ProQuest Number: 10183532

ProQuest 10183532

Sample Sizes and Related Problems in Multivariate Archaeology

Simon Westwood

A thesis submitted in partial fulfillment of the
requirements of The Nottingham Trent University
for the degree of Master of Philosophy

January 2003

Contents

Abstract

The number of samples available for the statistical analysis of archaeological scientific data, specifically chemical compositional data, is typically small and determined by practical considerations. However the number of variables measured is often high. We investigate a number of sample size problems that can arise in the multivariate analysis of compositional data with limited sample sizes.

Initially current trends in published articles on the multivariate analysis of compositional data are reviewed. We report that analyses are typically undertaken on data with between 8 and 20 variables with sample sizes in the region of 30 to 100, and are commonly analysed using principal components analysis or cluster analysis. We investigate the claims that projection pursuit is a 'sharper' tool than principal components analysis for the analysis of multivariate compositional data, but reject it on the grounds that a) limited sample size can result in the detection of spurious structure frequently, and b) the length of time required to fully examine results is excessive.

In trivariate lead-isotope ratio studies it is generally accepted that a minimum of 20 observations is adequate to define a lead-isotope field. Using simulated and actual data with a greater number of observations than have previously been available, we question this assumption and demonstrate that 40 or more observations are required to demonstrate non-normality in some cases. Our approach to determining sample sizes in lead-isotope data comprises direct testing of normality and assessment of modality. Our method of assessing modality is to generate kernel density estimates of data and count modes, this also allows us to perform a comparison between different methods of kernel density estimation. This work suggests that adaptive kernel density estimates are better able to estimate the density of an unknown population. We use this insight to extend a formal test of modality using kernel density estimation that provides more interpretable results.

Acknowledgements

My thanks go to my supervisors, Dr. Mike Baxter, Dr. Christian Beardah and other members of the former Department of Mathematics, Statistics and Operational Research at The Nottingham Trent University who have provided much support and encouragement during this research.

Chapter 1

Introduction

In scientific archaeology, or archaeometry, data are often collected on the chemical composition of artefacts. Such data may be used to investigate the relationship between the chemistry of an artefact and the site of its manufacture or of the raw materials used in the manufacturing process. This can, for example, be used to draw archaeological inferences about trade patterns if the origin, and hence distribution, of artefacts can be inferred from their chemistry.

Typically data are multivariate and are commonly analysed using multivariate statistical methods that result in graphical output designed to show structure (for example groups). Typically samples may be collected from a variety of sources and analysed statistically in the hope, and expectation, that they are chemically separable in multivariate space and that this can be demonstrated in a low number of dimensions. The techniques used include, but are not limited to, principal components analysis and cluster analysis which can be used to examine low dimensional views or projections of the data for structure. It is common that the sample size is determined by practical considerations, such as the cost of analysis or the availability of specimens, and as a consequence the number of samples available is often quite limited. Where structure is very obvious it is likely that relatively small sample sizes will be successful in displaying this. With less obvious structure larger samples may be needed.

The reliability of the conclusions that can be drawn from data, or the ability to detect patterns in the data having archaeological meaning, can be compromised by a small sample size. The presence of grouping in data (or its absence) is often a focus of interest, and for large enough samples this is often manifest through the presence of multi-modality of the data which can have archaeological meaning. This leads on to

1

the question of 'what sample size is needed?'

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 being used to investigate. Nevertheless our ultimate aim is to provide guidelines.

In chapter 2 we begin by taking an overview of current and previous trends in the analysis of multivariate archaeological data. This is provided by a review of published analyses making use of multivariate statistical techniques. The review is broken into several sections, which gives insight into the sample size typically used in analyses, the number of variables that are commonly used, and the techniques that are used to analyse the data.

Chapter 3 defines common notation used throughout the thesis and introduces some of the commonly used methods of multivariate analysis. We define the Kernel Density Estimate (KDE), which is a tool for estimating the density function of a population from which a set of data have been sampled. The density function gives an estimate of the distribution of a population which can be used to investigate the possibility of structure. The concept of exploratory data analysis is introduced, which encompasses the commonly used methods of multivariate data analysis.

Lead-isotope ratio studies in archaeology are a specific area of interest (Westwood, Baxter and Beardah, 1998; Baxter, Beardah and Westwood, 2000), where a sample of trivariate observations can be used to define a 3-dimensional lead isotope field. Previously there has been debate regarding the assumption of trivariate normality that is often made for some of the calculations performed for provenancing artefacts from an unknown origin. The consensus has been that 20 samples are sufficient to define a lead isotope field (Pollard and Heron, 1996), but we shall argue that this is only the case if it can be assumed that the field has a normal distribution. This is rarely tested in a satisfactory way and sample sizes are often too small to permit adequate testing. Stos-Gale *et al.* (1996) and Gale *et al.* (1997) published data with larger sample sizes than had previously been available, and this allows the normality assumption to be tested. In chapter 4 we make use of both simulated and real data to investigate the sample sizes that are required to detect structure in the data using both tests of normality and by counting modes in data directly. 'Structure' in this instance

is interpreted as multi-modality showing a significant departure from normality. We demonstrate that a sample size of 20 is usually insufficient to detect what can be quite serious departures from normality.

One approach to investigating sample size requirements in lead-isotope ratio studies involves counting the number of modes in KDEs of the data, however we conclude that this approach raises a number of issues and that more formal methods of investigating modality should be investigated. Silverman (1981) proposed a test of univariate modality that makes use of KDEs which we investigate in chapter 5. Experiences gained during the mode counting work undertaken for lead-isotope data suggests that adaptive KDEs provide a more accurate estimate of distribution of the population that data are sampled from than non-adaptive KDEs. We propose an 'adaptive' version of the test of Silverman (1981) and investigate its performance in relation to the original test of Silverman (1981).

In chapter 2 we find that many of the analyses of artefact compositional data make use of principal components analysis to generate a graphical representation of multivariate data, which is hoped will show structure. Jones and Sibson (1987) describe principal components analysis as 'something of a blunt instrument' and make the suggestion that projection pursuit is a sharper tool for the analysis of data. With the sample sizes typically available, we investigate the possibility of using projection pursuit instead of principal components analysis as a tool to show structure in data. In chapter 6 we investigate the claims that projection pursuit is 'a sharper tool' for a number of multivariate applications. In particular, the work on lead-isotope ratio studies in chapter 4 makes use of a test of multivariate normality which seeks the most non-normal one-dimensional view of data. This can be viewed as a form of projection pursuit and is discussed in chapters 4 and 6.

Chapter 2

A Review of the use of Multivariate Analysis for Artefact Compositional Data

2.1 Introduction

Prior to addressing any specific issues relating to sample size, we undertake a review of the current trends in the use of multivariate statistical methods in the field of Archaeology, and more specifically archaeometry. Our intention is to gain an understanding of the kinds of sample sizes and numbers of variables which are typically made available for statisticians to perform analyses on. The review also gives us opportunity to investigate trends in the use of multivariate statistics over time, as well the kinds of materials analysed and the methods of chemical analysis which are used.

The journal *Archaeometry* is the leading journal in the field of chemical analysis of archaeological data and contains a total of 69 articles between 1975 and the end of 1999 that report on the use of multivariate statistics, which are used as the basis for this review. Details of the 69 articles are presented in Table 2.1. References for the 69 articles used in the review are listed in the second half of the bibliography, separately from other references. Although other journals contain articles on the statistical analysis of archaeometric data, *Archaeometry* provides the richest source of articles on the subject. Prior to 1975, very few accounts of the use of multivariate statistics are published in archaeological journals. The review spans a 25 year period, this enables us to investigate trends over five 5 year intervals.

Within the field of archaeometry, the extraction of samples and subsequent chemical

Paper	Chemical analysis techniques used and variables measured	Materials studied	Sample size range	Dimension range	Statistical techniques	No. analyses
Bieber et al. (1976)	NAA (17)	Ceramics	63	17	CA	1
Hammond, Harbottle and Gazard (1976)	NAA (21)	Ceramics	88 - 91	19 - 21	CA	2
Attas, Yaffe and Fossey (1977)	NAA (19)	Ceramics	113	19	CA, DA	2
Ericson and Kimberlin (1977)	NAA (16)	Stone / Rock	65	13 - 16	CA, DA	4
Alvey and Laxton (1978)	XRF (10)	Ceramics	54	10	CA	1
Lambert, McLoughlin and Leonard (1978)	XPS (10)	Ceramics	71	10	CA	1
Birgül et al. (1979)	XRF (14)	Ceramics	99 - 400	7 - 14	PCA, CA, DA	4
Christie et al (1979)	AAS (4), XRF (6)	Glass	19	10	PCA, CA	2
Hedges and Salter (1979)	Electron Microprobe Analysis (17)	Metals	22 - 25	17	DA	2
Lambert et al. (1979)	AAS (12)	Misc - Bone	136	12	CA	1
Hatcher et al. (1980)	AAS (9), ICP-OES (9)	Ceramics	24 - 58	9	PCA	3
Krywonos et al. (1980)	NAA (5)	Ceramics	46	5	CA	1
Tubb, Parker and Nickless (1980)	AAS (9)	Ceramics	48	3 - 9	PCA, CA	8
Bimson, Laneice and Leese (1982)	XRF (4)	Stone / Rock	181	3	DA	1
Blasius et al. (1983)	XRF (8 in total), AAS	Ceramics - Tiles	95	3	CA	1
Burmester (1983)	Mass spectrometry	Misc - Lacquers	46	12	CA, DA	2
Craddock et al. (1983)	AAS (7), Colorimetric methods (1 in total)	Metals	271	7	DA	1
Philip and Ottaway (1983)	Measurements	Misc - Dimensions	32 - 40	7	CA	2
Pike and Fulford (1983)	NAA (11)	Ceramics	14 - 42	11	PCA	2
Carter and Frurip (1985)	XRF (7), Measurements (5)	Metals	150 - 245	5 - 13	DA	4
Kuleff, Djingova and Djingov (1985)	NAA (27)	Glass	30	25 - 27	CA	2
Cox and Gillies (1986)	XRF (12)	Glass	16 - 27	12	CA	2
Schubert (1986)	XRF (27)	Ceramics	10 - 12	12 - 27	CA	6
Newman and Nielsen (1987)	XRF (13)	Stone / Rock	34	13	CA	1
Rauret et al. (1987)	AAS (2), ICP-OES (11 in total), Flame Photometry	Glass	44	5 - 6	PCA, CA	4
Kilikoglou, Maniatis and Grimanis (1988)	NAA (20)	Ceramics	19 - 49	20	CA	4

Paper	Chemical analysis techniques used and variables measured	Materials studied	Sample size range	Dimension range	Statistical techniques	No. analyses
Mello, Monna and Oddone (1988)	NAA (18)	Stone / Rock	74	19	DA	1
Mommsen, Kreuser and Weber (1988)	NAA (14)	Ceramics	50 - 68	11 - 14	CA	3
Topping and Mackenzie (1988)	NAA (at least 9)	Ceramics	36	9	CA	1
Baxter (1989)	XRF (12)	Glass	27	12	PCA	1
Mirti et al. (1990)	AAS (8) ICP-AES (2) FES (2)	Ceramics	45	12	PCA, CA	4
Baxter (1991)	AAS (4) XRF (6) XRF (12) ICP-OES (15)	Glass	19 - 59	9 - 26	CA	16
Capannesi et al. (1991)	NAA (19)	Ceramics	50	8 - 19	PCA, DA	4
Gunneweg et al. (1991)	NAA (26)	Ceramics	14 - 81	14	CA	2
Baxter (1992)	ICP-OES (15)	Glass	44	7 - 15	CA, PCA	4
Bello and Martin (1992)	FES (20) AAS (2)	Stone / Rock	41	22	CA, DA	1
Djingova and Kuleff (1992)	NAA (32)	Glass	20 - 34	29 - 31	CA, DA	5
Sayre et al. (1992)	LIA (4)	LIA	22 - 43	3	DA	2
Glinsman and Hayek (1993)	XRF (9)	Metals	202	9	CA, DA	1
Kallithrakas-Kontos (1993)	PIXE (11)	Metals	41 - 53	8	CA	2
Kerschner et al. (1993)	NAA (28)	Ceramics	29	22 - 24	CA, DA	3
Mirti, Casoli and Appolonia (1993)	ICP-OES (15) FES (2)	Glass	55	8 - 14	CA	1
Tangri and Wright (1993)	Simulated compositional data	Misc - Simulated Glass	14		PCA	1
Carter and Powell (1994)	Measurements (35)	Misc - Dimensions	25	4 - 35	DA	7
Holmes et al. (1994)	NAA (23)	Stone / Rock	22 - 85	10-14	DA	4
Neff (1994)	NAA (19)	Ceramics	241	17	PCA	1
Pollard and Hatcher (1994)	AAS (9)	Ceramics	71 - 74	8	PCA, DA	2
Yap and Hua (1994)	XRF (7) AAS (3)	Ceramics	8 - 36	7	CA	4
Argyropoulous (1995)	NAA (20) ICP-AES (20)	Ceramics	59 - 120	20	CA, DA	4
Baxter et al. (1995)	ICP-AES (15)	Glass	115 - 120	14	PCA, DA	4
Mirti et al. (1995)	ICP-AES (9) FES (2)	Ceramics	31	9	PCA, CA	2

6

Paper	Chemical analysis techniques used and variables measured	Materials studied	Sample size range	Dimension range	Statistical techniques	No. analyses
Troja et al (1996)	XRF (12), AAS (1), Titration (1), ICP-OES and NAA (26)	Ceramics	13	28	PCA, CA	2
Broodie and Steel (1996)	AAS (9)	Ceramics	19 - 267	9	PCA, CA	6
Molera et al. (1996)	XRD (22 in total), XRF	Ceramics	88	9	CA	1
Castellano et al. (1996)	ICP-AES (10)	Ceramics	35	10	PCA, CA	2
Taylor and Robinson (1996a)	NAA (24)	Ceramics	33 - 75	24	CA	5
Taylor and Robinson (1996b)	NAA (24)	Ceramics	13 - 93	24	CA	4
Vednrell-Saz (1996)	XRD (16), ICP-AES (5), AAS (3)	Ceramics - Mortars	15 - 24	8	CA	2
Hall and Yablonsky (1997)	XRF (13), EPMA (7)	Glass	14	5	PCA, CA	2
Rotunno, Sabbatini and Corrente (1997)	AAS (16)	Ceramics	66	16	PCA, CA	2
Taylor, Robinson and Gibbins (1997)	NAA (24)	Ceramics	43 - 227	19 - 24	CA. DA	4
Bell and Croson (1998)	ICP-OES (16)	Metals - Slag	25	16	PCA, CA, DA	3
Doherty and Maske (1998)	ICP-OES (29)	Stone / Rock	65	29	PCA	1
Mirti, Aceto and Preacco Ancona (1998)	ICP-OES (13), FES (2)	Ceramics	137 - 157	15	PCA, CA	4
Truncer, Glascock and Neff (1998)	NAA (17)	Ceramics	48 - 81	17	PCA, DA	6
Yu and Miao (1998)	XRF (13)	Ceramics	28 - 66	13	PCA, DA	3
Adan-Bayewitz, Asaro and Giauque (1999)	NAA (21), XRF (17)	Ceramics	34 - 223	12 - 13	PCA	4
Adriaens et al. (1999)	XRF (17)	Powder	9 - 4500	8 - 17	PCA, CA	2
Koh Choo et al. (1999)	XRD (10)	Ceramics	63	8 - 10	PCA	2

Table 2.1: Papers from the journal *Archaeometry*, published between 1975 and 1999, that report on the use of multivariate statistical techniques. The table details the technique used to determine chemical composition, number of variables each method of chemical analysis determines, material being studied, range of sample size, dimensions, number of statistical analyses reported and multivariate techniques used in statistical analysis. A key to abbreviations is shown in table 2.2.

7

Abbreviation	Explanation
AAS	Atomic Absorption Spectrometry (Pollard and Heron 1996:26-31)
CA	Cluster Analysis (section 3.3.3)
DA	Discriminant Analysis (section 3.3.2)
EPMA	Electron Probe Micro Analyzer
FES	Flame Emission Spectrometry
ICP-OES / ICP-AES	Inductively Coupled Plasma Optical/Atomic Emission Spectrometry (Pollard and Hatcher 1986:31-36)
LIA	Lead Isotope Analysis (chapter 4)
NAA	Neutron Activation Analysis (Pollard and Heron 1996:54-61)
PIXE	Proton-Induced X-Ray Emission (Pollard and Heron 1996:53-54)
PCA	Principal Components Analysis (section 3.3.1)
XRD	X-Ray Diffraction
XRF	X-Ray Fluorescence Spectrometry (Pollard and Heron 1996:41-49)

Table 2.2: Key to abbreviations used in Table 2.1.

analysis of specimens can be problematic and often determines the number of samples available for statistical analysis. For example Pollard and Heron (1996:302) discuss some of the problems associated with extracting compositional data from metallic artefacts. Other factors which may influence the number of samples available include the cost of performing the chemical analysis and physical availability of samples.

This chapter is broken into several sections, initially the types of materials studied and any trends in the methods used to determine their chemical composition are reported on. Later sections discuss trends in sample sizes, numbers of dimensions used and the statistical techniques that are used to analyse data.

2.2 Materials and their chemical analysis

The papers detailed in table 2.1 can be categorised as analyses on one of the following materials:

- ceramics and clays;

- glass;

- stone and rock;

	1975-79	1980-84	1985-89	1990-94	1996-99	Total
NAA	4	2	5	6	7	24
ICP-OES / ICP-AES		1	1	4	9	15
AAS	2	4	1	5	4	16
FES				3	2	5
XRF	3	2	5	4	6	20
Other	2	3	2	4	5	16
Total for Period	11	12	14	26	33	96

Table 2.3: Frequency of the use of methods of chemical analysis from papers reporting on the use of multivariate methods in the the journal *Archaeometry* between 1975 and 1999.

- metals;

- miscellaneous analyses, for example analysis of bone, lacquers and geometric measurements (such as dimensions of artefacts).

Ceramics are the most frequently analysed material, accounting for 55% of papers within the review, the analysis of glass is discussed in 16% of articles, however the papers reporting on the analysis of glass commonly discuss several statistical analyses. The analysis of stone and rock, metals and miscellaneous materials are reported on in far fewer articles, stone and rock are discussed in 13% of papers and metals 9%.

Table 2.3 shows the frequency with which methods of chemical analysis are used within the papers within the review, the table is broken into five 5 year periods to allow trends over time to be investigated. It is not uncommon that a paper details the use of several methods of chemical analysis to generate data on the composition of specimens. Detailed descriptions of each of the chemical techniques can be found in Pollard and Heron (1996:20-80).

Neutron Activation Analysis (NAA, Pollard and Heron 1996:54-61) is used throughout the review period, however in later years the proportion of articles using it decreases as other (cheaper) techniques become more commonly used. Prior to 1985, NAA was used for nearly one third of the reported chemical analyses, however in the 1990-94 period, it is only used in 23% of papers. Despite the availability of other methods, NAA is still the most frequently used method over the same period. Pollard and Heron (1996, 54) suggest that NAA was the standard method for producing multi-chemical analyses until the advent of ICP and PIXE in the early 1990's. X-Ray Fluorescence

Spectrometry (XRF, Pollard and Heron 1996:41-49) follows a similar trend in use to NAA and accounts for nearly as many analyses.

Inductively Coupled Plasma Optical/Atomic Emission Spectrometry (ICP-OES or ICP-AES, Pollard and Hatcher 1986:31-36) is rarely used prior to 1990, however the technique has since gained in popularity. Similarly, Flame Emission Spectrometry (FES) is not reported on prior to 1990. Atomic Absorption Spectrometry (AAS, Pollard and Heron 1996:26-31) appears to be used throughout the review period. Its use is more popular in later articles (after 1990) however proportionally it accounts for fewer of the total analyses.

