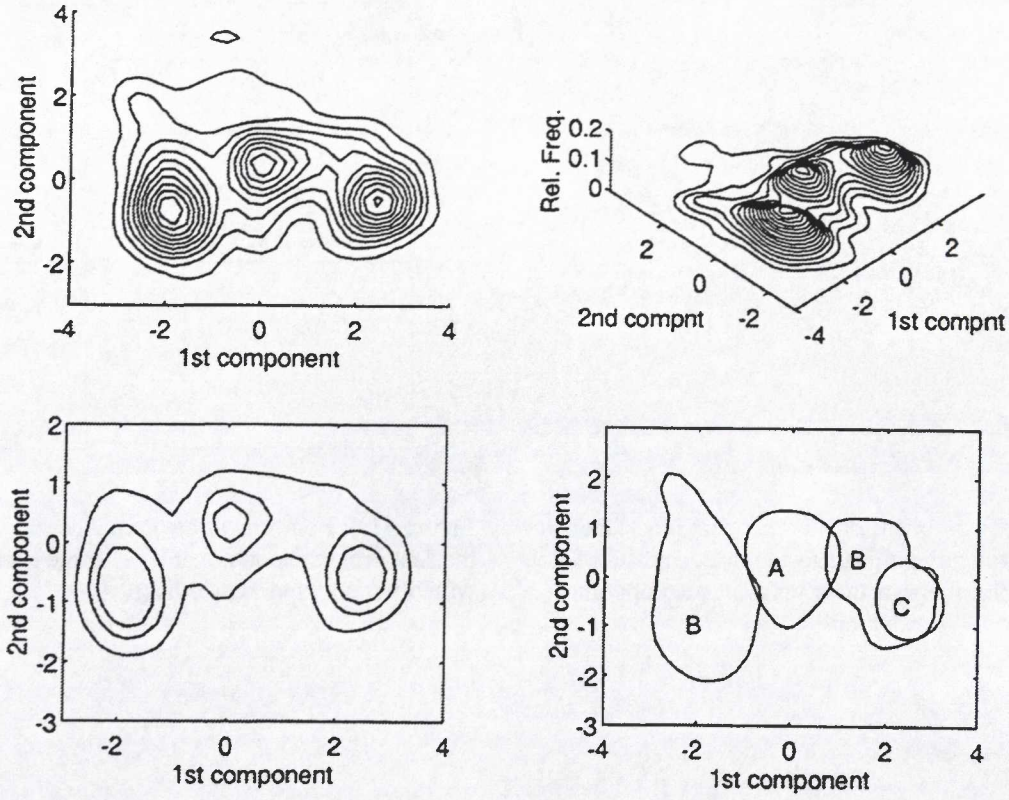

Figure 11.6: Four alternative contourings of the bivariate kernel density estimate of Figures 11.4 and 11.5. The upper figures present two views using contours spaced at equal intervals. The lower left figure shows the contours of the 'quartiles' of the data that enclose the most dense 25%, 50% and 75% of the data. The final figure shows contours, calculated separately for each of BV's groups, that enclose 75% of each group.

11.4.4 Density estimation and contouring in two dimensions

The univariate kernel density estimates of Figure 11.3 are two-dimensional representations of one-dimensional data. The ideas extend readily to bivariate density estimates that can be used to get three-dimensional perspective plots of two-dimensional data. Figures 11.4 and 11.5 show perspective plots, based on the normal kernel with window width 0.3611, of the first against second component. The first view is chosen to show the clear separation into three groups; the other view is from a different angle and shows features obscured in Figure 11.4. Since any single view can obscure features of the data, views from several different angles should usually be inspected. Different styles of plotting are available and colour, where different colours correspond to different heights, may also be used.

The bivariate density estimates can be used as the basis for obtaining contour maps of the data and several possibilities are shown in Figure 11.6. The upper left figure views the data from 'overhead' while the upper right figure views the contoured data from an angle, similar to the views in Figure 11.4. Contours are spaced at equal 'heights' in both cases.

An alternative approach is to define the contours to encompass prespecified amounts of the data. In Figure 11.6 the lower left figure shows the contours that enclose 25%, 50% and 75% of the data points. This exploits an idea of Bowman and Foster (1993) who base the contouring on the ranked heights of the density estimate at each point. Each of the three figures noted so far clearly identifies three groups in the data. An advantage of Bowman and Foster's approach, used by them, is that it is readily applied to pre-defined sub-groups within the data. The lower right figure in Figure 11.6 shows the 75% contours for each of the three groups defined by BV. This shows that their group A is reasonably distinct, but that group B splits into two disjoint groups, one of which is coincident with group C.

A potential theoretical problem with the methods used here is that the tails of the distribution can have an undue influence on the results obtained. One response to this is to use an adaptive kernel density estimate in which the window width varies according to the location of a point within the density. This can be programmed in MATLAB and has been applied to the data used here, but has little effect on the results obtained.

