Sample Size Problems with Multivariate Archaeometric Data

S. Westwood
Dept. of Mathematics, Statistics and Operational Research. The Nottingham Trent University
Clifton Campus, Nottingham NG11 8NS, U.K.
E-mail: sw@maths.ntu.ac.uk

M.J.Baxter
Dept. of Mathematics, Statistics and Operational Research. The Nottingham Trent University
Clifton Campus, Nottingham NG11 8NS, U.K.
E-mail: mjb@maths.ntu.ac.uk

C.C. Beardah
Dept. of Mathematics, Statistics and Operational Research. The Nottingham Trent University
Clifton Campus, Nottingham NG11 8NS, U.K.
E-mail: c.beardahs@maths.ntu.ac.uk

Introduction

In this research note, a brief account is given of some recent work, investigating sample size requirements for some specific archaeometric problems.

The original motivation for this work arose in the context of an investigation into the normality, or otherwise, of lead isotope ratio fields (Baxter, 1998). A specimen from an ore-body can be characterised by measurements of three lead isotope ratios, and n such specimens can be used, to estimate the lead isotope field for the ore-body. This is a three-dimensional construct. In using the data on such fields, in provenancing studies for example, it is sometimes assumed that fields have a trivariate, normal distribution (Sayre, et al., 1992).

The analyses in Baxter (1998), suggested that normality was the exception rather than the rule. This may not previously have been recognised, because the sample sizes typically available, in conjunction with the methods of statistical analysis used, were too small to detect departures from normality. It is thus of interest, to ask how large sample sizes need to be to detect non-normality. This is particularly so, since there seems to be widespread acceptance, that a well selected sample of size 20, from an ore-body, is an 'agreeable minimum' (Pollard and Heron, 1996). We shall argue that this is only so, if the lead isotope field is normally distributed; if it is not, somewhat larger samples are needed to detect the non-normality.

The issue of sample size is of concern in a more general setting. In practice, sample sizes are often determined by practical considerations, such as cost of analysis or availability of specimens. This is so, in the study of artefact compositional data where, for example, the reported use of samples greater than 100 is uncommon. Often, data are analysed using multivariate statistical methods such as cluster or principal component analysis (PCA), that result in graphical output, designed to show structure (e.g. groups) in the data, or its absence. Where structure is very obvious, it is likely that relatively small samples will be successful in displaying this (and also the case, that multivariate methodology may be unnecessary). With less obvious structure, larger samples may be needed, and the question, 'how large?', is then of interest.

In a sense this is an impossible question to answer, since the answer depends on the precise, but unknown, form of the structure, that the data are designed to investigate. Nevertheless, it may be possible to suggest guidelines, and in the remainder of this paper, we outline some possible approaches, that we have explored.

Sample sizes for lead isotope data

The statistical tests used in Baxter (1998), for several of the larger data sets, published by Stos-Gale et al. (1996), suggested that the data were non-normal. Given that this is established, kernel density estimates (KDEs) provide a useful tool for displaying the form of non-normality. KDEs can be thought of, as smoothed histograms and are discussed, in an archaeological context, in Baxter, et al. (1997). In Figure 1, a KDE is shown for one of the univariate ratios for the Lavrion field, based on 59 observations; more generally, such a KDE might be based on a linear combination of the ratios. The KDE looks non-normal and is bi-modal.

In asking what sample sizes are needed to detect structure in multivariate data sets, the term 'structure' needs to be defined. One model, for lack of structure, is that the data have a (multivariate) normal distribution. Here, we shall define structure to be a departure from normality, that manifests itself as multi-modality. This is possibly restrictive, but many published analyses of compositional data are primarily interested in this kind of structure (as shown in PCA or discriminant analysis plots, for example).

The problem then is, given a sample from a population with a multi-modal distribution, what sample size is needed to detect the multi-modality? The answer clearly depends on the form of the multi-modality, and we have approached this in two ways.

(a) Non-normal, multi-modal distributions have been simulated, using mixtures of normal distributions. The 'populations' generated in the simulation are repeatedly sub-sampled, for some fixed sample size, and the number of occasions, on which multi-modality is detected, is determined. This exercise is repeated for different sample sizes, to find at what point the detection of non-normality becomes reasonably certain.

(b) Real data sets, in which multi-modality is evident, are sub-sampled in a similar way, to determine at which sample size there is a failure to detect multi-modality.

It is necessary to establish a methodology, for determining whether or not a specific sample exhibits multi-modality, and two approaches have been used. In the first approach, tests of normality have been used, and the power - the proportion of times that the test correctly rejects the null hypothesis of normality - for different sample sizes, and kinds of multi-modality investigated. A test of normality based on a KDE estimate, developed by Bowman (1992), for the univariate case, and extended to the multivariate case, by Bowman and Foster (1994), has been used. In the second approach, a KDE of a sample is obtained and the number of modes counted. This latter approach has presented a number of difficulties, that are discussed in section 3.