The method of chemical analysis of ceramics follows similar trends to those discussed above. NAA is used more frequently than other techniques with ICP-OES and XRF being used more frequently in later years. There are far fewer published analyses on glass than ceramics, however ICP-OES and XRF are used more frequently, in part because very few articles report on the analysis until after 1985.

2.3   Sample sizes

Table 2.4 shows the average sample size (n), by year, reported on in papers within the review, this data is also presented graphically in figure 2.1. An analysis by Adriaens *et al.* (1999) stands out immediately due to the $n = 4500$ samples used, we have removed this value of n from our analyses.

Of the remaining 198 samples, 86% of analyses are performed on fewer than 100 samples and 15% are performed on fewer than 30 samples. The mean sample size is 63 with a median of 44.

Figure 2.1 shows the reported sample sizes graphically. Figure 2.1a shows a box and whisker plot of sample sizes used by year, figure 2.1b shows the same information as a dotplot. Initial observation suggests that there are no obvious trends in sample size used over time. Pre 1986 there is a large degree of variation in sample sizes with fewer analyses than post 1986. Figure 2.1b shows that after 1986 the upper quartile range generally increases over time which suggests that, for some analyses, more samples are available. Despite this the mean sample size tends to vary considerably.

Figure 2.1c shows a box and whisker plot of sample sizes recorded in each of five

Year	Average n	Average d	Average n/d	Average d Measured	No. Analyses
1976	81	19	4.2	19	3
1977	81	16	5	17.5	6
1978	63	10	6.3	10	2
1979	135	10	17.8	13	9
1980	47	6	9.2	11	12
1981	-	-	-	-	-
1982	181	3	60.3	4	1
1983	73	9	11.7	9	8
1984	-	-	-	-	-
1985	158	15	19.7	19	6
1986	14	16	1	19	8
1987	42	7	7	13	5
1988	41	16	2.7	15	9
1989	27	12	2.3	12	1
1990	45	10	4.7	12	4
1991	41	13	3.7	27	22
1992	35	18	3.9	18	12
1993	57	14	5.3	16	8
1994	47	10	4.8	19	18
1995	89	15	5.8	22	10
1996	67	17	6.1	22	22
1997	70	16	4.2	20	8
1998	74	16	4.7	18	17
1999 [1]	58	11	5.2	22	7

[1] Adriaens *et al.* (1999) perform an analysis on $n = 4500$ with $d = 8$. This value has been removed as an outlier.

Table 2.4: Average sample sizes (n), number of variables (d) and mean n/d ratio used in multivariate statistical analyses reported on in the journal *Archaeometry* between 1975 and 1999.
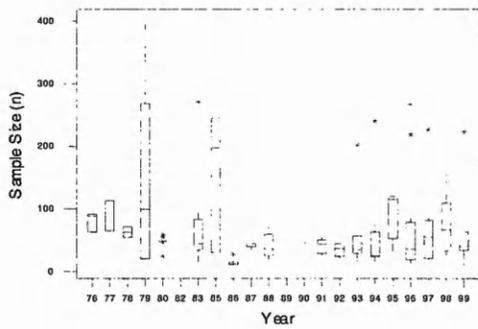
(a) Box and Whisker plot of n by year
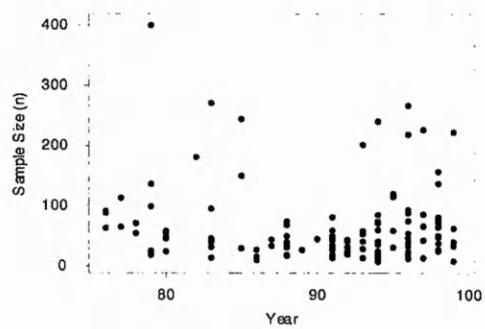
(b) Dotplot of n by year
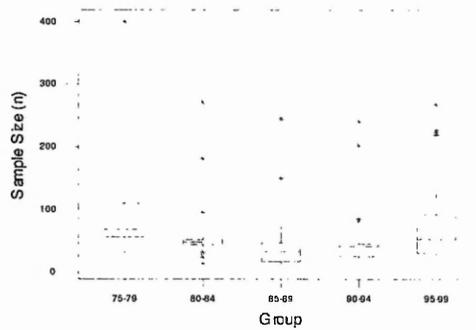
(c) Box and Whisker plot of n by 5
year period

Figure 2.1: Plots showing the sample sizes reported in articles making use of multi-variate statistical analyses in the journal *Archaeometry* between 1975 and 1999.

Year	Average n for Ceramics	Average n for Glass
1975		
1976	81	
1977	113	
1978	62	
1979	249	19
1980	47	
1981		
1982		
1983	50	
1984		
1985		30
1986	11	21
1987		44
1988	37	
1989		27
1990	45	
1991	49	37
1992		34
1993	29	58
1994	66	
1995	70	117
1996	67	
1997	88	14
1998	85	
1999	76	

Table 2.5: Average sample sizes (n) for analyses on ceramics and glass reported on in the journal *Archaeometry* between 1975 and 1999.

5 year periods. The intention is to assess if there are any general trends over a longer period of time than a year. Post 1985 there appears to be a slight upwards trend in the average sample size with the lower quartile range also increasing slightly which suggests more samples are available for more analyses. Pre 1985 the average sample size is greater than in later years. Generally the spread of n values used for statistical analyses tend to be fairly constant over 5 year periods.

For analyses of ceramics and glass, similar variation in sample sizes is evident (table 2.5).
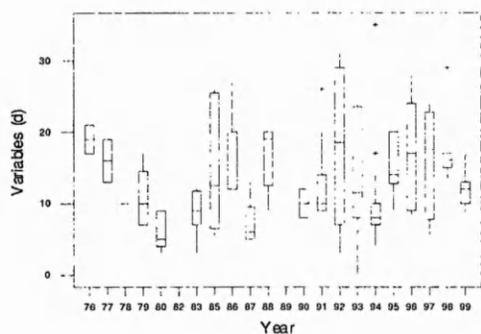
(a) Box and Whisker plot of d by year
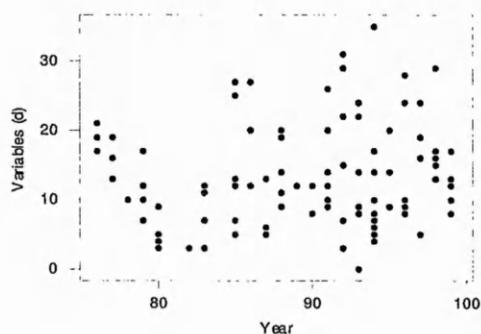
(b) Dotplot of d by year

Figure 2.2: Plots showing the number of variables (d) by year used in statistical analyses in the journal *Archaeometry* between 1975 and 1999.

2.4 Numbers of variables

Figure 2.2a shows a box and whisker plot detailing the number of variables, d, used in published statistical analyses. The box and whisker plot shows considerable variation in the number of variables used. The mean number of variables used is 13 with a median of 12. 50% of analyses use between 9 and 17 variables with an overall range of between 3 and 35 variables.

Figure 2.2b shows a dot plot of d against year and suggests a slight tendency to use more variables in analyses in later years. The plot also suggests that since 1990 the majority of analyses use a minimum of 8 variables, with it not being uncommon to see analyses on as many as 20 variables. In 1976 to 1978 there were relatively few analyses undertaken, however all that were published were performed on data with between 10 and 20 variables.

As part of the analysis we have also looked at the number of variables used with different multivariate techniques. Although there are no particular trends over time, of interest is the fact that cluster analyses are commonly performed with the widest range of variables, 50% of analyses using between 8 and 20 variables with a mean of $d = 14$. Principal components analyses are carried out on a much smaller range of variables, with 50% of analyses using between $d = 9$ and $d = 13$ with a mean of $d = 11$.

In the majority of articles, all variables that are measured are used, at some point, in the statistical analysis. In some cases, the approach is to use all variables to gain

14
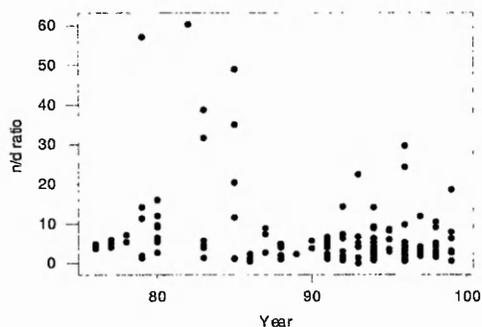
(a) Dotplot of n/d by year
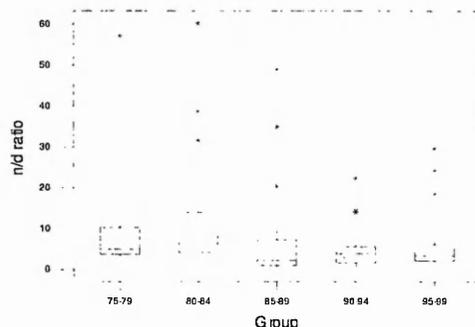
(b) Box and whisker plot of n/d by 5 year period

Figure 2.3: Dotplot of n/d by year and box and whisker plot showing n/d values by 5 year groups.

an initial understanding of the data and then remove variables which are unhelpful to the analysis, for example those that are deemed unreliable because they are close to detection limits. Some elements cannot be reliably measured in small quantities with certain methods of chemical analysis, and are said to be close to detection limits. A large proportion of analyses continue to use all variables throughout the analysis.

2.5 The sample size to variables (n/d) ratio

Some statistical methods cannot be used reliably on sparse data, for example to make use of Mahalanobis distance it is necessary to have more samples than variables and preferably $n > 3d$ or $5d$ within groups for which Mahalanobis distance is calculated. A similar n/d ratio is needed for reliable estimation of correlation coefficients for principal components analysis. As such, analyses undertaken where the sample size to variables ratio is small may be misleading

The mean n/d ratio over the entire period of the review is 6.5, with a median of 4. In total, 40% of articles have an n/d ratio of less than or equal to 3, 50% with less than or equal to 4 and 60% with an n/d ratio of 5 or lower.

Table 2.4 shows the average n/d ratio by year and is presented graphically in figure 2.3a as a dot plot. There is a slight suggestion that as time progresses, there is a tendency for more articles to report on the analyses of data with a smaller n/d ratio, which suggests that for some cases the number of samples and variables used are not

15

increasing at the same rate. However, the overall range of n/d remains fairly constant throughout the review period. Prior to 1985 there is far more variability in the n/d ratio compared to later years, there are also fewer reported analyses.

Figure 2.3b shows a box and whisker plot grouped by 5 year periods. The median by period does not fall below 4 which suggests that even when more variables are used, more samples are also used in the statistical analysis. In some cases the lower quartile (i.e. 25% of observations falling below this value) raises concerns. In the 1985-89 period the lower quartile lies at 1.06 and it is 1.7 during the period 1990-94, analysis of data with such low n/d ratios may be problematic with certain statistics, for example PCA. Of the analyses reported on in the review period, the lower quartile range for analyses using PCA is 2.5, this suggests that an acceptable number of samples and variables are used in the majority of articles.

2.6 Summary of methods

Table 2.6 shows the frequency with which each multivariate statistical method (cluster analysis, discriminant analysis and principal components analysis) are discussed in articles. Cluster analysis is used consistently throughout the review period and is in fact the most popular technique. The n/d ratio for data undergoing a cluster analysis has steadily increased since about 1990 with a minimum value of 2.5, reaching 4 in 1996 and 7 in 1998. Sample sizes have increased slightly since 1995, however the number of dimensions used in analyses has varied considerably.

Discriminant analysis is used less frequently, and usually in conjunction with other techniques, although Bimson, Laneice and Leese (1982) do make use of discriminant analysis without other statistical methods.

Principal components analysis becomes popular from 1979 onwards and grows in popularity from 1994 onwards. The average n/d ratio is generally above 3 and does not drop below 2.

2.7 Discussion

The published use of multivariate statistical analysis has become more common over the previous 25 years, no doubt due to the availability of accessible tools to perform

Year	Papers	Cluster Analysis (CA)	Discriminant Analysis (DA)	Principal Components Analysis (PCA)
76	2	2		
77	2	2	2	
78	2	2		
79	4	3	2	2
80	3	2		2
81				
82	1		1	
83	5	3	2	1
84				
85	2	1	1	
86	2	2		
87	2	2		1
88	4	3	1	
89	1			1
90	1	1		1
91	3	2	1	1
92	4	3	3	1
93	5	4	2	1
94	5	1	3	2
95	3	2	2	2
96	7	7		3
97	3	3	1	2
98	5	2	3	5
99	3	1		3

Table 2.6: Frequency with which the use of multivariate statistical methods are reported on in the journal *Archaeometry* between 1975 and 1999.

statistical analysis. Also, the advent of new methods of chemical analysis, such as ICP-OES, has meant that the number of variables measured, and used in multivariate statistical analysis, has increased during recent years, with the exception of the few published articles prior to 1978. Newer methods of chemical analysis have given opportunity to measure more variables. The n/d ratio suggests that, in general, as the number of variables increases the number of samples has increased at a similar same rate.

The use of principal components analysis has grown in popularity since the mid 1990's, however Jones and Sibson (1987) have made a suggestion that projection pursuit is a 'sharper tool' for the investigation of structure in multivariate data. In chapter 6 we investigate the possibility of using projection pursuit, instead of principal components analysis, to detect structure in data with smaller sample sizes.

Chapter 3

Statistical and Mathematical Background

In this chapter notation and core mathematical and statistical techniques are introduced. Initially, common notation is defined which is used throughout this thesis. Following this, Kernel Density Estimates (KDEs) are discussed.

In later sections, the mathematics of exploratory data analysis (EDA) are briefly introduced, specifically Principal Components Analysis (PCA), Cluster Analysis (CA) and Linear Discriminant Analysis (DA) are discussed.

3.1   Notation

Throughout, $X_1, X_2, \ldots, X_n$ denotes a univariate sample of n observations sampled from a population with unknown density function f. Also, $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ denotes a d-variate data set of n observations with an unknown density f, with $\mathbf{X}_i = \{X_{i1}, X_{i2}, \ldots, X_{id}\}^T$ used to denote the components of $\mathbf{X}_i$.

The multivariate normal probability density function (pdf) is defined as

$$N(\mathbf{x};\ \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-d/2} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right) \qquad (3.1)$$

where $\mathbf{x} \in \Re^d$, $\boldsymbol{\mu}$ is the mean and $\boldsymbol{\Sigma}$ is the covariance matrix of the distribution.

3.2   Kernel Density Estimates (KDEs)

3.2.1   Estimation of an unknown density

Kernel density estimates (KDEs), at their simplest, can be thought of as an alternative to a histogram for presentation of data. They typically provide a smoother represen-
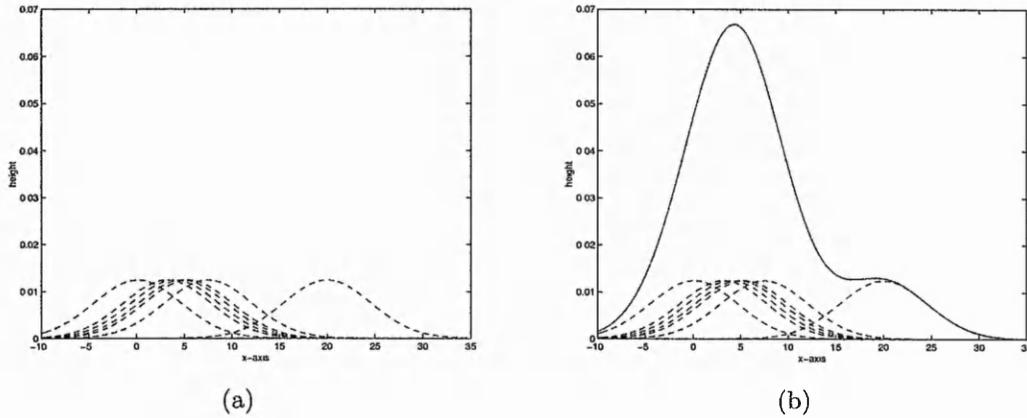
Figure 3.1: A KDE is formed by placing a 'bump' centred on each observation on the x-axis. The shape of the KDE is then given by summing the height of all bumps at each point on the x-axis.

tation of the data, which is more aesthetically pleasing, and their appearance does not depend on the choice of an origin. The use of KDEs allows presentation of multiple density estimates on a single plot which allows for easier comparison of data.

Given n observations, $X_1, X_2, \ldots, X_n$, placed on the x-axis, a univariate KDE can be constructed by placing a 'bump' centred on each X_i and summing the height of the bumps at each point on the x-axis. The shape of each 'bump' is defined by a function, the kernel function $K(x)$, and the spread of the bumps is determined by a window-width or bandwidth parameter, h, which is analogous to the bin width used in construction of histograms and controls the overall smoothness of the KDE. Silverman (1986:43) suggests that the choice of kernel is less important than the choice of h, and in fact it is often desirable to choose $K(x)$ for computational considerations.

Consider a data set with $n = 7$ observations, $\{0.1, 3.2, 4.0, 4.8, 5.4, 7.5, 20.0\}$. Figure 3.1a shows a normal gaussian 'bump' centred on each X_i with a spread of 4.5902 (determined using the normal scale h selection method discussed below). It is then necessary to sum the height of all bumps to determine the height of the resulting KDE, shown in figure 3.1b. For computational ease, it is common to sum the height of the KDE at regular intervals on the x-axis.

Mathematically, a univariate KDE is given by

$$\hat{f}(x; h) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - X_i}{h}),$$ (3.2)

20

where $K(x)$ is a kernel function which satisfies the condition $\int K(x)dx = 1$ and h is the window-width or bandwidth, discussed in the following section. It is common to use a univariate normal pdf as a choice of $K(x)$,

$$K(x) = (2\pi)^{-1/2}\exp(-\frac{1}{2}x^2). \qquad (3.3)$$

Equation 3.2 can be generalised to a d-dimensional kernel density estimate (Scott, 1992, 153)

$$\hat{f}(\mathbf{x}; H) = \frac{1}{n}|H|^{-1/2}\sum_{i=1}^{n}K(H^{-1/2}(\mathbf{x} - \mathbf{X}_i)) \qquad (3.4)$$

where K is a kernel function that satisfies $\int K(\mathbf{x})d\mathbf{x} = 1$ and is often chosen to be the standard multivariate normal density function

$$K(\mathbf{x}) = (2\pi)^{-d/2}\exp(-\frac{1}{2}\mathbf{x}^T\mathbf{x}) \qquad (3.5)$$

and H is a $d \times d$ symmetric positive definite matrix which determines the "smoothness" of the KDE.

3.2.2 Choice of smoothing parameter

For $d = 1$ the choice of h (sometimes known as the window-width) has been widely studied (Wand and Jones, 1995, 58-88).

A smoothing parameter is usually chosen to give a kernel density estimate which is in some way "optimal", generally we want the kernel density estimate to be as 'close' as possible to the true underlying density, f. The usual measure of closeness is mean integrated squared error (MISE, Silverman 1986, 35-6) which gives a measure of the global accuracy of $\hat{f}$ as an estimate of f. MISE is defined as

$$MISE(\hat{f}) = E\int\{\hat{f}(x) - f(x)\}^2dx. \qquad (3.6)$$

Wand and Jones (1995:19) point out that direct use of MISE expressions "depend on the bandwidth in a complex way. This makes it difficult to interpret the influence of bandwidth on the performance of the kernel density estimates". They advocate the use of asymptotic mean integrated squared error (AMISE) which is a large sample approximation of MISE. AMISE expressions are easier to manipulate than those for

MISE and AMISE "allows a deeper appreciation of the role of the bandwidth". AMISE is defined as

$$AMISE(\hat{f}) = \frac{1}{nh}R(K) + \frac{1}{4}h^4\mu_2(K)^2R(f''), \qquad (3.7)$$

where f'' denotes the second derivative of f,

$$R(K) = \int_{x\epsilon\Re} K(x)^2 dx = \frac{1}{2\sqrt{\pi}}$$

and

$$\mu_2(K) = \int_{x\epsilon\Re} x^2 K(x) dx = 1$$

for the normal kernel defined in equation 3.3.