11.5 Alternatives To MATLAB

MATLAB was used in this investigation for a number of reasons; it was available; one of us (CCB) has considerable experience of its use; and it was suited to the task to hand. Other general advantages are that it provides an integrated environment within which all the analyses can be undertaken with relative ease, and is a commercial package that is well supported and quite widely available in the University sector at least. A disadvantage for some is that for the applications described here some programming, and hence a knowledge of the underlying mathematics, is needed.

Since undertaking most of the work described here the S-PLUS package (Becker *et al.* 1988; Chambers & Hastie 1993) became available to us and can also be used for some of the analyses described.

S-PLUS has a similar 'flavour' to MATLAB but is designed with statistical applications in mind, so that some applications, such as univariate kernel density estimation, are immediately available without the need for programming. Facilities for the interactive rotation and labelling of three-dimensional plots are superior to those in MATLAB. For more sophisticated analyses a similar programming effort to that required in MATLAB may be needed.

Both packages support a variety of computing platforms, including both Windows on PCs and Unix based systems. (The software development work reported here was produced on a 33MHz 486 PC using MATLAB version 4 for Windows.) Both packages are based around a Graphical User Interface (GUI) involving windows, menus and pointers. As a result the 'feel' of the packages is similar, and is relatively independent of the particular platform being used. However the performance of either package benefits from the greater processing power of a Unix workstation.

An alternative strategy, for those without access to these packages, might be to carry out the PCA in some other package such as MINITAB or STATGRAPHICS, and save the component scores. These can then be used as input into graphics packages that produce results identical to, or analogous to, those presented in this paper.

11.6 Conclusions

The main purpose of this paper has been to illustrate some possible approaches to the display of results from a PCA that, in some cases, are quite uncommon and have certainly had little or no application in archaeological practice. The usefulness of the techniques discussed, in comparison to the presentation of the usual component plots, may be open to question. For smallish data sets with clear structure some of the techniques described add little to what is usually done; for large data sets our experience with the example given here, and other data, encourages us to believe that the ideas have some merit. The usefulness, for interactive exploratory analysis, of colour for labelling points is unquestionable, though publishing constraints make the publication of such analyses difficult.

That some of the techniques are not widely used outside the mainstream statistical literature is undoubtedly related to the fact that easily used software, allowing routine application, has not been widely available. An assessment of the impact of interesting, and potentially useful, developments reported in past proceedings of Computer Applications in Archaeology conferences (e.g. Bayesian methodology; graphical modelling) would probably identify this as a common problem. The development of commercially available, and supported, packages such as MATLAB and S-PLUS is likely to encourage the more widespread use of 'modern' methods for data analysis and display.

Appendix

Listings can be provided for univariate and bivariate kernel density estimation, and for perspective and contour plots of the latter. These can be obtained by e-mailing the second author whose address is

mat3beardcc@ntu.ac.uk

References

BARRERA, J. & VELDE, B. 1989. 'A Study Of French Medieval Glass Composition', *Archaeologie Medievale* , 19, 81–130.

BAXTER, M. J. 1994. *Exploratory Multivariate Analysis in Archaeology*, Edinburgh University Press, Edinburgh.

BAXTER, M. J., BEARDAH, C. C. & FREESTONE, I. C. (in preparation), 'A re-analysis of some French Medieval glass compositions with an application'.

BOWMAN, A. & FOSTER, P. 1993. 'Density based exploration of bivariate data', *Statistics and Computing*, 3, 171–7.

BECKER, R. A., CHAMBERS, J. M. & WILKS, A. R. 1988. *The New S Language*. Wadsworth & Brooks/Cole, Pacific Grove, California.

CHAMBERS, J. M. & HASTIE, T. J. (eds.). 1993. *Statistical Models in S*. Chapman and Hall, New York.

FREESTONE, I. C. (n.d.). *An analytical investigation of glass from the retable of Westminster Abbey*.

HODSON, F. R. 1969. 'Searching for structure within multivariate archaeological data', *World Archaeology*, 1, 90–105.

HUGHES, M. 1991, 'Provenance studies of Spanish Medieval tin-glazed pottery by neutron activation analysis', in Budd, P., B. Chapman, C. Jackson, R. Janaway, & B. Ottoway, (eds.), *Archaeological Sciences 1989*, Oxbow Books, Oxford

RAURET G., E. CASASSAS, F. X. RIUS & M. MUNOZ 1987. 'Cluster analysis applied to spectrochemical data of European Medieval stained glass', *Archaeometry*, 29, 240–9.

SCOTT, D. W. 1992. *Multivariate Density Estimation*, Wiley, New York.

SILVERMAN, B. W. 1986. *Density Estimation*, Chapman and Hall, London.