Repeatedly, taking sub-samples, of size 20 from data on the univariate ratio shown in Figure 1, suggested that the power of the test of normality was about 20%. To achieve a power of 70%, a sample size of around 45 was needed. This result is consistent with those, arising from the experiments conducted on simulated mixtures, where a sample size of 20 was inadequate for detecting non-normality, in the presence of significant overlap between the components of a mixture.

Gale et al. (1997), published data from the Larnaca axis in Cyprus, for 73 specimens, and discussed this in Stos-Gale, et al. (1997). The specimens came from nine different deposits, and bivariate plots of the ratios showed a clearly non-normal, multi-modal structure, associated with the different deposits. Conducting a similar exercise, to that described in the previous paragraph, shows that for 70% power, a sample size of over 30 is needed.

The importance of these results, is that they suggest that sample sizes recommended in the literature, may be too small. If the data for a field are normal, then a sample size of 20 may be adequate, to delineate the field, but if data are non-normal, then much larger samples may be needed, to detect and display this.

Mode counting

In the analyses just discussed, the sampled populations were multi-modal, and tests for normality were used to detect this. It is possible that samples from the population will be detected as non-normal, but will not necessarily exhibit the multi-modality of the population. In practice, an assessment of structure would often be made more directly, on the basis of visual inspection of the data, in the form of a histogram, KDE or bivariate plot. In other words, after the creation of some visual display, modes are counted. It is of some interest to ask if this approach, as opposed to formally testing normality, gives rise to similar conclusions.

In principle, it should be possible to repeatedly sub-sample data, from a population known to be multi-modal, count the number of modes in a sample, and, estimate the sample size needed to 'capture' the true modality, some fixed proportion of the time. Putting this idea into practice is a non-trivial problem. To begin with, there is no uniquely 'correct' way of determining the number of modes in a sample. Our approach has been to fit an adaptive kernel density estimate (Silverman, 1986), using a pilot smoothing parameter, determined by a method described in Wand and Jones (1995, 74).

For a single sample, visual inspection is usually sufficient to establish the number of modes, though there are sometimes borderline cases, where the decision is not straightforward. For the kinds of structure we are interested in, small modes at the periphery of a plot, corresponding to a small group of outliers for example, would be discounted in assessing the main structure in a data set. Devising methods of automatically counting modes - given a KDE estimate - is not straighforward, because of the difficulty of establishing rules, mimicking human decision making in a consistent way. Some automatic procedure is necessary, if thousands of simulated data sets are to be inspected.

We have experimented with a number of methods, including the use of neural networks, and this work is still at an early stage. First impressions are that mode counting gives similar, or possibly better, results, compared with testing for normality, in the sense that similar or smaller samples may suffice to detect structure. This may, however, be a consequence of the particular test of normality used, and further investigation is needed.

Multivariate problems

Section 2, and other work not described here, has concentrated on the univariate case. The real challenge is to extend the ideas developed there to the multivariate case, and in this section, some possible approaches are outlined.

The problem is that, of determining what sample sizes are needed to detect multimodality in p-dimensional data sets, where p may be large (> 20 is increasingly common). A direct attack on this problem is unlikely to succeed, because of the 'curse of dimensionality', so that some form of initial data reduction is almost certainly essential. The main approach investigated, so far, has been to perform a PCA, and then to extend the methods used, for the univariate case, to the bivariate PCA plot.

Figure 2 shows a plot of the first two components, in a PCA analysis of about 230 specimens of archaeological glass, using 11 elements. Heyworth (1991) classified the glasses by colour. There are two main concentrations of points on the plot; the dense central concentration consists mainly of light-blue glass, while the less dense cloud, to the left, consists mainly of light-green glass.

The data set, used here, is much larger than many used in practice. For example, of five multivariate analyses reported in four papers in *Archaeometry* 38 (I), four use a sample size of less than 40, three of which are less than 20. Repeatedly, sub-sampling from the component scores, shown in Figure 2, and testing for bivariate normality using the statistic developed by Bowman and Foster (1994), suggests that a sample of size 25, gives a power of about 60%, whereas a sample of size 50, gives a power close to 100%. These sample sizes were used, because Bowman and Foster (1994) provide critical values for them; their results are being

extended, so that intermediate sample sizes can be investigated.

Investigations of simulated data, where the possibilities are much richer, than in the univariate case, and mode counting, have still to be undertaken, as have studies, based on other real and structured data sets.