The minimising value of equation 3.7 can be shown to be

$$h_{AMISE} = \left[\frac{R(K)}{\mu_2(K)^2 R(f'')n}\right]^{1/5} \qquad (3.8)$$

by differentiating 3.7 with respect to h and setting the derivative equal to 0 (Wand and Jones, 1995:22).

If the true density f is normal with variance σ^2 and the normal kernel is used, equation 3.8 can be reduced to the Normal Scale Rule, h_{NS},

$$h_{NS} = 1.06n^{-1/5}\hat{\sigma} \qquad (3.9)$$

where $\hat{\sigma}$ is an estimate of σ. For non-normal, multi-modal densities this will over-smooth the data, which can result in real structure being hidden by the estimate (as illustrated in figure 4.7 in section 4.2.2)

An alternative with good properties, that typically results in smaller h, is the 'solve-the-equation' (STE) selection procedure of Sheather and Jones (1991). The solve-the-equation estimate does not assume the true density is normal like h_{NS}, instead it estimates $R(f'')$. Wand and Jones (1995:74) discuss the method in detail, however essentially STE estimation of h is based on equation 3.8. Wand and Jones (1995) prefer to use the notation $\psi_s = \int f^{(s)}(x)f(x)dx$, where $f^{(s)}$ is the sth derivative of f, as it is 'more straightforward to generalise to the multivariate case'. $R(f'')$ is replaced with a kernel density estimate, and h is calculated subject to the condition

$$h = \left[\frac{R(K)}{\mu_2(K)^2\hat{\psi}_4(\gamma(h))n}\right]^{1/5}.$$

In practice, Wand and Jones provide the following procedure for calculating a two-stage solve-the-equation bandwidth estimator, h_{STE}.

1. Estimate ψ_6 and ψ_8 using $\hat{\psi}_6^{NS} = -15/(16\pi^{1/2})/\hat{\sigma}^7)$ and $\hat{\psi}_8^{NS} = 105/(32\pi^{1/2})/\hat{\sigma}^9)$.

2. Estimate ψ_4 and ψ_6 using the kernel estimators $\hat{\psi}_4(g_1)$ and $\hat{\psi}_6(g_2)$ where

$$g_1 = \{-2K^{(4)}(0)/(\mu_2(K)\hat{\psi}_6^{NS}n)\}^{1/7}$$

and

$$g_2 = \{-2K^{(6)}(0)/(\mu_2(K)\hat{\psi}_8^{NS}n)\}^{1/9}.$$

3. Estimate ψ_4 using the kernel density estimator $\hat{\psi}_4(\gamma(h))$ where

$$\gamma(h) = \left[\frac{2K^{(4)}(0)\mu_2(K)\hat{\psi}_4(g_1)}{-\hat{\psi}_6(g_2)R(K)} \right]^{1/7} h^{5/7} \tag{3.10}$$

and

$$\begin{aligned} \hat{\psi}_4(g) &= n^{-1}\sum_{i=0}^{n}\psi_4, \\ &= n^{-2}\sum_{i=0}^{n}\sum_{j=0}^{n}\frac{1}{g}K^{(4)}\left(\frac{X_i - X_j}{g}\right). \end{aligned}$$

4. The bandwidth is then the solution to

$$h = \left[\frac{R(K)}{\mu_2(K)^2\hat{\psi}_4(\gamma(h))n} \right]^{1/5} \tag{3.11}$$

which can be calculated using numerical routines.

Theory is less well developed for choosing H when $d > 1$. Wand and Jones (1993) have studied the case $d = 2$ and conclude that taking

$$H = \begin{pmatrix} h_1^2 & 0 \\ 0 & h_2^2 \end{pmatrix}$$

is often adequate, but that the further simplification $h_1 = h_2$ is not. Separate estimation of h_1 and h_2 using univariate methods is possible, however it is preferable to determine them simultaneously. In section 4.2 the (2-stage) bivariate direct plug-in (DPI2) method has been used, this is similar to the solve-the-equation method discussed above, however a less sophisticated method of estimating $R(f'')$ is used. Wand and Jones (1995:71-74) discuss direct plug-in methods in the univariate case with bivariate generalizations discussed by Wand and Jones (1995:105-108).

23

3.2.3 Adaptive KDEs

In the approaches described so far, the degree of smoothing applied is the same regardless of how sparse or densely distributed the data are. In some applications it may be preferable to smooth more in regions of low density, and this can be done using *adaptive* KDEs.

The basic procedure used for a d-dimensional KDE is that described in Silverman (1986, 91). A univariate adaptive KDE can be generated using the following procedure. Given a *pilot* estimate of h, a KDE is obtained, $\tilde{f}(x)$. This allows *local bandwidth factors*, λ_i, to be defined which determine by how much the density will vary at each point estimated

$$\lambda_i = \{\tilde{f}(X_i)/g\}^{-\alpha} \tag{3.12}$$

where g is the geometric mean of the $\tilde{f}(X_i)$:

$$\log(g) = n^{-1} \sum_{i=1}^{n} \log \tilde{f}(X_i) \tag{3.13}$$

and α is the *sensitivity parameter*, $0 \leq \alpha \leq 1$. We use $\alpha = 1/d$, as used by Breiman, Meisel and Purcell (1977) which ensures that the number of observations caught 'by the scaled kernel' will be approximately the same in all parts of the density.

The KDE can then be recalculated using different values of h at each data point that depend on the sensitivity parameter

$$\hat{f}(x) = n^{-1} \sum_{i=1}^{n} (h\lambda_i)^{-1} K\{(h\lambda_i)^{-1}(x - X_i)\}. \tag{3.14}$$

Returning to the univariate example in section 3.2.1 with $n = 7$ observations. An adaptive KDE can be constructed by initially calculating a pilot estimate of h, here we have used the normal scale rule to select a pilot estimate of 4.5902 as before. The local bandwidth parameters, λ_i, are calculated from 3.12 to give

$$\lambda = \{1.0121, 0.7449, 0.7294, 0.7317, 0.7452, 0.8841, 3.7722\}$$

which define the spread of each of the 7 individual 'bumps' respectively. A bump is placed at each X_i with spread $\lambda_i h$, as illustrated in figure 3.2a. Notice that the bump to the far right of the plot has far greater spread than in the previous example, the intention is to apply more smoothing to areas of low density so that sparse areas do
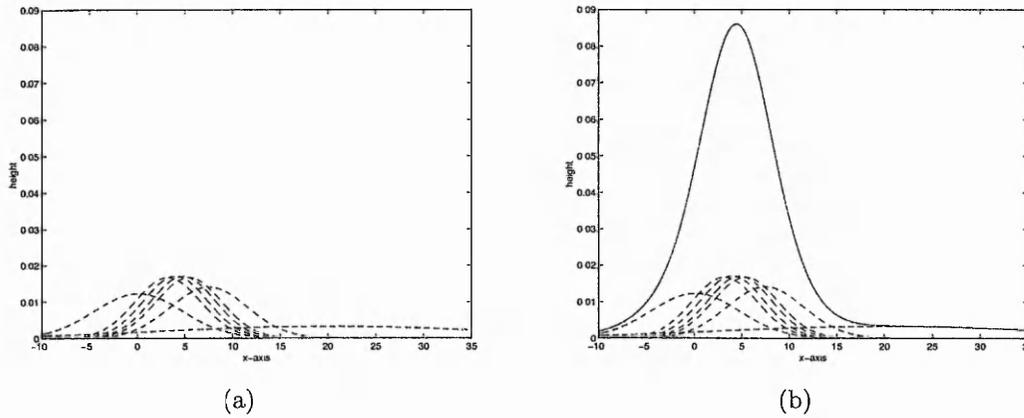
Figure 3.2: An adaptive KDE is formed by placing a 'bump' centred on each observation on the x-axis, the spread of the bump is dependent on the distribution of data. The shape of the KDE is then given by summing the height of all bumps at each point on the x-axis.

not result in multiple modes. The height of the KDE is then found by summing the heights of the individual 'bumps' at points along the x-axis, as illustrated in figure 3.2b.

3.2.4   Software to implement KDEs

Beardah and Baxter (1995) implement routines for the creation of KDEs in the MAT-LAB computer package. Their comprehensive suite of routines implements 1-, 2- and 3-dimensional KDE generation and a range of bandwidth selection routines. Computer routines used in the following chapters are built on the core functionality provided by these routines. A download is available from

    http://science.ntu.ac.uk/msor/ccb/

A similar suite of routines has been made available by Bowman and Azzalini (1997) for the S-Plus (Venables and Ripley, 1999) package. Downloads for Windows and UNIX versions of S-Plus are available from

    http://www.stats.gla.ac.uk/~adrian/sm/

3.3 Exploratory data analysis

Exploratory data analysis (EDA) allows investigation of structure in data, for example to investigate clustering or outliers, often through the use of graphical techniques. Typically EDA will begin with an informal graphical display of the data, such as a draftsman plot (bivariate plots of all combinations of variables) which gives a feel for initial relationships, potential outliers and may suggest approaches for more a detailed investigation of structure.

Although informative, and often giving valuable insight into potential structure within data, these informal graphical techniques are commonly used for initial investigation only. Frequently, techniques such as principal components analysis (PCA), linear discriminant analysis (DA) or cluster analysis (CA) are applied to the data following an initial graphical analysis. EDA techniques commonly attempt to display interesting structure in a limited number of dimensions to allow direct visual investigation into potential structure. These techniques are briefly introduced in the following sections. In all cases, it is common practice to transform the data in some way, so that one variable is no more dominant than any other. Commonly one of the following methods of transformation is used

1. Centring scales the mean of data to be zero;

2. Scaling removes size effects in measurements so that one variable is no more dominant than any other. One possible form of scaling is to subtract the mean, i.e. $Y_i = X_i - \bar{x}$ which results in unit variance. Another possible form of scaling is to divide by the maximum value, i.e. $Y_i = X_i/\max(x)$;

3. Standardisation is when data are scaled and centred

$$Y_i = \frac{X_i - \bar{x}}{s} \qquad (3.15)$$

for $i = 1, 2, \ldots, n$ where $\bar{x}$ is the mean and s^2 is the estimated variance. Standardisation allows data to be compared to an $N(0, 1)$ normal distribution;

4. Logarithmic transformations, for example $Y_i = \ln(X_i)$, are sometimes used to normalise data, this also has the effect of making the variances of data similar.

26
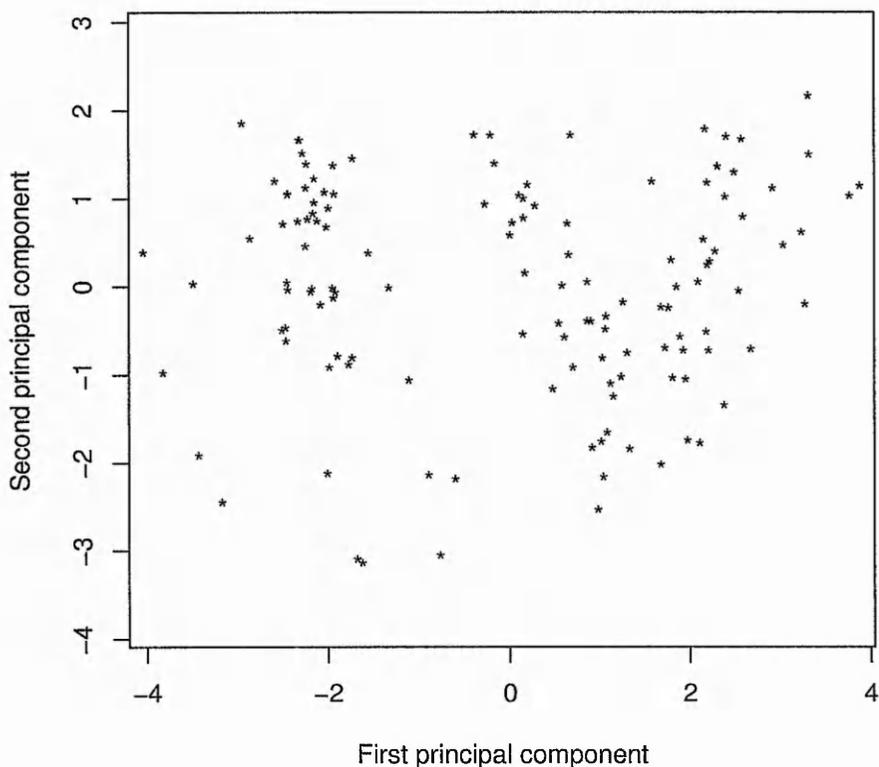
Figure 3.3: Plot of the first 2 principal components of the chemical compositional data of 130 oriental greenwares presented by Pollard and Hatcher (1986). 3 outliers were removed prior to computing the principal components.

3.3.1 Principal components analysis

Principal components analysis (PCA) is one of the simplest EDA techniques. The intention is to transform d variables, $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_d$ to d new variables $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_d$ which are uncorrelated.

Specifically,

$$\mathbf{Z}_i = \sum_{j=1}^{d} \alpha_{ji} \mathbf{X}_j \tag{3.16}$$

for $i = 1, 2, \ldots d$; α_{ji} are scalars constrained such that $\sum_{i=1}^{d} \alpha_{ji} = 1$ and selected such that the $\mathbf{Z}_i$ are uncorrelated, with $\mathbf{Z}_1$ having maximum variance, $\mathbf{Z}_2$ having second maximum variance subject to being uncorrelated to $\mathbf{Z}_1$ and so on.

The hope is that the first few variables will account for the majority of the variance

within the data and as such reveal the main structure present, however as will be discussed in later chapters this is not always the case. Results are commonly presented graphically by plotting the scores $\mathbf{Z}_2$ against $\mathbf{Z}_1$ as illustrated in the example which follows. It is also common to work with scaled data, either standardised, which gives equal weighting to each $\mathbf{X}_i$, using or a logarithmic transformation which has the effect of making the standard deviation of variables more similar.

Pollard and Hatcher (1986) analyse the chemical composition of 133 oriental greenwares which are suspected to have originated from several areas of manufacture. A total of 9 chemical concentrations are measured. Prior to the analysis, 3 clear outliers were removed. When the data are standardised and subjected to a principal components analysis (figure 3.3), two clear clusters are visible within the first two principal components, with the possibility that the larger of the clusters could be separated into two further groups. Pollard and Hatcher relate these two main clusters to the date of manufacture of the greenwares.

3.3.2 Linear discriminant analysis

Discriminant analysis addresses the problem of maximising the difference between groups within data. Given a d-variate data set with g groups, a discriminant analysis can be used to define $g - 1$ functions of the original variables which maximise the difference between known or suspected groups. As with PCA, new variables are defined thus

$$\mathbf{F}_i = \sum_{j=1}^{d} \alpha_{ji} \mathbf{X}_j \tag{3.17}$$

for $i = 1, 2, \cdots, g - 1$. The values of the coefficients α_{ji} are the eigenvalues of

$$(\mathbf{T} - \mathbf{W})\mathbf{W}^{-1}$$

where $\mathbf{T}$ is the total sample matrix of sums of squares and cross products and $\mathbf{W}$ is the within-sample matrix of sums of squares and cross products (Manly, 1994:49). As with PCA, the hope is that the first few functions are sufficient to allow graphical representation of the differences between groups. Unlike PCA, the results of discriminant analysis are not dependant on having standardised data.

Foy (1985) published chemical compositional data for 46 specimens of Medieval French glass found at three known sites. Concentrations of 9 elements are recorded.

28
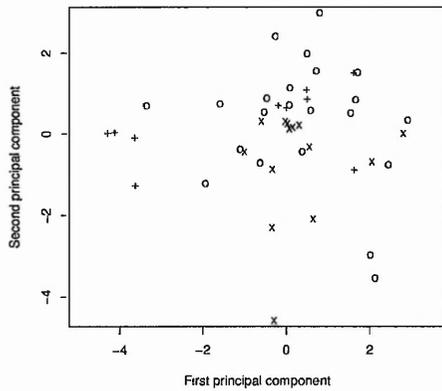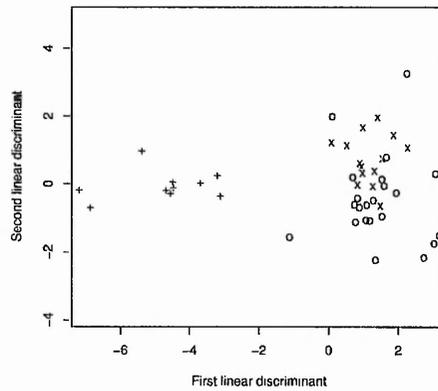
(a) Principal Components Analysis   (b) Discriminant Analysis

Figure 3.4: Plots of the first 2 principal components and linear discriminant functions for the French Medieval glass compositional data of Foy (1985). Point labels identify which of the three sites the glass originates from, detailed in Foy (1985).

Figure 3.4a shows the first two principal components of standardised data, which do not show any separation between specimens from the 3 sites. Figure 3.4b shows the first 2 discriminant functions and suggests one of the groups can be separated from the other two.

3.3.3   Cluster analysis

Cluster analysis is designed such that for a sample of size n with measurements on d variables, similar cases are grouped together. In the current context the intention is to group samples with similar chemical composition. There are a great number of methods which are categorised as cluster analysis, however the general process is two stage.

Initially a coefficient of (dis-)similarity is calculated. Typically d_{ik} is used to denote distance, however to avoid confusion with our notation for dimensionality we will adopt the notation m_{ik}. It is quite common for squared Euclidean distance, m_{ik}^2, or Euclidean distance, m_{ik}, to be used where

$$m_{ik}^2 = \sum_{j=1}^{d}(X_{ij} - X_{kj})^2$$

29

however it is possible to use other measures of distance, such as Manhattan distance

$$m_{ik} = \sum_{j=1}^{d} \mid X_{ij} - X_{kj} \mid .$$

Secondly some form of clustering algorithm is applied to the data, of which there are two general forms, hierarchical clustering and partitioning.

In hierarchical clustering, the process begins with n clusters consisting of one case each and combines cases using a measure of similarity. The resulting clusters are then linked using the same measure of similarity until only a single group remains. Common clustering algorithms include, but are not limited to:

1. Nearest neighbour or single linkage where the distance between two clusters is given by the distance between the two closest neighbours. At any time, the two most similar groups are combined.

2. Complete linkage, also known as furthest neighbour, measures the distance between two clusters as the maximum distance between observations in each of 2 clusters. The two clusters are combined if the distance between them is found to be minimal.

3. Ward's method, which attempts to minimise the sum of squares of any two clusters. Specifically, if $\bar{\mathbf{X}}_k$ is the mean of the k'th variable in a cluster then the variability of the cluster is defined as

$$S = \sum_{i=1}^{d} \sum_{k=1}^{d} (X_{ik} - \bar{\mathbf{X}}_k)^2. \tag{3.18}$$

S is calculated and summed for all clusters to get total variability T. The two clusters are combined that produce the smallest increase in T.

Hierarchical clustering is often represented graphically in a tree-like structure called a dendrogram which displays the hierarchal similarity of objects.

Partitioning is a second method of clustering observations and begins with a known number of groups. It differs from hierarchal methods in that observations can be added to and removed from clusters, thus allowing constant refining of cluster membership. K-means clustering, one method of partitioning, will attempt to determine which of the

30

k defined groups observations should belong to. Given that initially k arbitrary cluster centroids are defined, the technique attempts to minimise the variability in clusters by moving objects into more appropriate clusters. The process continues until group variability cannot be further minimised.

Results from K-means clustering are generally represented in tabular form detailing group membership and cluster centroids, with the performance of the technique is measured by the separation of group centroids. A drawback is that the technique's performance can be affected by the initial choice of group centroids.

Figure 3.5 shows a dendrogram (clustering tree) for the oriental greenware data of Pollard and Hatcher (1986) as discussed in section 3.3.1. The dendrogram illustrated was created using the Euclidean distance measure and the complete linkage method. In order to assign the final group membership, the dendrogram is *cut* at some level of (dis-)similarity, the branches below the cut represent the clusters. In this particular case, Pollard and Hatcher (1986) cut the dendrogram at about 80 to give 2 clusters which relate to those discovered using principal components analysis in section 3.3.1. They then go on to further investigate the 2 clusters in isolation using both cluster analysis and discriminant analysis.

Figure 3.5: Dendrogram showing the cluster membership for the oriental greenwares of Pollard and Hatcher (1986). Cluster analysis was performed using the Euclidean distance measure with observations clustered using the complete linkage method.

Chapter 4

Lead Isotope Analysis

In this chapter, we investigate a specific sample size related problem which arises in Lead Isotope Analysis. We make use of one-, two- and three-dimensional techniques to investigate the problem, which could be applied to any low-dimensional problem.

4.1 Introduction

Since the 1930's, instrumental methods for chemical analysis of metals have been utilised in an attempt to chemically "fingerprint" the origin of metals. Large programmes of analysis of prehistoric metal objects have been undertaken in this time, the intention being to determine the origin of the ore source which in turn will help reconstruct prehistoric economic trade patterns. The use of instrumental methods to analyse metallic objects is fraught with problems. Pollard and Heron (1996:302) note that "the relationship between the trace element composition of a metalliferous ore and that of the metal object derived from it is an extremely complicated one, which is influenced by a number of factors". Aside from the practical difficulties of extracting samples for analysis from precious objects, Pollard and Heron discuss two less obvious problems. Firstly they note that the composition of an ancient mineral deposit may differ from the mineral composition of ore extracted from it. The second complication arises when the metal is extracted from the ore using some kind of furnace technology. Pollard and Heron (1996:304) note that it may have been necessary to crush or wash the ore, which may influence its mineralogical composition. Additional minerals that may have been added to the ore during the furnace process further complicate the mineralogical composition of the metal. Pollard and Heron (1996, 305) conclude

that the measured chemical composition of the metal artefact is a complex function incorporating many factors and go on to suggest that it is no wonder few archaeological chemists have faith in trace element fingerprinting.

Another approach is to utilise isotopic compositions of metals, Pollard and Heron (1996:302) note that most metallic elements exist naturally as different isotopes, that is atoms of the same element that have identical chemical characteristics but vary in atomic weight. For most metals there is little difference between the quantities of each isotope across the surface of the earth, however lead (Pb) is unique in that it has a large range of natural isotopic compositions, due to the fact that three of its four stable isotopes lie at the end of the major radioactive decay chains (Pollard and Heron, 1996, 306-312). The four stable isotopes in lead are ^{206}Pb, ^{207}Pb, ^{208}Pb and ^{204}Pb, the latter being non-radiogenic in origin. Pollard and Heron (1996:306-312) detail the radioactive decay chain and determine that the decay chain of each isotope has a half-life in excess of $0.7 * 10^9$ years. The ratios of the stable isotopes of lead vary measurably from metal deposit to metal deposit and the fact that they appear unaffected by furnace process, provides a major breakthrough in the scientific study of the origin of metal objects.

In geochemistry it is conventional to use the ratios ^{206}Pb/^{204}Pb, ^{207}Pb/^{204}Pb and ^{208}Pb/^{204}Pb since ^{204}Pb is non-radiogenic and these ratios occur in the equations for the isotopic evolution of ore bodies (Pollard and Heron, 1996, 314-315). There is also a practical reason for using these ratios which relates to the method used to measure the isotopic ratios. Pollard and Heron (1996:312) discuss how modern mass spectrometrists use a method called *thermal ionisation mass spectrometry* (TIMS) to take high precision measurements, and one way of achieving the precision necessary is to measure the isotope compositions simultaneously as ratios, this minimises fluctuations in the measuring process.

Brill and Wampler (1967) pioneered the use of lead isotope analysis in archaeology. For reasons that were 'not explained' (Pollard and Heron, 1996, 322), Brill and Wampler used the ratios ^{208}Pb /^{206}Pb, ^{207}Pb /^{206}Pb and ^{206}Pb /^{204}Pb which have subsequently been adopted as the convention within archaeological applications. These ratios are measured directly using TIMS, Pollard and Heron (1996:326) note that TIMS is capable of making the measurements to an absolute 95% error of 0.05% in ^{207}Pb

34
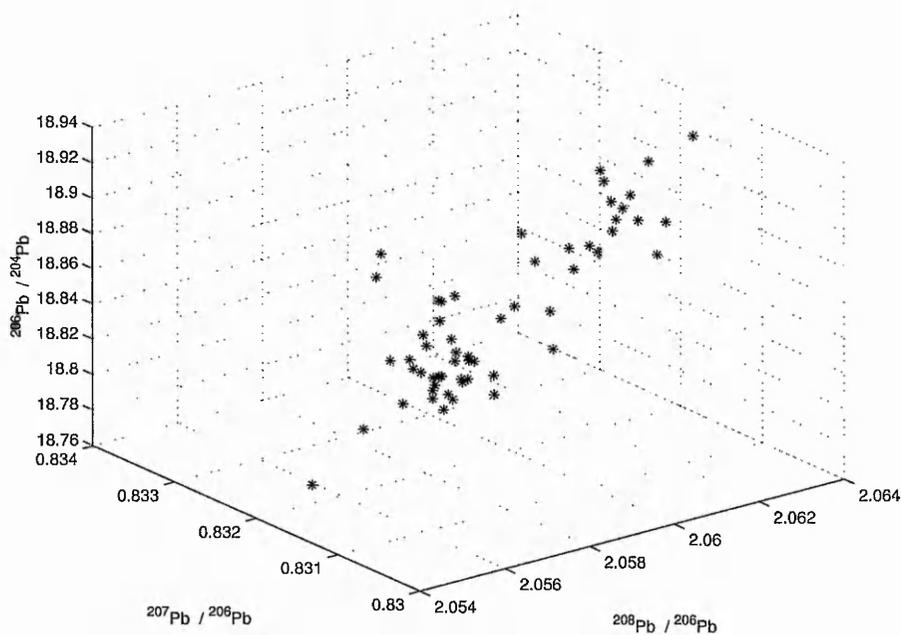
Figure 4.1: 3-dimensional plot (point cloud) of the Lavrion field using data from Stos-Gale *et al.* (1996).

/^{206}Pb and 0.1% for the other 2 ratios.

Reedy and Reedy (1992) make the suggestion that the three lead isotope ratios should be converted to measurements on each of the four isotope concentrations. Pollard and Heron (1996:328) suggest that although this is statistically sensible, it is "unlikely to be helpful in light of the method used to produce the data (i.e. ratio measurements)".

A specimen from an ore body is defined by measurements on all 3 ratios. Given a sample of n specimens, it is possible to estimate the *lead isotope field* for the data which may be represented as a three-dimensional (trivariate) point cloud as illustrated in figure 4.1. The field is commonly represented as bivariate plots of ratio pairs with confidence ellipsoids marking its estimated extent (e.g. Sayre *et al.*, 1992; Gale and Stos-Gale, 1992). The confidence ellipse is determined based on the assumption that the two variables are bivariate normal. Essentially an ellipse is plotted around the data with an orientation determined by the correlation of the variables. For each observation, a probability can be calculated which determines how likely it is lie within the ellipse. Commonly 90% or 95% confidence ellipses are used. Stos-Gale and Gale
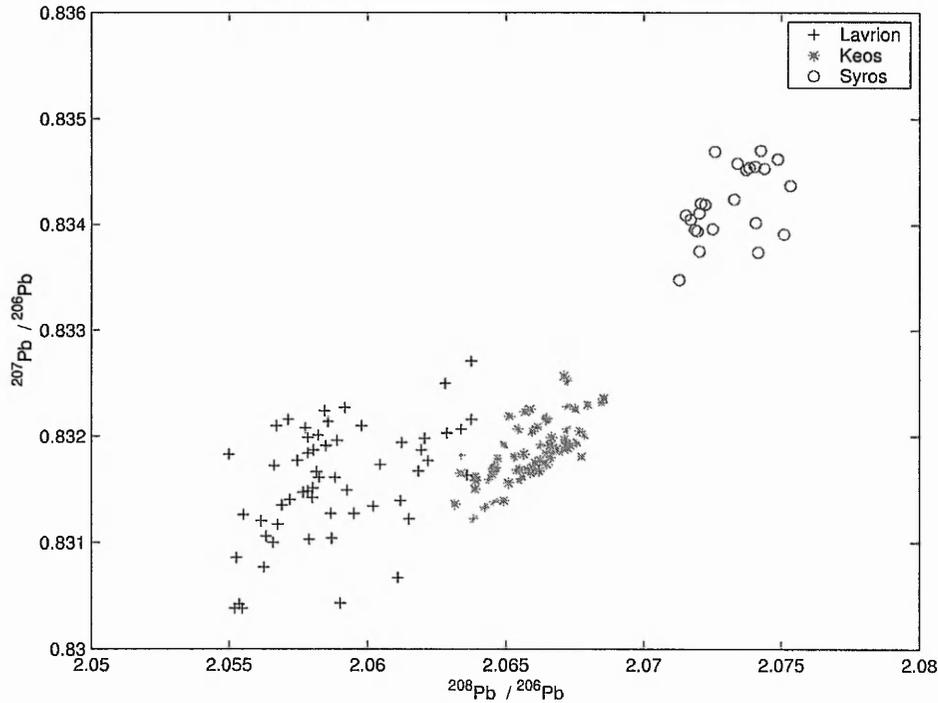
Figure 4.2: Bivariate plot of ^{208}Pb /^{206}Pb against ^{207}Pb /^{206}Pb for the Lavrion, Keos and Syros fields which shows how different ore fields can be distinguished. Data from Stos-Gale *et al.* (1996) and Gale *et al.* (1997).

(1994:112) make use of discriminant analysis to determine the relative probabilities of a sample coming from different fields.

Another approach is adopted by Sayre *et al.* (1992, 1995). They determine the (Mahalanobis) distance of an artefact from an ore-body, and convert this to a probability which is used to assess if an ore body could be the source of the artefact. This probability calculation also requires the assumption of normality.

If fields for the ore-bodies are distinct in two-dimensional space, it follows that they are distinct in three-dimensional space. Given distinct fields, there is the possibility that the lead isotope signature for an artefact may be matched to a field which will allow investigation into its provenance. This is demonstrated in the bivariate plot shown in figure 4.2 where it can be seen that the Syros, Lavrion and Keos fields are separated from each other. Although lead isotope analysis is less problematic than other forms of metallurgic analysis, the amount of data available is still relatively small. As a consequence, the numbers of samples which are used to define the fields is usually

minimal.

In the following sections the question of sample size in lead isotope analysis is addressed. We begin by looking at the univariate ratios in isolation and continue to extend this for bivariate pairs of ratios and finally examine the trivariate data directly. It has previously been the contention that 20 samples are sufficient to define a lead isotope field (Pollard and Heron, 1996). We will utilise some of the large data sets published by Stos-Gale *et al.* (1996) and Gale *et al.* (1997) and apply various techniques to determine minimum sample sizes required to detect structure evident in the full data set. We demonstrate that, in some cases, 20 samples is inadequate and go on to question the assumption of normality.

4.1.1 The question of normality

Recently there has been much debate surrounding lead isotope analysis; Budd *et al.* (1993, 1995, 1996), Stos-Gale *et al.* (1997) and Baxter, Beardah and Westwood (2000) have questioned some of the commonly used methods of provenancing lead isotope data. Budd *et al.* (1993) raise concerns over the method used to construct the confidence ellipse, they suggest that it "creates a misleading impression of the separation of the source fields". The same group have also questioned the assumption of trivariate normality (Scaife *et al.*, 1996) which is an essential assumption in order to construct the confidence ellipse. Our primary interest is in this question of trivariate normality, and in sample sizes which are required to demonstrate this.

Gale and Stos-Gale (1993) and Sayre *et al.* (1992) assert that lead isotope data are (trivariate) normal. Gale and Stos-Gale (1993) use statistical tests of univariate normality of the three ratios (or marginals) considered in isolation. Univariate normality of the three individual ratios alone does not, however, mean that data are trivariate normal. The converse is however true, if there is a departure from univariate normality, there will also be a departure from trivariate normality. Sayre *et al.* (1992) examine the principal component scores for the ratios. They note that they 'often create histograms of the distributions $\cdots$ for source group specimens along the characteristic vectors for the groups. If the specimens of a group are normally distributed along each of its characteristic vectors one can conclude that the group is normally distributed

$\cdots$. These histograms for our groups do approximate Gaussian curves.' This approach is limited by the inefficiency of the histogram as a method of assessing normality, particularly with small samples. In the following examples, it will be demonstrated that normality of the individual univariate marginals in isolation does not imply that the data are trivariate normal.

It is quite common for confidence ellipsoids to be used to define the individual lead isotope fields, it is also common for probability calculations to be performed to determine which field a sample belongs to (Gale and Stos-Gale, 1992; Sayre *et al.*, 1992, 2001). Both techniques require the assumption of normality and thus may be misleading. Baxter, Beardah and Westwood (2000) discuss the importance of the assumption of normality, they point out that the assumption is "rarely, if ever, exactly true but is often a sufficiently good approximation that the methodology is not compromised".

Some statistical methods are "robust" to the assumption of normality, that is the output of the method is insensitive to departures from normality, however Baxter, Beardah and Westwood (2000) also stress that robustness of methods should not be assumed. Later in the paper it is illustrated how the assumption of normality can affect probability calculations, specifically they illustrate how the assumption of normality can result in misleading group memberships. Baxter and Gale (1998) discuss the robustness of discriminant analysis (advocated by Sayre *et al.* 1992) and conclude that it is robust to the assumption of normality.

Results on the univariate case reported on by Baxter, Beardah and Westwood (2000:975) are based on results from the univariate case presented in section 4.2. Also Westwood, Baxter and Beardah (1998) was jointly written and based on results presented in the following sections.

4.1.2 The sample size issue

In addition to specific issues related to the assumption of normality, discussed above, we argue that the lack of sufficient numbers of specimens is cause for concern, as acknowledged by Sayre *et al.* (1992). An obvious consequence of insufficient samples is the inability to detect structure in data, for example departures from normality.

Baxter, Beardah and Westwood (2000) discuss a specific example relating to the

Cyprus field. The field was originally defined by 43 specimens from several sources on the island and was defined using confidence ellipsoids (Stos-Gale et al., 1997). Until extensive new data was made available by Gale *et al.* (1997) and Stos-Gale *et al.* (1997), there was much published debate (e.g. Budd *et al.*, 1995) over the possibility of sub-dividing the field. The availability of more extensive data confirmed that the field could in fact be sub-divided into different deposits. Baxter, Beardah and Westwood (2000:974) state "the important point here is that, because of an inadequate sample size, it could not be recognised that the Cyprus field was non-normal and extremely multimodal. Had larger sample sizes been available at an earlier date much argument could have been avoided".

Although there appears to be little agreement on any area of lead isotope analysis, the general opinion appears to be that a sample size of $n = 20$ is an acceptable minimum (Pollard and Heron, 1996). Sayre *et al.* (1992, 97) state that 'the spread of uncertainty about a source field steadily contracts as the number of specimens describing the source becomes large, tending to level off when one has something in the order of 20 such data points'. Pollard and Heron (1996:328) summarise by noting that 20 geologically well selected ore samples is an 'agreeable minimum'.

Baxter, Beardah and Westwood (2000) suggest that a more "theoretical" justification is provided by Harbottle (1976) for this choice. Harbottle (1976) suggests that for probability calculations that require stable estimation of the covariance matrix with d variables, $n > 5d$ is a desirable minimum. For $d = 3$, a value of n of less than 20 emerges as a desirable minimum. This rule also assumes that data are normally distributed.

Sayre *et al.* (1992) define fields with as few as 10 samples, however more recently Stos-Gale *et al.* (1996) and Gale *et al.* (1997) have made available large data sets, some with over 50 observations.

In the following section, we show that a sample size of 20, or even 40 in some cases, is not sufficient to expose the non-normal structure in certain data. If data are sampled from a normal distribution, 20 may well be an acceptable minimum, however if specimens are sampled from a non-normal distribution, 20 may be seriously inadequate.

4.2 The one-dimensional case

4.2.1 Testing for departures from normality

Bowman (1992) proposed a test of univariate normality based on the use of kernel density estimates. Given that $X_1, X_2, \ldots, X_n$ is a univariate sample with n observations, the test simply measures the "distance" between a KDE for the data, and the expected density under the assumption of normality. The resultant integrated squared error (ISE) statistic is given by

$$\int \{N(x, 1 + h^2) - \hat{f}(x; h)\}^2 dx. \tag{4.1}$$

where $\hat{f}(x; h)$ is a kernel density estimate of x with window-width h (equation 3.2). $N(x, 1 + h^2)$ is the normal density with mean 0 and variance $1 + h^2$. Bowman (1992) notes that the comparison of the KDE is made against $1 + h^2$ as this is the "expected value of $\hat{f}(x; h)$ under the null hypothesis" (of normality). The univariate normal density function (pdf) is defined as

$$N(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}} |\sigma|^{-\frac{1}{2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma}\right). \tag{4.2}$$

Data are initially standardised to have zero mean and unit variance, this allows direct comparison against a normal density with zero mean and unit variance.

The statistic can be evaluated numerically or analytically using results given in Bowman (1992), who also provides critical values using

$$h = \left\{\frac{4}{3n}\right\}^{1/5} \tag{4.3}$$

which is in fact the normal scale rule (equation 3.9) with $\sigma = 1$, which is AMISE-optimal under the null hypothesis of normality.

Assessing performance of the 1D ISE statistic

Bowman (1992) produced power curves to demonstrate how the ISE statistic performed for a specific mixture of normal distributions of the form

$$0.5N(-\mu, 1) + 0.5N(\mu, 1)$$

for $-3 \leq \mu \leq 3$ and sample sizes of $n = 20$ and 50. These distributions are formed by sampling $n/2$ observations from $N(-\mu, 1)$ and $n/2$ from $N(\mu, 1)$. The resulting

40
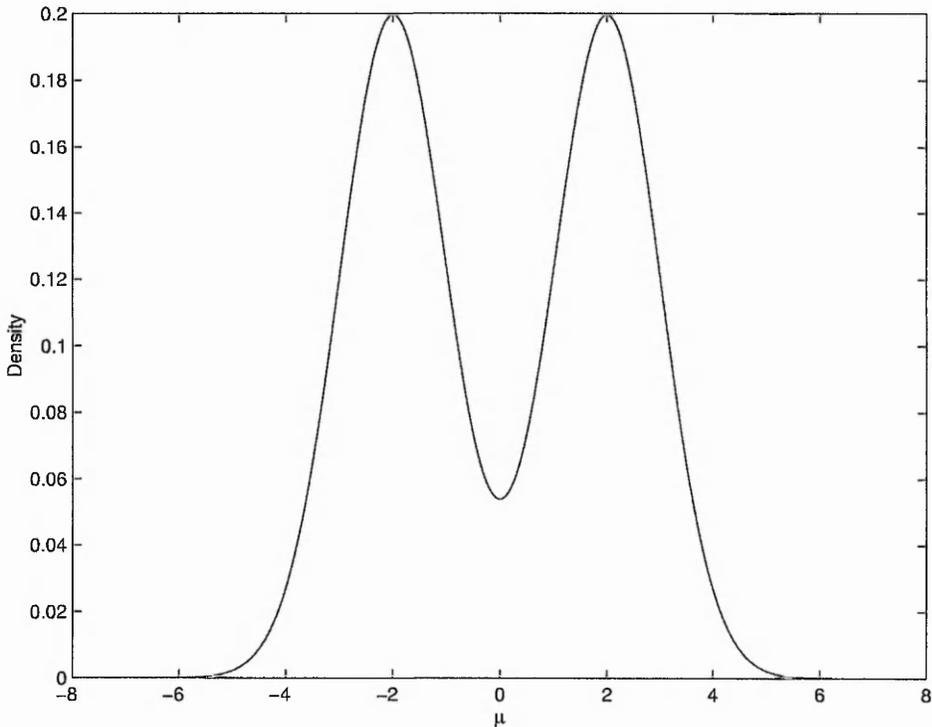
Figure 4.3: Bimodal distribution $0.5N(-\mu, 1) + 0.5N(\mu, 1)$ with $\mu = 2$, one of the distributions used by Bowman (1992) as part of a study of the power of the univariate ISE statistic.

distributions are bimodal with 2 identical normal pdf's separated by 2μ. Figure 4.3 shows the mixture $0.5N(-\mu, 1) + 0.5N(\mu, 1)$ with $\mu = 2$ (that is a separation of 4).