An alternative to the use of PCA, that is also under investigation, is the use of projection pursuit (PP) methodology, of which PCA is a special case. PP methods have been around for some time (Jones and Sibson, 1987) but, with the exception noted below, do not seem to have been applied to archaeometric problems. The basic idea is simple. Whereas in PCA, linear combinations of the data are chosen to maximise variance, in PP methods they are chosen to optimise some index of 'interestingness'. As Simonoff (1996, 117) notes, normality may be regarded as uninteresting, so any statistic suitable for testing for normality might be used as an index. The idea is illustrated in Baxter (1998), albeit without using the term 'projection pursuit'. The univariate, Shapiro-Wilk statistic is widely regarded as one of the best omnibus tests of normality. The multivariate extension of Malkovich and Afifi (1973) seeks the linear combination of p variables, that minimises the univariate statistic. Baxter (1998) uses the statistic, along with others, to test for trivariate normality, in three-dimensional data sets. The minimising combination identifies a particular view of the data, that can be displayed using a univariate KDE, to visualise the form of non-normality.

In the context of sample size problems, PP potentially provides a 'sharper' view of the data than PCA. If, empirically, this can be shown to be the case, it suggests that smaller sample sizes may be needed, to identify structure, than if PCA is the chosen method of analysis.

Summary

In this paper, we have reported on work - still very much in progress - that is attempting to grapple with the problem of sample size requirements in archaeometric study. Our approach has been based on a mixture of simulation and case studies of real data, and has mainly looked at univariate problems, so far. Results suggest that in one specific area of application - lead isotope ratio analysis - sample size recommendations, commonly given in the literature, may be much too small. The controversy surrounding the interpretation of Cypriot lead isotope data (e.g. Stos-Gale et al., 1997), is at least partially attributable to the inadequacy of the sample size, 43, on which (until recently) interpretations were based. That 43 was inadequate, given the true complexity of the Cypriot field(s), has only become readily apparent with much more data collection.
The work described here is being extended to the more difficult multivariate case, and a number of possible avenues of enquiry have been identified in the paper.

Bibliography

BAXTER, M.J. (1998), "On the multivariate normality of lead isotope fields", (to appear).

BAXTER, M.J., BEARDAH C.C. & WRIGHT R.V.S. (1997), "Some archaeological applications of kernel density estimates", *Journal of Archaeological Science,* 24, pp. 347-354.

BOWMAN, A.W. (1992), "Density based tests of goodness-of-fit", *Journal of Statistical Computation and Simulation,* 40, pp. 1-13.

BOWMAN, A.W. & FOSTER P.J. (1993), "Adaptive smoothing and density-based tests of multivariate normality", *Journal of the American Statistical Association* , 88, pp. 529-537.

GALE N.H., STOS-GALE Z.A., MALIOTIS G. & ANNETTS N. (1997), "Lead isotope data from the Isotrace laboratory, Oxford: Archaeometry data base 4, ores from Cyprus", *Archaeometry,* 39, pp. 237-246.

HEYWORTH, M.P. (1991), *An Archaeological and Compositional Study of Early Medieval Glass from North-West Europe*, unpublished PhD thesis, University of Bradford, UK.

JONES, M.C. AND SIBSON R. (1987), "What is projection pursuit?", *Journal of the Royal Statistical Society,* A150, pp. 1-36.

MALKOVICH, J.F. & AFIFI A.A. (1973), "On tests for multivariate normality", *Journal of the American Statistical Association* , 68, pp. 176-179.

POLLARD, A.M. & HERON C. (1996), *Archaeological Chemistry*, Royal Society of Chemistry, Cambridge.

SAYRE, E.V., YENER K.A., JOEL E.C. & BARNES I.L. (1992), "Statistical evaluation of the presently accumulated lead isotope data from Anatolia and surrounding regions", *Archaeometry,* 34, pp. 73-105.

SILVERMAN, B. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

SIMONOFF, J.S. (1996), *Smoothing Methods in Statistics*, Springer-Verlag, New York.

STOS-GALE Z.A., GALE N.H. & ANNETTS N. (1996), "Lead isotope data from the Isotrace laboratory, Oxford: Archaeometry data base 3, ores from the Aegean, part 1", *Archaeometry,* 38, pp. 381-390.

STOS-GALE Z.A., MALIOTIS G., GALE N.H.& ANNETTS N. (1997), "Lead isotope characteristics of the Cyprus copper ore deposits applied to provenance studies of copper oxhide ingots", *Archaeometry,* 39, pp. 83-123.

WAND M.P. & JONES M.C. (1995), *Kernel Smoothing*, Chapman and Hall, London.

List of Figures in CD-ROM.

Figure 1. An adaptive kernel density estimate of the distribution of the $^{206}Pb/^{204}Pb$ lead isotope ratio for the Lavrion field. (Data source: Stos-Gale et al., 1996)

Figure 2. Principal component, of standardized data, based on the chemical composition of 227 specimens of glass from Saxon Southampton. Plotting symbols distinguish between light-blue and light green glass. (Data source: Heyworth, 1991)