This original, limited, simulation study has been extended to determine how the ISE statistic performs for a wider range of univariate distributions.

Bowman's (1992) mixture distribution can be generalised to generate a wider variety of bimodal distributions

$$(1 - p)N(0, 1) + pN(\mu, \sigma^2). \tag{4.4}$$

$N(\mu, \sigma^2)$ is a general univariate normal pdf with mean μ and variance σ^2. The value of p, the mixing parameter, determines the proportion of samples taken from each of the components, $0 \leq p \leq 1$. Given that n is the number of samples in the simulated distribution, $n(1 - p)$ samples are taken from $N(0, 1)$ and np from $N(\mu, \sigma^2)$. The resulting distributions are typically bimodal and have a mode at 0. Depending on the value of σ^2 a second mode is present at μ. The conditions for bimodality, which depends on μ and σ^2, are defined in equation (4.6).

41

This model can be further generalised to simulate multi-modal distributions:

$$\sum_{i=1}^{c} p_i N(\mu_i, \sigma_i^2) \tag{4.5}$$

where $\sum_{i=1}^{c} p_i = 1$ and c is the number of components forming the distribution.

Using equation (4.4), the number of possible distributions which can be simulated is endless. The distributions range from those which appear unimodal, for example taking $\mu = 0$, to the obviously bimodal distributions which can be simulated with $\mu \geq 10$. In order to limit the number of possible distributions to a manageable number, the simulation study is restricted to distributions with two components, as defined in equation (4.4). Also, μ is constrained to lie between 0 and 6. This gives a range of distributions from those which cannot be differentiated from unimodal (e.g. $\sigma^2 = 1$ and $\mu = 1$) to a clearly bimodal distribution with $\mu = 6$. It would be expected that further separation of the components would make little difference to the performance of the ISE statistic. Sample sizes of $20 \leq n \leq 100$ were used.

The extended simulation study which we undertake consisted of generating 1000 samples of size n from each of 24 mixture distributions of the type defined in equation (4.4). Results are presented in table B.1 in Appendix B and presented as cumulative frequency curves where appropriate. The curves show the separation of components plotted (μ) against the proportion of repetitions for which the hypothesis of normality is rejected (power).

Initially, the effect of varying sample size was examined and values of $p = 0.5$ and $\sigma^2 = 1$ were used. Everitt and Hand (1981) state that univariate mixture distributions of the form

$$(1-p)N(\mu_1, \sigma_1^2) + pN(\mu_2, \sigma_2^2)$$

have one unique mode if

$$|\mu_1 - \mu_2| \leq 2 \min(\sigma_1^2, \sigma_2^2). \tag{4.6}$$

For the mixture distributions used here, $\mu_1 = 0$, $\mu_2 = \mu$, $\sigma_1^2 = 1$ and $\sigma_2^2 = \sigma^2$. In the present example (with $\sigma^2 = 1$), $\mu \geq 2$ will result in a bimodal distribution. The power curve in figure 4.4 shows that for $\mu \geq 2$, the power of the statistic increases as n does. Smaller sample sizes require that components be more separated for the hypothesis of normality to be rejected more frequently.

42
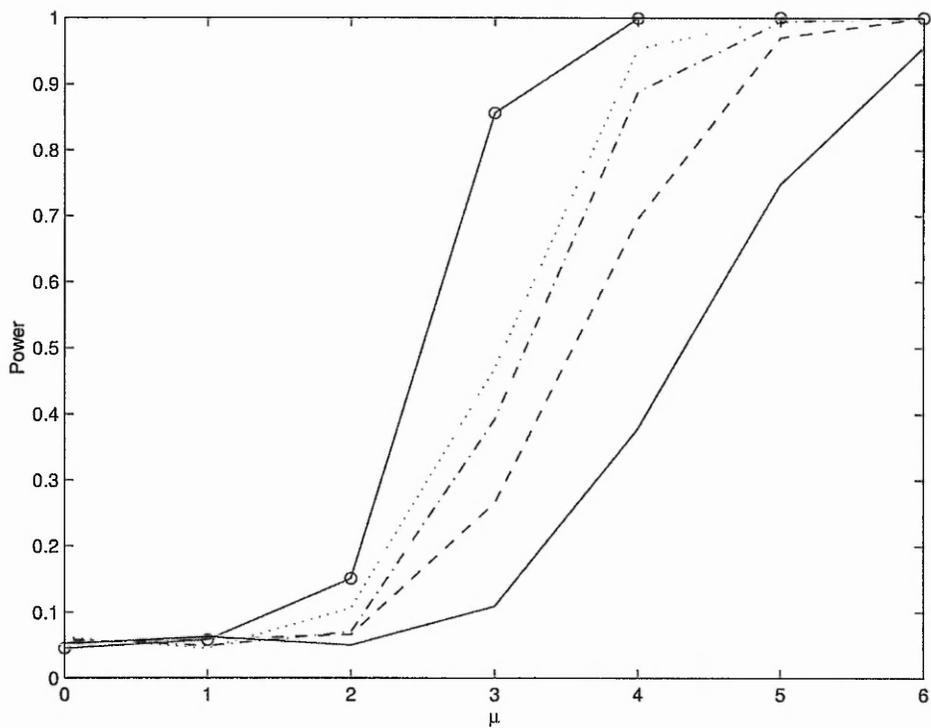
Figure 4.4: Cumulative frequency curves showing the power of the ISE statistic for mixtures of the form $0.5N(0, 1) + 0.5N(\mu, 1)$ for $1 \le \mu \le 6$. The solid line represents power for $n = 20$, dashed for $n = 30$, dash-dot for $n = 40$, dotted for $n = 50$ and solid with 'o's for $n = 100$.

(a) $n = 25$ (b) $n = 100$
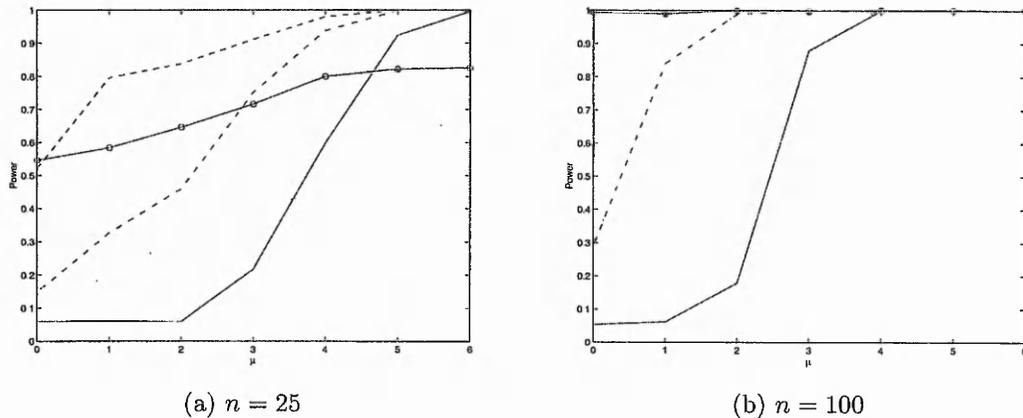
Figure 4.5: Cumulative frequency curves showing the power of the ISE statistic for mixtures of the form $0.5N(0,1) + 0.5N(\mu, \sigma^2)$ for $1 \leq \mu \leq 6$. The solid line represents power for $\sigma^2 = 1$ (bimodal for $\mu \geq 2$), dashed for $\sigma^2 = 0.25$ (bimodal for $\mu \geq 2$), dash-dot for $\sigma^2 = 0.5$ (bimodal for $\mu \geq 2$), dotted for $\sigma^2 = 2$ (bimodal for $\mu \geq 4$) and solid with 'o's for $\sigma^2 = 4$ (bimodal for $\mu \geq 8$).

In general, varying p has little effect on the power of the statistic other than for $p = 0$ and 1 which results in a unimodal population. For smaller sample sizes, varying p does result in fractionally lower power, however for larger n virtually no difference is observed.

Simulating distributions with long tails is achieved by decreasing σ^2, this results in a much narrower mode. Increasing σ^2 results in the second distribution becoming kurtotic, that is a lower flatter mode. For smaller sample sizes (e.g. 25), the power is low compared to larger n (e.g. 100) as illustrated in figure 4.5. The implication of this is that sample sizes of ≈ 20, as commonly used in lead isotope analysis, are not sufficient to detect certain forms of structure directly using the ISE statistic, larger sample sizes are needed in some cases. In particular kurtotic distributions and those with very little separation between components require considerably more samples to be able to reproduce the structure present in the population.

4.2.2 Counting modes

A more direct approach than testing for normality is simply to count the number of modes in a KDE, $\hat{f}(x)$. For example, a $N(\mu, \sigma^2)$ distribution is unimodal as it contains a single component. If we count the number of modes in a KDE based upon a sample

from a distribution and find that more than 1 are present, we can assume a departure from normality.

The difficulty is in assessing whether or not what seems visually apparent reflects real population structure. Assessing the modality of a KDE can be achieved in a number of ways, for example one obvious approach is to locate the turning points of $\hat{f}(x)$ by finding solutions to $\hat{f}'(x) = 0$. We have used a more numerical approach, presented below, which we found to be far less computationally expensive.

Given a vector of KDE heights evaluated at regular intervals; let $\hat{f}_k = \hat{f}(s_k)$ where s_k defines the mesh of values at which the KDE height is calculated. Then a mode is detected when $\hat{f}_k > \hat{f}_{k-1}$ and $\hat{f}_k > \hat{f}_{k+1}$.

Experimentation has suggested that both forms of mode counting proposed above are extremely sensitive to outliers. Whereas visual examination of KDEs allows us to ignore what is deemed to be spurious structure, for example outliers, replicating this procedure mathematically is a more complex problem.

We have attempted to explore alternate approaches to mode counting, specifically we have investigated the possibility of using neural networks. A neural network can be thought of as a simplified model of the human brain which can be "trained" to detect patterns within data, the intention being to mimic the subjective ability to categorise patterns. Very simplistically a number of inputs are used to generate a number of outputs, and the network is "trained" by example. Our experimentation consisted of training a neural network to determine if a KDE has 1 or 2 modes. We input the height of a KDE at regular intervals and the network outputs a flag to indicate if the KDE is unimodal or bimodal. Our results with neural networks were unsatisfactory at times; although the network was able to successfully classify a proportion of samples, it had a tendency to incorrectly classify clearly bimodal data as unimodal and clearly unimodal data as bimodal.

In order to address the problem of outliers, most of which commonly appear in the tails of the estimated distribution, we have opted to ignore modes found below a pre-determined height threshold (unless otherwise stated any mode whose height is below 10% of the maximum height of the KDE is ignored). Figure 4.6a shows an example of a mode which falls below the threshold which is ignored, with figure 4.6b showing a

(a) Mode ignored (below 10% threshold)

(b) Mode not ignored (above 10% threshold)
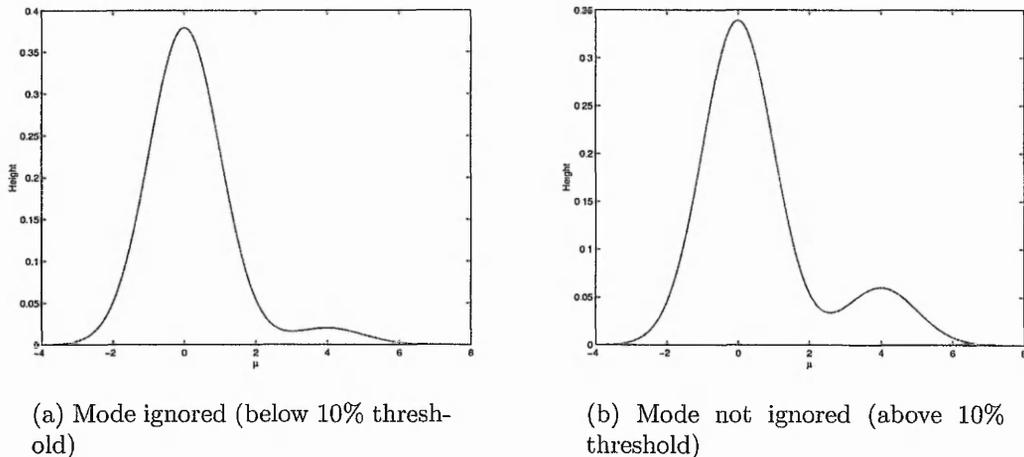
Figure 4.6: Example illustrating how potential outliers are removed from bump hunting. A mode is ignored if it is lower than 10% of the maximum height of the KDE.

mode which is not ignored.

Assessing performance of 1D mode counting

In order to assess the performance of the mode counting procedure (using the numerical approach) discussed above, we use the same distributions used to assess the performance of the ISE statistic in section 4.2.1. Specifically we have constructed a family of distributions of the form

$$(1 - p)N(0, 1) + pN(\mu, \sigma^2) \tag{4.7}$$

which range from those which cannot be distinguished from unimodal to clearly bimodal distributions. The samples generated for the ISE statistic were saved and have been re-used for mode counting, meaning 1000 repetitions are used for each distribution and sample size combination. Results are presented in tabular form in appendix B.

The mode counting results serve three purposes, initially to determine performance of the mode counting procedure; secondly to allow comparison between non-adaptive (section 3.2.1) and adaptive KDEs (section 3.2.3) and finally to compare performance of different algorithms for the selection of the smoothing parameter, h (section 3.2.2). Results from each of the four methods of generating KDEs (non-adaptive and adaptive with h selected via h_{NS} and h_{STE}) are presented in individual tables B.2 to B.5 in appendix B

46

Silverman (1986; 102) observes that adaptive KDEs are insensitive to the pilot value of h used. Our experience tends to disagree with this observation and we argue that the pilot value of h forms an important choice in the construction of an adaptive KDE. For example, figure 4.7a and b show 2 adaptive kernel density estimates of the same sample of size $n = 100$ from the mixture of distributions

$$0.4N(0,1) + 0.6N(2/3, 5/2).$$

Figure 4.7a has a pilot h value computed using the Normal Scale Rule and figure 4.7b has a pilot h value computed using the Sheather and Jones (1991) technique. The pilot estimate of h_{STE} emphasises the mode to the left of the plot whereas the pilot estimate of h_{NS} makes the mode appear far less prominent.

Generally the Sheather and Jones (1991) Solve-the-Equation (h_{STE}) algorithm tends to produce smaller h values than the normal scale method (h_{NS}). Non-adaptive KDEs with h_{STE} tend to be more 'jagged' due to the relative under-smoothing from the small value of h (figure 4.7d), whereas non-adaptive KDEs with h_{NS} (figure 4.7c) tend to be smoother in appearance. The window-widths generated by h_{NS} tend to be larger than those from h_{STE}. With the additional smoothing applied by the adaptive KDE, the pilot estimate of h_{NS} tends to over smooth compared to the non-adaptive KDE, whereas adaptive KDEs using a pilot estimate of h_{STE} appear smoother than the non-adaptive KDE with the same value window-width.

The general trend of our results is that, as sample size or component separation increases, the proportion of times which the correct number of modes are detected increases as would be expected. In general, for smaller sample sizes, the adaptive h_{STE} method tends to perform better than other methods, however for larger n there is less difference between adaptive and non-adaptive KDEs or h selection algorithms.

The effect of varying sample size is noticeable, particularly when the normal scale h selection method is used. In general, an adaptive KDE with a pilot estimate generated by h_{STE} produces better results, with over 50% of cases being detected as bimodal when $\mu = 3$, irrespective of sample size. Figure 4.8 shows cumulative frequency curves for adaptive h_{STE} and non-adaptive h_{NS} which show the proportion of replications for which bimodality is detected against the separation of components (μ). Samples are from the distribution $0.5N(0,1) + 0.5N(0,\mu)$ for $0 \leq \mu \leq 6$.
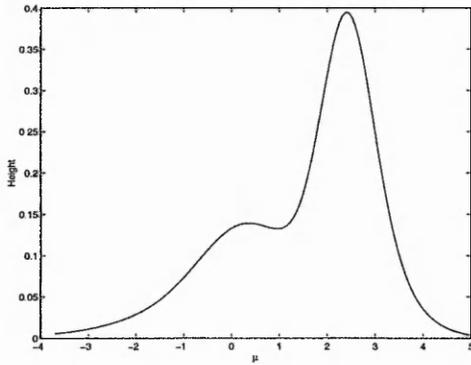
(a) Adaptive KDE with pilot estimate of $h_{NS} = 0.6104$
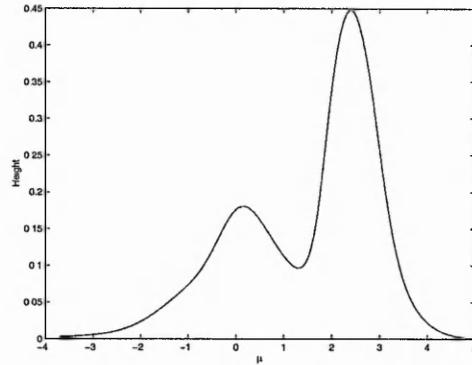
(b) Adaptive KDE with pilot estimate of $h_{STE} = 0.3443$
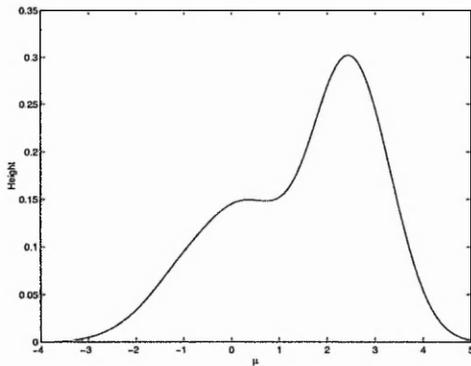
(c) Non-adaptive KDE with $h_{NS} = 0.6104$
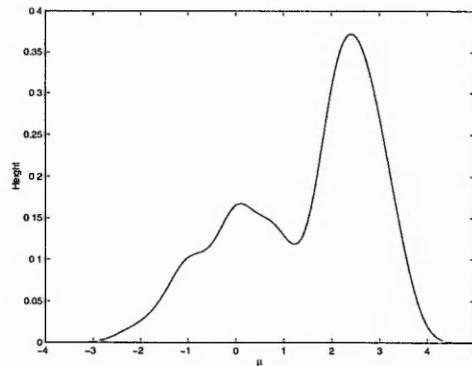
(d) Non-adaptive KDE with $h_{STE} = 0.3443$

Figure 4.7: KDEs generated for 100 observations from the mixture $0.4N(0,1) + 0.6N(2/3, 5/2)$.

(a) Adaptive, h_{STE} (b) Non-adaptive, h_{NS}
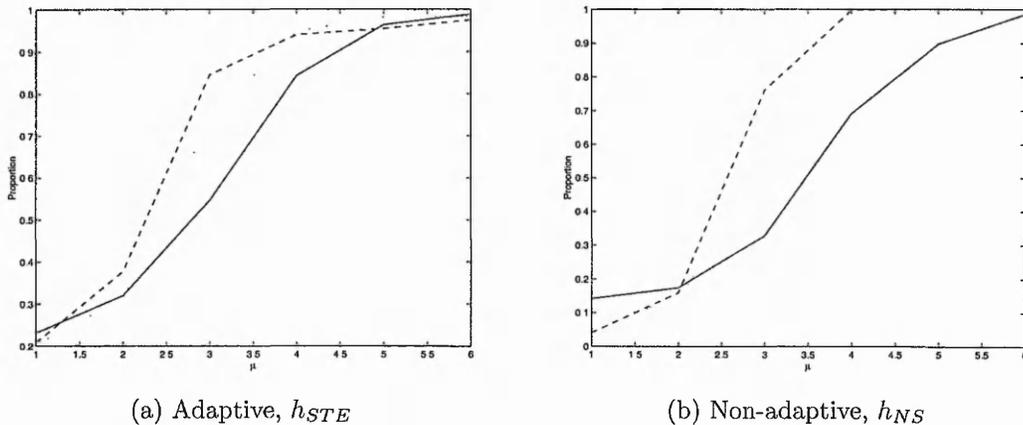
Figure 4.8: Performance of mode counting for adaptive KDEs with h selected via h_{STE} and non-adaptive KDEs with h via h_{NS}. Curves show proportion of replications for which bimodality is detected against separation of components. Solid line represents $n = 20$, dotted represents $n = 50$ and dashed represents $n = 100$. Samples are from the distribution $0.5N(0,1) + 0.5N(0,\mu)$ for $0 \leq \mu \leq 6$.

Varying p has far more effect than it had on the ISE test of normality. For smaller sample sizes, $\mu \geq 3.5$ tends to result in approximately 50% of samples being detected as bimodal when $p = 0.3$. For sample sizes of 50, this increases to over 60% (when h_{STE} is used). There is generally little to choose between adaptive and non-adaptive KDE generation, however normal scale estimates of h tend to outperform h_{STE} when $p = 0.1$.

4.2.3   Other approaches to mode counting

The approach adopted for assessing the modality of data discussed above is primarily numerical, the intention of the approach was to quickly assess how many modes were present in the data. It was hoped that this would mimic the visual interpretation which we make, however this was not as successful as would have been hoped. We suspect that this is in part because of spurious structure, which is often visible in the tails of KDEs as small modes. Our attempts at ignoring these small modes, i.e. those which are less than 10% of the height of the highest point of the KDE, has not always been successful with some spurious structure not being ignored. More formal methods of mode counting have been proposed.

One of the most commonly used formal tests of modality is that of Silverman (1981)

who proposes the use of KDEs. Modality is sequentially tested for

H_0 : The number of modes is k

H_1 : The number of modes is greater than k

starting with $k = 1$ and increasing until we are satisfied that H_0 is accepted. For each stage, a KDE is formed with k modes, then repeated bootstrap samples are taken from the data and again used to generate KDEs with the same window-width. The significance level of the test is given by the number of times bootstrap samples lead to k-modal KDEs. This test is discussed more fully in chapter 5, along with suggestions for possible improvements. Efron and Tibshirani (1997) extend the original work of Silverman by incorporating Bayesian approaches.

Another approach is that of 'excess mass'. In this approach, modes are ignored if they are deemed to be insignificant (as measured by the proportion of 'mass' of the estimated density which they occupy). Discussion of this approach can be found in Muller and Sawitzki (1991). Other tests include the DIP test of unimodality (Hartigan and Hartigan, 1985), the RUNT test for multimodality (Hartigan and Mohanty, 1992) and the MAP test (Rozal and Hartigan, 1994). Although theory for these tests receives much discussion in statistical literature, few applications appear to have been published. A possible extension of the work presented here would be a comparative study of these tests.

4.3   One-dimensional applications

In this section we report on the application of the ISE statistic (discussed in section 4.2.1) and the mode counting procedure (discussed in section 4.2.2) to the 1D lead-isotope problem.

Several of the larger lead-isotope data sets in Stos-Gale *et al.* (1996) and Gale *et al.* (1997) that exhibit non-normality were used to assess the sample size required to detect the non-normality of the original data set. Sub-samples were taken (without replacement) from the Larnaca data ($n = 63$) (Gale *et al.*, 1997). Figure 4.9 shows an adaptive KDE of the third ratio and it clearly appears to be multi-modal and non-normal.

Repeated sampling (1000 times) for $n = 20$ from the third ratio of the Larnaca data
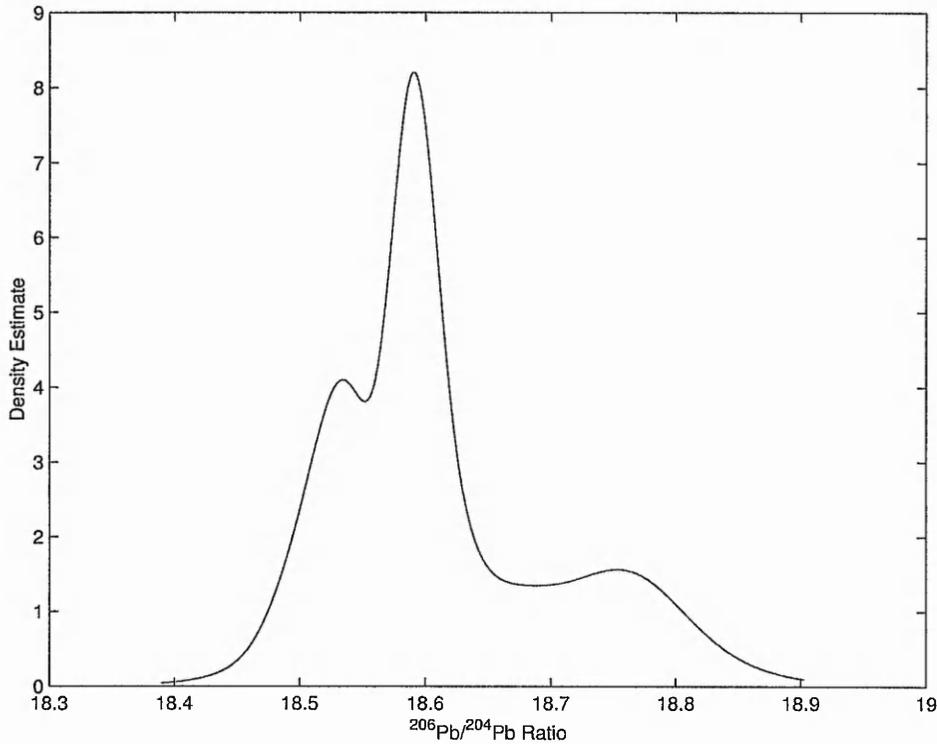
Figure 4.9: Adaptive KDE of the ^{206}Pb/^{204}Pb ratio of the Larnaca axis, $n = 63$ (Gale *et al.*, 1997). h selected via solve-the-equation.

resulted in the rejection of normality about 50% of the time using the ISE test; this rose to about 90% for $n = 35$. The value of the ISE statistic for all $n = 63$ observations is 0.0215 which is significant at the 1% level.

Figure 4.10 shows an adaptive KDE of the third ratio of the Lavrion data ($n = 59$) (Gale *et al.*, 1996). For $n = 20$, normality was rejected just over 15% of the time, for such clearly bimodal data this is quite low; for $n = 50$ normality was rejected about 80% of the time. A "mode counting" analysis on this data, using adaptive KDEs with h_{STE} producing the pilot estimate, suggests that a sample size of 20 gives rise to bimodal KDEs about 40% of the time, which is about double the success rate for the normal testing approach. The ISE statistic for all 59 observations is 0.0116 which is also significant at the 1% level.

Although examining the margins of the three-dimensional data allows non-normality to be detected in these instances, this will not generally be the case. Although direct testing of trivariate normality is possible, an alternative approach can be to find a univariate 'view' of the data which is non-normal. If such a 'view' of the data can be

51
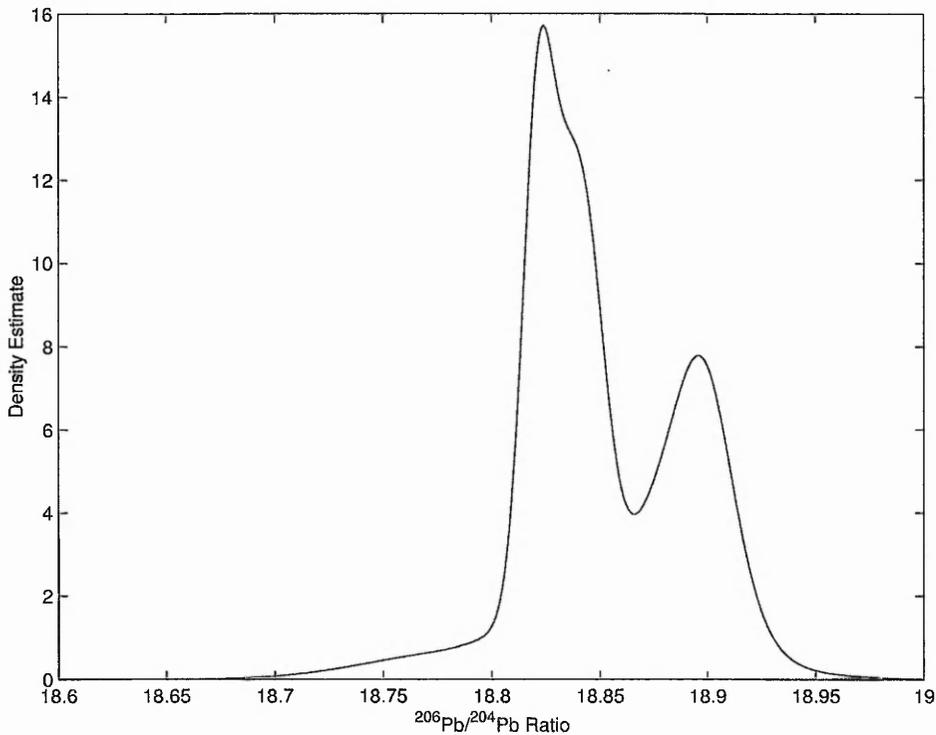
Figure 4.10: Adaptive KDE of the ^{206}Pb/^{204}Pb ratio of the Lavrion ore-field, $n = 59$ (Stos-Gale *et al.*, 1996). h selected via solve-the-equation.

found it follows that the hypothesis of trivariate normality cannot be accepted.

Figure 4.11 shows a KDE of a 'projection' of 2 of the 3 variables (0.1874^{208}Pb/^{206}Pb -0.9823^{207}Pb/^{206}Pb) from the Keos ore field (Stos-Gale *et al.*, 1996) which consists of 62 observations. This linear combination of variables was suggested by Malkovich and Afifi's (1973) multivariate modification of the Shapiro-Wilk Statistic,

$$W = (\Sigma \alpha_i X_i)^2 / \Sigma (X_i - \bar{x})^2 \qquad (4.8)$$

where $X_1, X_2, \ldots, X_n$ are ordered samples of n observations, and the coefficients, α_i, depend on the covariance matrix of the order statistics of a sample of standard normal random variables (Baxter, 1999). Malkovich and Afifi (1973) extend this to the multivariate case, specifically they calculate the minimum value of W over all possible linear combinations of variables, which is a non-trivial calculation and solved by 'brute force' methods. This method, which scans high dimensional data to produce a low-dimensional view which is 'interesting', is one form of projection pursuit methodology, a topic which is discussed more fully in chapter 6.

Sub-samples were taken (without replacement) from the projected data and tested

52
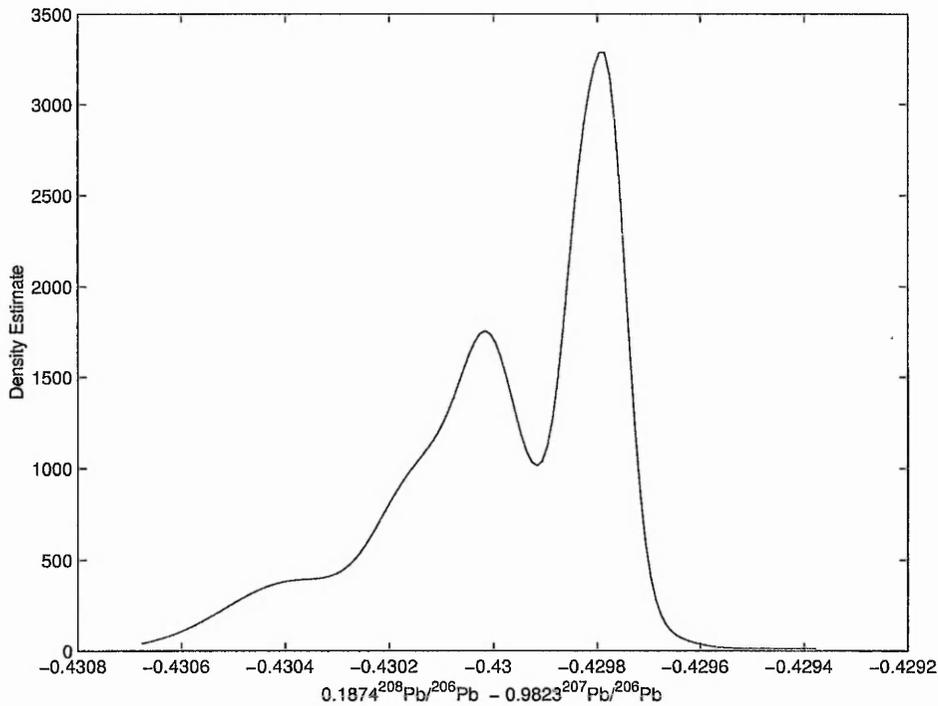
Figure 4.11: Adaptive KDE for a linear combination of two lead isotope ratios from the Keos ore field, $n = 62$ (Stos-Gale *et al.*, 1996), 0.1874^{208}Pb$/^{206}$Pb-0.9823^{207}Pb$/^{206}$Pb. h selected via the h_{STE} procedure.

for departures from normality using the ISE statistic. A sample size of $n = 20$ results in the hypothesis of normality being rejected 28% of the time; for a sample size of 40 this figure rises to 70%.

It is only very recently that sufficiently large data sets have been made available to investigate normality in this way. The analyses just cited, and the studies on simulated data in previous sections, suggest that 20 can be a seriously inadequate number of samples with which to define a lead isotope field. In some of the examples double the number of samples are necessary to detect the departures from normality. Since the assumption of normality is central in some approaches to statistically handling lead-isotope data (Sayre *et al.*, 1992; Sayre *et al.*, 2001), this is an important finding. In some cases, sample sizes at least double those often recommended may be needed to detect problems with the assumptions used in statistical analyses.

4.4 The two-dimensional case

4.4.1 Testing for departures from normality

Bowman and Foster (1993) extend the original work of Bowman (1992) on the ISE statistic and propose a multivariate form. Essentially the statistic remains unchanged,

$$\int \{N_d(\mathbf{x}, (1 + h^2)\mathbf{I}_d) - \hat{f}(\mathbf{x}, h^2\mathbf{I}_d)\}^2 d\mathbf{x} \tag{4.9}$$

where $N_d(\mathbf{x}, (1 + h^2)\mathbf{I}_d)$ denotes the normal pdf (equation 3.1) with covariance matrix $(1 + h^2)\mathbf{I}_d$ evaluated at $\mathbf{x}$ and $\hat{f}(\mathbf{x}, h^2\mathbf{I}_d)$ is the multivariate kernel density estimate as defined in equation 3.4, created with the normal kernel. As in the univariate case, data are initially standardised.

Critical values are again provided for the statistic for a range of d and n at the 5% level using the AMISE-optimal value of

$$h = \left\{ \frac{4}{(d+2)n} \right\}^{1/(d+4)}. \tag{4.10}$$

Assessing performance of the 2D ISE statistic

Bivariate mixture distributions can be created in much the same way as univariate distributions. The multivariate normal distribution (as defined in equation 3.1) with $d = 2$ can be written as $N(\boldsymbol{\mu}, \Sigma^2)$ where $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ is the mean of the distribution and $\Sigma^2 = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{bmatrix}$ is the covariance matrix. A bivariate mixture distribution can then be written as

$$p_1 N(\boldsymbol{\mu}_1, \Sigma_1^2) + p_2 N(\boldsymbol{\mu}_2, \Sigma_2^2) + \cdots + p_c N(\boldsymbol{\mu}_c, \Sigma_c^2)$$

where $\sum_{i=1}^{c} p_i = 1$ are the mixing proportions.

In the bivariate case, the number of distributions which could be created is enormous. Wand and Jones (1993) work with a set of 12 bivariate mixtures which demonstrate various forms of structure (detailed in appendix A). For the present purposes, these 12 distributions have been used. As with the univariate distributions, 1000 subsamples of varying size have been taken and we test using the ISE statistic H_0 : The distribution of the population from which the sample is selected is normal against H_1 : The distribution of the population from which the sample is selected is not normal.

54

Sample Size	15	20	25	50	100
(A) Uncorrelated Normal	0.072	0.053	0.055	0.037	0.041
(B) Correlated Normal	0.050	0.059	0.066	0.040	0.036
(C) Skewed	0.222	0.244	0.339	0.504	0.816
(D) Kurtotic	0.176	0.235	0.265	0.436	0.846
(E) Bimodal I	0.053	0.072	0.097	0.159	0.519
(F) Bimodal II	0.475	0.931	1.0	1.0	1.0
(G) Bimodal III	0.302	0.572	0.842	1.0	1.0
(H) Bimodal IV	0.237	0.429	0.657	0.991	1.0
(I) Trimodal I	0.192	0.321	0.486	0.855	0.999
(J) Trimodal II	0.105	0.170	0.230	0.530	0.967
(K) Trimodal III	0.263	0.449	0.643	0.923	1.0
(L) Quadrimodal	0.104	0.204	0.371	0.738	0.991

Table 4.1: Proportion of 1000 replications for which the ISE statistic rejects the hypothesis of normality (at the 5% level) for mixtures detailed by Wand and Jones (1993). Details of the 12 mixtures, A to L, can be found in appendix A.

Table 4.1 details the results of the simulation study. Distributions A and B, which are not mixtures, show power of approximately 5% irrespective of sample size, which is to be expected. For the other distributions, as n increases, so does the power, although for the remaining 2 univariate distributions, C and D, power is low for $n \leq 50$.

Mixtures F, G, H and K, which show the greatest separation between components also have the highest powers for sample sizes of over 25. For larger sample sizes, detection of non-normality is almost certain.

In mixtures I and L the separation of two components is at least partially obscured by a third or fourth component. The power is moderately good for sample sizes of 50 (though, for L, non-normality would not be detected about a quarter of the time).

For mixture J, with two distinct modes but considerable overlap between components, a sample size of 100 is needed for good power, and power is poor for the smaller sample sizes. Mixture E is qualitatively similar to mixture J, but performance is even poorer, with a sample of size 100 not really adequate for detecting normality.

In general, it can be concluded that for mixtures which consist of components which are clearly separated by a region of low density, a sample size of 50 will often be sufficient to reject the hypothesis of normality a high proportion of the time. Smaller sample sizes, sometimes as low as 25, may be good enough to reject the hypothesis of normality a moderate proportion of the time.

Where there is a clear overlap between components or in cases where two distinct components are linked by a third, larger sample sizes are needed to obtain high power.

4.4.2 Counting modes

The univariate mode counting procedure discussed in section 4.2.2 can be directly extended to the bivariate case. In the bivariate case, each point defining the height of a KDE is surrounded by 8 neighbours in a 3×3 grid. A point can thus be defined as a mode if none of its 8 neighbours have a greater height. Again, spurious structure, commonly found in the tails, is ignored by means of a predetermined threshold value.

As well as simply counting the number of modes present, it is also possible to determine if modes are in approximately the correct location. For an m-mode mixture a KDE based on a simulated sample is considered to be 'successful' if the m modes are detected in approximately the right positions (within a radius of $\boldsymbol{\mu}_c/2$ where $\boldsymbol{\mu}_c$ is the known mean of the component) regardless of the total number of modes detected. This gives us an opportunity to determine how accurately a KDE is able to reproduce the characteristics a known population from a limited sample.

Assessing performance of the 2D mode counting procedure

The 12 bivariate mixtures used above to assess the performance of the ISE statistic are used as the basis for a study into the performance of the mode counting procedure detailed above. The 12 mixtures detailed in appendix A were repeatedly sampled from 1000 times for varying sample sizes. Results are presented in tabular form in tables B.6 to B.11 in appendix B. Each table shows results for one of the six methods of generating the bivariate KDEs examined, i.e. non-adaptive and adaptive KDEs with the window-width calculated by h_{NS}, h_{STE} and h_{DPI2}.

In general, the probability of detecting the correct number of modes was relatively low. There was a tendency for performance not to increase with sample size; we suspect that spurious structure in regions of low density is partially to blame for this. Our attempt to automatically remove modes which are not important by ignoring any which are below a defined threshold was not as successful as in the univariate case. In some instances, the 2 stage direct plug-in window-width selection routine out-performed the univariate routines, however this was not evident in all cases.

Modes were generally located in approximately the correct location, irrespective of which window-width selection routine was used. For the three bimodal mixtures with clearly separated components and modes (F, G, H) $n = 100$ gives a success rate of about 85-90%. With $n = 50$ the success rate is between about 65-80%, with the lowest value associated with the obviously separated mixture F.

For the overlapping bimodal, and tri- and quadrimodal distributions the success is somewhat lower being, at best, about 60% for $n = 100$ and very much lower in some cases, particularly for the quadrimodal mixture.

The bivariate mode counting procedure's performance has not been satisfactory and highlights that the process discussed above needs further development to be of use. The results presented in tables B.6 to B.11 in appendix B suggest that the correct number of modes are rarely detected, with spurious modes causing considerable problems for our automatic mode counting. As such, some form of more intelligent outlier removal, or a more sophisticated mode counting process is required.

Currently a point (i.e. a location at which the height of a KDE is evaluated) is a mode if it has no higher neighbour and is above the threshold. This is perhaps an overly simplistic definition of a mode as it does not take into account anything other than directly neighbouring points. Complications arise because spurious structure does not always fall below the threshold. If the threshold is too high, there is a possibility that valid structure will be ignored, however if the threshold value is too low there is a possibility of spurious structure appearing in the tails of the KDE, sometimes appearing as several small modes which can generally be assumed to be unimportant. Figure 4.12 shows two examples of what might be considered spurious structure. These are presented as one-dimensional KDEs although similar can occur in any number of dimensions. Figure 4.12a shows a 'bump' at about $x = 18$; the mode counting procedure would count this as a valid mode, however it might be classed as unimportant when examined by a human. Figure 4.12b shows 2 modes in the right hand tail of the KDE which are close together, both would likely be ignored as spurious.

A slightly more sophisticated approach to mode counting than that used above would be to look at areas of the KDE, instead of just looking at a single point and its neighbours. The hope is by looking at a slightly larger area, small spurious modes
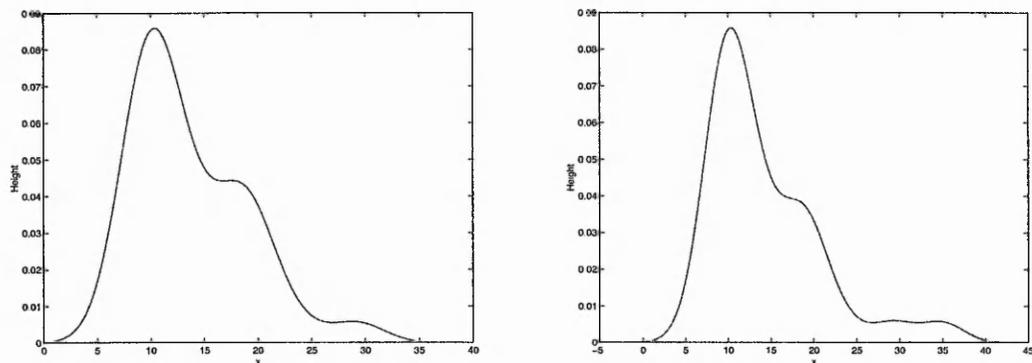
(a) KDE showing a 'bump' which may be classified as spurious

(b) KDE showing spurious modes in the tails

Figure 4.12: One-dimensional KDEs showing examples of spurious structure.

can be ignored. For example, look at blocks of 2 or 3 points, for the bivariate case a grid of 2×2 or 3×3 points could be used. For each block of points, compare their average height to neighbouring blocks, if the average height is greater than all of the neighbouring blocks, a mode is present. This would help to remove some of the forms of spurious structure which were illustrated in figure 4.12.

In section 4.2.3, alternative methods to univariate mode counting were discussed. Not all of these techniques can be generalised to the bivariate case, however the concept of 'excess mass' could be extended to the bivariate case. A further alternative would be to apply the EMMIX technique (McLachlan *et al.*, 1999). This technique can be used to determine the component normal pdf's that form the population. This method would not necessarily determine the number of modes in the population however once the components are determined simulated versions of the population could be used to further investigate modality.

4.5 Two-dimensional applications

In this section the techniques discussed above are applied to the bivariate case. For the larger data sets of Stos-Gale *et al.* (1996) and Gale *et al.* (1997), the three bivariate pairs of ratios

$$^{208}\text{Pb} / ^{206}\text{Pb} \ v ^{207}\text{Pb} / ^{206}\text{Pb},$$

$$^{208}\text{Pb} / ^{206}\text{Pb} \ v \ ^{206}\text{Pb} / ^{204}\text{Pb},$$

^{207}Pb $/^{206}$Pb v ^{206}Pb $/^{204}$Pb

are repeatedly sub-sampled 1000 times for varying sample size, n. Each bivariate sub-sample is subsequently tested with the ISE statistic (section 4.4.1).

The 3 individual univariate ratios of the Keos field (Stos-Gale $et\ al.$ 1996) do not suggest a departure from univariate normality, with ISE statistics for the 3 ratios being 0.003, 0.0024 and 0.0015 respectively, all of which are not significant. In section 4.3 we demonstrated that using the Malkovich and Afifi (1973) extension to the Shaprio-Wilk statistic, it was possible to find a linear combination of the ^{208}Pb/^{206}Pb and ^{207}Pb/^{206}Pb ratios which was clearly non-normal and bimodal (figure 4.11). This is sufficient to show that trivariate normality cannot be assumed.

Figure 4.13 shows a bivariate plot of the two ratios ^{208}Pb/^{206}Pb v ^{207}Pb/^{206}Pb which the Malkovich and Afifi (1973) statistic combined to create the non-normal univariate view. We have repeatedly sub-sampled without replacement from these ratios for varying sample sizes and tested for departures from normality with the ISE statistic. For $n = 30$ the hypothesis of normality is rejected about 35% of the time, increasing to 84% for sample sizes of $n = 50$. For all $n = 62$ cases, the ISE statistic is 0.0078 which is significant at the the 5% and 1% levels.

For the other 2 pairs of ratios, the hypothesis of normality is accepted no more than 10% of the time irrespective of sample size.

Although the non-normality of the Keos field has previously been established by finding the most non-normal linear combination of variables, this example illustrates that the non-normality of the field can also be established by examining bivariate pairs of ratios.

The univariate ratios of the Larnaca data ($n = 73$, Gale $et\ al.$, 1997) are clearly non-normal with values of the ISE statistic of 0.0189, 0.0345 and 0.0215 which are all significant. Figure 4.14 shows contour plots of the ^{208}Pb/^{206}Pb v ^{207}Pb/^{206}Pb and ^{208}Pb/^{206}Pb v ^{206}Pb/^{204}Pb ratios of the Larnaca data and suggests both pairs of ratios are highly skewed and possibly bimodal.

Repeatedly sub-sampling 1000 times from the bivariate pair ^{208}Pb/^{206}Pb v ^{207}Pb/^{206}Pb for $n = 25$ results in rejection of the hypothesis of normality 70% of the time. For $n = 35$, the hypothesis of normality is rejected over 90% of the time. Similar results

59
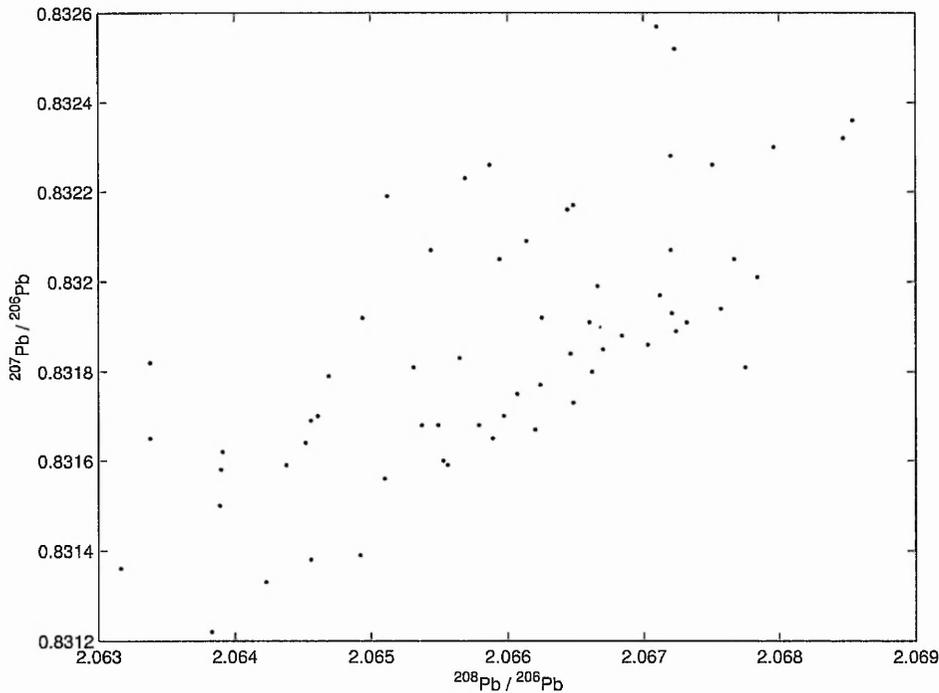
Figure 4.13: Bivariate scatter plot of ^{208}Pb/^{206}Pb v ^{207}Pb/^{206}Pb ratios of the Keos ore field, $n = 62$ (Stos-Gale *et al.*, 1996).

are evident when taking sub-samples from the ^{207}Pb /^{206}Pb v ^{206}Pb /^{204}Pb bivariate pair, with rejection of the hypothesis of normality in 90% of cases for $n \geq 25$.

Repeating this for the ^{208}Pb/^{206}Pb v ^{206}Pb/^{204}Pb bivariate pair suggests that for $n = 25$ the hypothesis of normality is not rejected in all but a handful of cases (less than 1%), however for $n = 55$ the hypothesis of normality is rejected 73% of the time.

The Solea axis has $n = 50$ observations (Gale *et al.*, 1997). As with the Keos data discussed above, the individual univariate ratios suggest no apparent departures from normality with ISE statistics of 0.0016, 0.0051 and 0.0027, none of which are significant at the usual levels. As before, we took 1000 sub-samples from the ^{208}Pb/^{206}Pb v ^{207}Pb/^{206}Pb pair of ratios. This resulted in the hypothesis of normality being rejected less than 40% of the time for $n = 20$ and over 75% of the time for $n = 30$. For sample sizes greater than 30, normality is rejected over 90% of the time. For the ^{208}Pb/^{206}Pb v ^{206}Pb/^{204}Pb bivariate pair of ratios, 40 or more samples are required to detect departures from normality over 80% of the time.

The examples above have demonstrated that in some cases bivariate non-normality is very apparent, even with sample sizes of 20. In other cases, larger sample sizes,

(a) $^{208}Pb/^{206}Pb$ v $^{207}Pb/^{206}Pb$ (b) $^{208}Pb/^{206}Pb$ v $^{206}Pb/^{204}Pb$
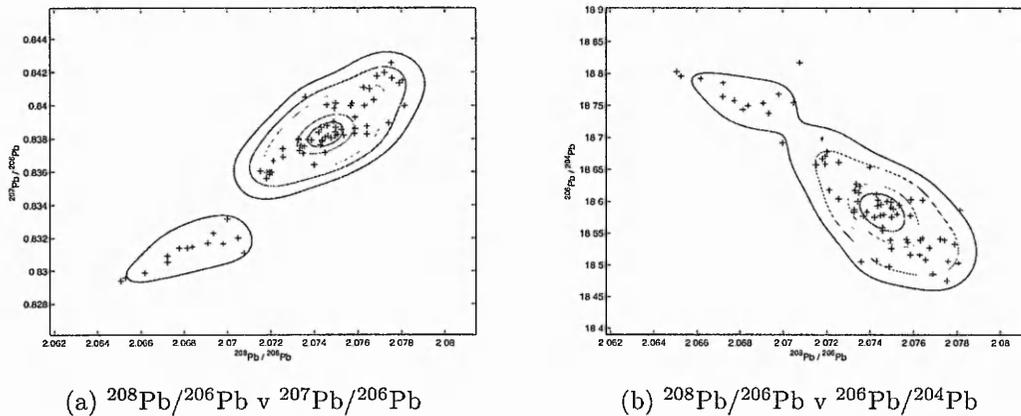
Figure 4.14: Contour and scatter plots of two of the bivariate pairs of ratios of the Larnaca field, $n = 73$ (Gale $et~al.$, 1997). In both cases h is calculated by the DPI2 routine.

for example $n \geq 35$ are required to detect departures from normality a substantial proportion of the time. This further illustrates that 20 samples may be insufficient to correctly represent structure. We have also illustrated that, although lead isotope data may have three ratios that, marginally, do not appear to be non-normal, bivariate combinations of the ratios can highlight very apparent departures from normality.

4.6 The three-dimensional case

The ISE statistic discussed in section 4.2.1 and 4.4.1 can be directly extended into higher dimensions, however as dimensionality increases, the computational complexity also increases. Given sufficient critical values, the analyses undertaken in section 4.3 and 4.5 can be extended to the trivariate case.

The mode counting approach discussed in section 4.4.2 has not been extended to the trivariate case due to the issues previously detailed. Theoretically the technique can be extended to the trivariate case however the process would be extremely computationally complex. For the bivariate case a 3×3 grid of heights surrounding a point are examined. For the trivariate case a $3 \times 3 \times 3$ cube of points would be considered. Computationally this approach would be very demanding. At the time of writing, we are not aware of any direct measure of assessing the modality of trivariate data. The EMMIX technique (McLachlan $et~al.$, 1999) could be applied to trivariate

data to assess the components which form the population, which may give insight into the modality of the data.

4.7   Three-dimensional applications

Further confirmation of the results obtained in sections 4.3 and 4.5 comes from direct testing of three-dimensional normality. Baxter, Beardah and Westwood (2000) present 3-dimensional KDEs which provide evidence that the Lavrion lead isotope data (Gale *et al.*, 1996) is non-normal and bi-modal. A 3-dimensional KDE requires 4-dimensions to represent it, this method of data display is not easily appreciable on a 2-dimensional computer screen or printed page. Baxter, Beardah and Westwood (2000) display 3-dimensional KDEs with different percentage contour shells to show structure. Using the MATLAB routines of Beardah and Baxter (1995), the representation of a 3-dimensional KDE on a 2-dimensional computer screen can be enhanced, for example by rotating the percentage contour shells in 3-dimensional space, this gives an understanding of structure. Although it was not mentioned by Baxter, Beardah and Westwood (2000), it is possible to use animation to illustrate the extent of varying contour shells.

Baxter, Beardah and Westwood (2000) note that statistical tests of multivariate normality provide reassurance that the data are non-normal. Although specific results are not given, they cite Baxter and Gale (1998) and Baxter (1999) who discuss the statistical analysis of the data in detail. The work presented below serves to support that of Baxter, Beardah and Westwood (2000) by demonstrating departures from normality using direct testing of trivariate normality.

Considering initially the Lavrion ($n = 59$) field (Gale *et al.*, 1996). In section 4.3 we have demonstrated that direct testing of the univariate ratios suggested a clear departure from normality. Considering now the three-dimensional data as a whole, we have repeatedly sub-sampled (1000 times) for different sample sizes and used the ISE statistic to test for departures from normality. For sample sizes of $n = 20$, the null hypothesis of normality is rejected about 37% of the time rising to over 85% for sample sizes of 40 or larger.

The Keos ($n = 62$) field (Gale *et al.*, 1996) has been used to demonstrate that linear combinations of variables can be used to demonstrate non-normality, although the

individual ratios do show departures from normality. Again, taking 1000 sub-samples (without replacement), with a sample size of 20 the null hypothesis of normality is rejected approximately 20% of the time. For $n = 50$, this increases to nearly 70%.

For the Larnaca data, which is obviously non-normal on bivariate plots (Stos-Gale et al., 1997) much smaller samples suffice to detect departures from normality in a formal way. A sample size of $n = 25$ results in rejection of the hypothesis of normality nearly all the time.

These results confirm that a sample size of 20, widely quoted as appropriate for lead-isotope data, is inadequate to detect non-normality in some cases.

4.8 Conclusions

In the previous sections we have made use of some of the larger LIA (Lead Isotope Analysis) data sets of Stos-Gale et al. (1996) and Gale et al. (1997) and simulated populations to investigate the normality of the lead isotope fields. It is commonly assumed that lead isotope fields are trivariate normal (Scaife et al., 1996), an assumption required for calculation of confidence ellipsoids and probability calculations. As such many of the published conclusions depend on this assumption of trivariate normality.

We have demonstrated that the trivariate normality of LIA data cannot be assumed. Of the data analysed, we have demonstrated that there are several examples which show clear departures from normality in the form of multimodality, in some cases non-normality is evident with around 20 observations. For small departures from normality, a larger sample size is needed. The history of the Cyprus field (summarised in section 4.1.2) shows that even with very obvious multimodality (evident from later data collection), even $n > 40$ is not always sufficient to show departures from normality and has led to a lot of controversy. Conclusions from most techniques should be robust to relatively small departures from normality, however for the examples cited above, conclusions based on techniques which assume trivariate normality should be questioned.

Sayre et al. (1992) conclude that lead isotope fields are trivariate normal by testing the univariate normality of each of the three individual ratios in isolation. We have demonstrated, for the Keos field for example, that although the individual ratios may

63

be normally distributed, bivariate pairs of ratios and direct testing of trivariate normality using the ISE statistic shows clear departures from normality. We have also demonstrated that linear combinations of the individual ratios can be found which again show clear departures from normality. This approach can be considered as a form of projection pursuit which is discussed further in chapter 6.

The analyses undertaken are based on larger sample sizes than are usually available, Pollard and Heron (1996) suggest that 20 samples is an acceptable minimum. The analyses above suggest that a sample size of 20 is not always sufficient to detect departures from normality and that in some cases 40 or more samples are required to highlight non-normality.

4.9   Further work

Our approach to mode counting has raised a number of issues. Some suggested modifications to the current technique were discussed in section 4.4.2, however the numerical approach used to recognise relevant structure automatically has proven more complex than was originally anticipated. Alternative approaches may provide more satisfactory results, for example neural networks can be trained to recognise patterns in data. The intention is that neural networks are able to mimic simple human pattern matching which is far more subjective than using simple mode counting methods. Attempts at this have, however, raised many issues. It is felt that a more reliable approach may lie with the use of formal mode counting approaches, such as those discussed in section 4.2.3 and in the following chapter. An investigative study of these techniques would clearly be useful as much of the published literature is theoretical in nature.

The work on the bivariate case is limited and a fuller investigation would give deeper insight into sample size requirements for various forms of structure.

Our mode counting methodology for univariate and bivariate data involves identifying all modes and attempting to use some automated process to ignore spurious structure. An outcome of the work was an insight into the performance of different window-width selection routines, as well as the use of adaptive density estimates. Our, simplistic, approach to mode counting has however raised a number of issues which are discussed above. Better methods of mode counting could be explored, for example the

theory of excess mass could be extended to the 2-dimensional case. Extension beyond this would require dealing with non-visible dimensions.

Chapter 5

Mode Counting with Kernel Density Estimates

5.1 Introduction

The mode counting work discussed in section 4.2.2 suggested that adaptive KDEs outperform non-adaptive KDEs in that they "better" display the underlying structure of the population from a given sample.

In section 4.2.3 a number of methods of mode counting were discussed as alternatives to our numerical approach used in sections 4.2.2 and 4.4.2. One alternative is that of Silverman (1981) who makes use of kernel density estimates to estimate the true number of modes in a univariate population. Izenman and Sommer (1988) propose that Silverman's original test be extended to make use of adaptive density estimates which they also suggest are better able to display the underlying structure of a population. They do not however discuss implementation.

This chapter begins by exploring the original test of Silverman (1981) applied to a number of archaeometric applications and goes on to propose an adaptive version. The chapter concludes with a comparison of the two versions of the test. This is used to gauge whether our adaptive test is easier to interpret than the original test of Silverman (1981) and if it is of use for relatively small sample sizes.

5.2 Silverman's test of modality

Given a univariate sample, $X_1, X_2, \ldots, X_n$, it is possible to estimate the modality of the underlying population by sequentially testing for $1-, 2-, \ldots$ modes by fitting KDEs

to the data.

For the general case of k-modality we test H_0: The number of modes is k against H_1: The number of modes is greater than k. Silverman's (1981) approach, which we adopt, is to sequentially test for $k = 1, 2, \ldots$ modality. He begins by defining the k-critical window width h_{crit}^k as

$$h_{crit}^k = \inf\{h : \hat{f}(x, h) \text{ has at most } k \text{ modes}\}$$

which is the smallest possible value of h giving just k modes. Silverman uses a simple binary search procedure to find h_{crit}^k. He notes that in practice, an interval of (h_{lo}, h_{hi}) in which h_{crit}^k is known to lie can be halved in length by checking whether the value of $\frac{1}{2}(h_{lo} + h_{hi})$ leads to $k + 1$ modes. Silverman's original work does not discuss how modes are counted, we use our numerical method of mode counting as discussed in section 4.2.2. The mode counting is performed with no threshold value as we are interested in a KDE with just k modes.

The sample variance, $\hat{\sigma}^2$, should be estimated, thus :

1. Take a bootstrap sample from $X_1, X_2, \ldots, X_n$, i.e. sample n observations with replacement, and label the sample $Y_1, Y_2, \ldots, Y_n$.

2. Smooth the bootstrap sample, using

$$X_i^* = (1 + h_{crit}^k{}^2/\hat{\sigma}^2)^{-1/2}(Y_i - h_{crit}^k \epsilon_i) \tag{5.1}$$

where ϵ_i is a random variable from the standard $N(0, 1)$ pdf. This ensures that samples are independent.

3. For the sample of $X_1^*, X_2^*, \ldots, X_n^*$, find the window-width $\hat{h}$ such that a non-adaptive KDE (equation 3.2) has just k modes. Silverman (1981) notes that a computational shortcut is to construct a KDE of the smoothed bootstrap sample using h_{crit}^k and determine if this has more than k modes.

This process is repeated a large number of times; Silverman (1981) used 100 repetitions however Efron and Tibshirani (1993) suggest that 500 times is reasonable. We have experimented with the number of repetitions and believe that 500 provides acceptable results. A summary of our results comparing the number of repetitions are

67

presented in the following section. The significance of the test (or p value) is then the proportion of repetitions for which $h_{crit}^k > \hat{h}$ (or for which the number of modes of a KDE based upon $X_1^*, X_2^*, \ldots, X_n^*$ using h_{crit}^k is found to be more than k).

If the test is carried out sequentially by computing p_k (the significance level) for $k = 1, 2, \ldots$ until a sufficiently large p value is reached, it is possible to estimate the number of modes within the population. What is meant by 'large' is not clear. Izenman and Sommer (1988) suggest that a p value of 0.4 is reasonable (but should be used with caution). More commonly in statistics, a significance level of 5 or 10% level is used to support a hypothesis. Our experience has suggested that the p value is dependent on the structure of the data, in many cases a stopping rule of $p = 0.4$ does not support the hypothesis expected and a stopping rule of $p = 0.1$ appears to provide a conservative test, that is under-estimating the number of modes.

5.3 Examples

Very few published applications of Silverman's test of modality can be readily located. Silverman originally demonstrated the test on $n = 20$ observations made on chondrite meteors (Good and Gaskins, 1980). Izenman and Sommer (1988) extensively investigate $n = 485$ observations on philatelic data and Efron and Tibshirani (1993) reproduce a similar analysis, reaching slightly different conclusions.

We begin by replicating the analysis of Izenman and Sommer (1988) which illustrates a number of important points that arise in using the test. We then go on to apply the test to a number of lead isotope data sets and finish with an example based on the chemical composition of glass fragments.

5.3.1 Postal stamps data

Figure 5.1 shows two non-adaptive KDEs of the thickness of $n = 485$ postal stamps from the 1872 Hidalgo stamp issue of Mexico, as used by Izenman and Sommer (1988). The data used are univariate and consist of a measurement of the stamp thickness in millimetres. The aim of the analysis undertaken by Izenman and Sommer was to assess how many printings had occurred for this particular issue of stamps. Figure 5.1a uses an h value of 0.0046, calculated using the Normal Scale rule, and suggests the presence

(a) Normal Scale   (b) Solve-the-Equation
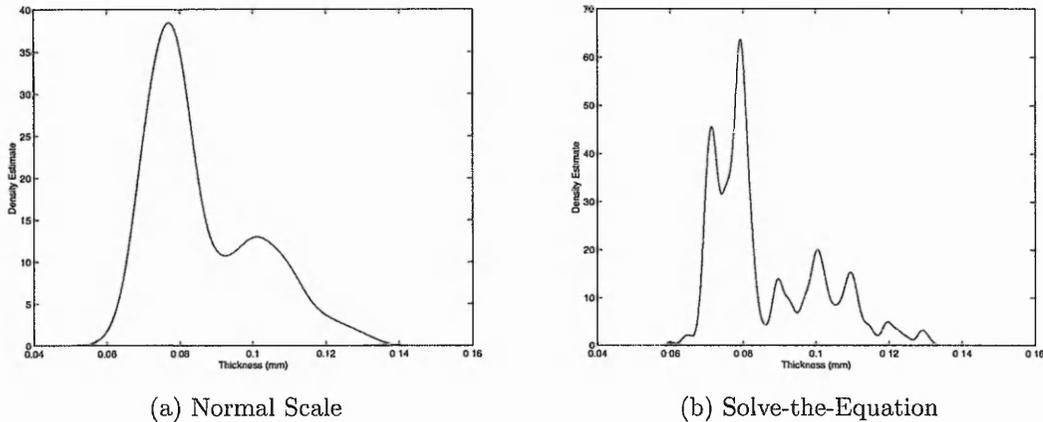
Figure 5.1: KDEs showing thickness measurements on 485 stamps generated using (a) $h_{NS} = 0.0046$ and (b) $h_{STE} = 0.0012$.

k	h^k_{crit}	p_k for 100 reps	p_k for 500 reps	p_k for 1000 reps
1	0.0068	0.000	0.004	0.001
2	0.0032	0.200	0.300	0.288
3	0.0030	0.040	0.044	0.057
4	0.0028	0.000	0.006	0.008
5	0.0026	0.000	0.002	0.003
6	0.0024	0.000	0.002	0.000
7	0.0015	0.410	0.514	0.492
8	0.0014	0.300	0.266	0.266
9	0.0011	0.580	0.560	0.539
10	0.0010	0.430	0.336	0.307

Table 5.1: Critical window widths and significance levels for the 1872 Hidalgo stamp thickness data.

of 2 modes, figure 5.1b uses an h value of 0.00123 calculated by the Solve-the-Equation method and clearly contains 9 modes with the 2 smaller modes to the left.

Silverman's test for modality was applied to 100, 500 and 1000 bootstrap samples of size 485, the resulting p values from sequentially testing for the presence of $k = 1, 2, 3, \ldots, 10$ modes are presented in table 5.1. The p value does not vary a great deal with a change in the number of repetitions, therefore we follow the example of Efron and Tibshirani (1993) and base our analysis on 500 repetitions.

The values in table 5.1 for 500 repetitions are similar to those printed in Izenman and Sommer (1988) and in section 16.5 of Efron and Tibshirani (1993) (although our value for $k = 9$, and that of Izenman and Sommer (1988), varies from that printed by Efron and Tibshirani by a considerable degree, we believe it may be due to a different

69

(a) $k = 1$ 　　　　　　　　　　　　(b) $k = 2$

(c) $k = 5$ 　　　　　　　　　　　　(d) $k = 7$
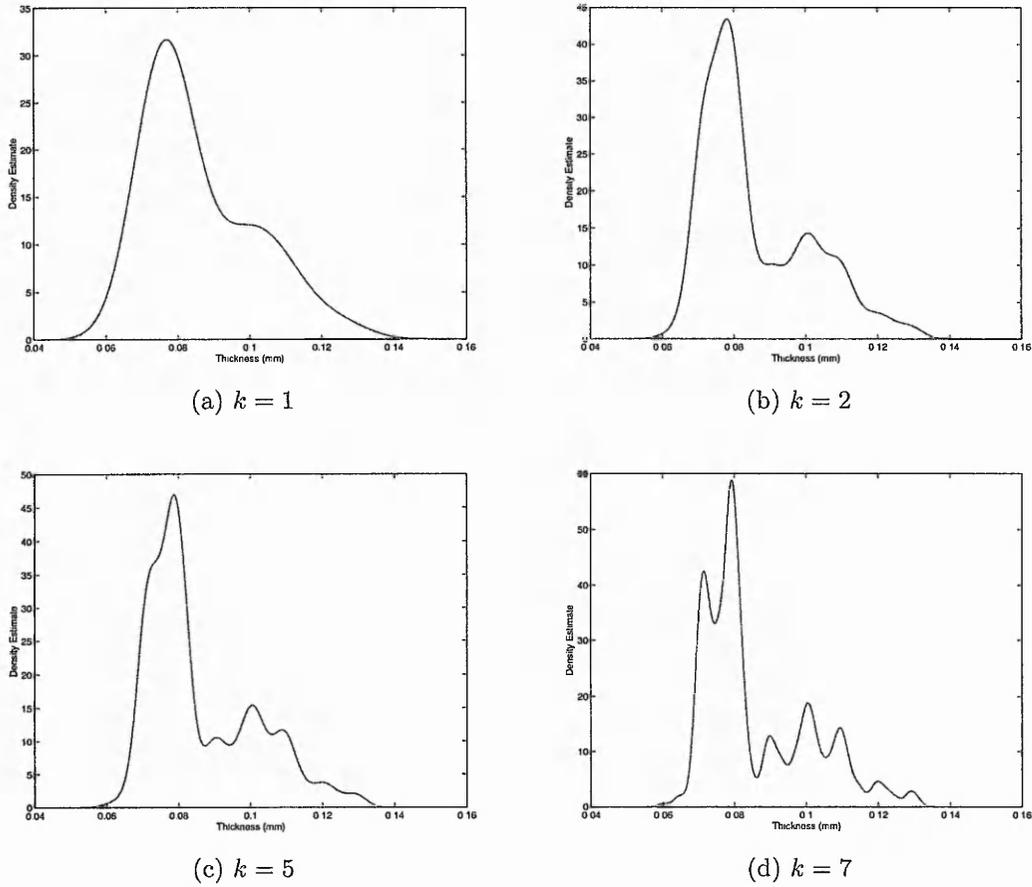
Figure 5.2: KDEs of the 1872 Hidalgo stamp thickness data with $n = 485$ constructed with h^k_{crit} for $k = 1, 2, 5$ and 7 where h^k_{crit} is the smallest possible value of the bandwidth resulting in a KDE with k modes.

h^9_{crit} value). If a 'stopping rule' (i.e. the point at which the hypothesis is accepted) of $p \geq 0.4$ is used based on the suggestion of Izenman and Sommer (1988), the results suggest the presence of 7 modes. Based on this, the hypothesis of 9 modes ($p_9 = 0.56$) would be rejected as $k = 7$ has already been accepted. Although figure 5.1a clearly suggests the presence of just two modes, the p_2 value of 0.3 is inconclusive (however Efron and Tibshirani (1993, 231-2) do not reject the hypothesis of bimodality, but suggest that 7 or even 9 modes could be considered as an alternate hypothesis).

Izenman and Sommer (1988) use $p \geq 0.4$ as a stopping rule. This suggests that they are willing to accept an incorrect null hypothesis 40% of the time. A more traditional approach is to investigate significance at the 5% or 10% level. Using this approach, $k = 2$ is the preferred hypothesis with a p_2 value of 0.3, which agrees with the KDE

in figure 5.1b. Silverman (1981) notes that the test provides a conservative estimate of modality, as such the hypothesis of $k = 2$ should be considered as a conservative estimate of modality. We can consider the hypothesis of $k = 7$ if we are willing to accept a 40% chance of being wrong.

Izenman and Sommer (1988) suggest it is helpful to examine the KDE at the critical window-width, h_{crit}^{k}, to help in determining an appropriate stopping rule. The KDEs for h_{crit}^{k} for the hypothesis of $k = 1, 2, 5$ and 7 are shown in figure 5.2. Considering figure 5.2a, this KDE suggests 1 mode, however there is also suggestion of an additional mode (or "bump"), the same can be seen in figure 5.2b and c. However, for 5.2d there are clearly 7 defined modes, as opposed to small bumps, which supports the hypothesis reached from the p values. For KDEs constructed from h_{crit}^{k} other than $k = 7$, there are always small, less defined modes. As such this method of assessing significance appears to be more useful than p values alone in this example.

5.3.2   Lead isotope data

In section 4.3, it was observed that a number of the lead isotope data sets from Stos-Gale *et al.* (1996) and Gale *et al.* (1997) exhibited signs of non-normality when their one-dimensional marginals were examined in isolation. Of the data sets examined, the $^{208}Pb/^{206}Pb$ ratio of the Lavrion field ($n = 59$) appears clearly bimodal when adaptive KDEs are used. Figure 5.3 shows non-adaptive KDEs which appear bimodal when h_{NS} is used, more clearly so for h_{STE}. There is a suggestion of trimodality with the possibility of a small mode to the left of the plot, however bimodality is the preferred interpretation. The $^{208}Pb/^{206}Pb$ ratio of the Larnaca field ($n = 73$) appears trimodal when adaptive KDEs are used (figure 4.9). On the other hand this also appears clearly bimodal when non-adaptive KDEs are used, as shown in figure 5.4.

Tables 5.2 and 5.3 show p values and critical window widths for the hypothesis of $k = 1, \ldots, 4$ modes in the case of the Lavrion $^{208}Pb/^{206}Pb$ and the Larnaca $^{208}Pb/^{206}Pb$ ratio respectively. The p values generated for testing the $^{208}Pb/^{206}Pb$ Lavrion ratio suggest that there are 2 modes ($p_2 = 0.588$) which agrees with the visual interpretation of the KDE. A conservative estimate of $k = 1$ is significant at the 10% level with a p_1 value of 0.264.
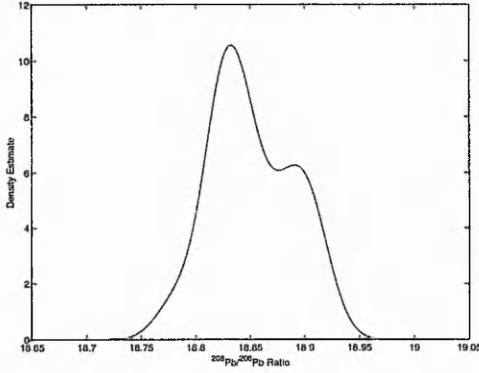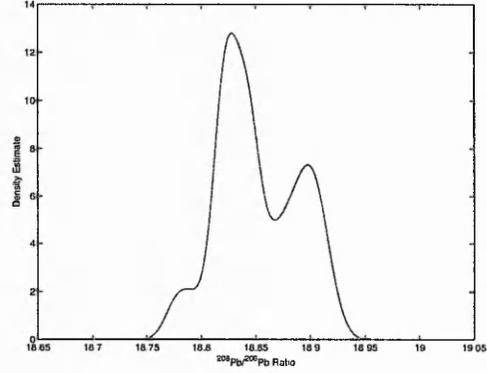
| (a) Normal Scale | (b) Solve-the-Equation |

Figure 5.3: KDEs showing the $^{208}Pb/^{206}Pb$ ratio of the Lavrion field generated using $h_{NS} = 0.0173$ and $h_{STE} = 0.0101$ (Stos-Gale *et al.*, 1996).
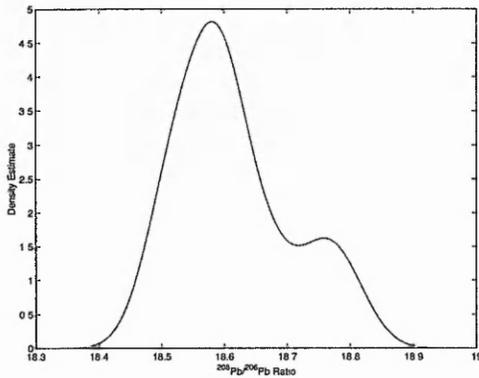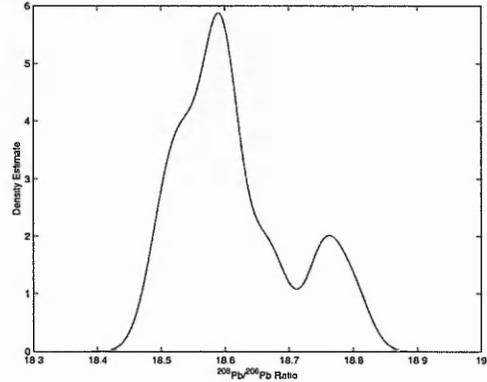
| (a) Normal Scale | (b) Solve-the-Equation |

Figure 5.4: KDEs showing the $^{208}Pb/^{206}Pb$ ratio of the Larnaca field generated using $h_{NS} = 0.0400$ and $h_{STE} = 0.0243$ (Gale *et al.*, 1997).

For the $^{208}Pb/^{206}Pb$ ratio of the Larnaca field, a p_2 value of 0.494 also supports a hypothesis of $k = 2$ modes, which agrees with the non-adaptive KDEs in figure 5.4. The adaptive KDE shown in figure 4.9 hints towards trimodality, however the non-adaptive KDEs appear to 'hide' this extra mode, which may be spurious structure.

Unlike the analysis of the stamp thickness data discussed in the previous section, in these examples it was found that KDEs of the data at the critical window widths were uninformative. Whereas with the stamps data, the KDE with h^7_{crit} had more defined modes than for other k, which helped accept the hypothesis of $k = 7$, the KDEs for the two data sets examined in this section did not follow this pattern.

k	h_{crit}^k	p_k
1	0.0010	0.264
2	5.8003e-4	0.588
3	5.3762e-4	0.268
4	4.3003e-4	0.186

Table 5.2: Critical window widths and significance levels for the 59 observations from the $^{208}Pb/^{206}Pb$ ratio of the Lavrion field based on 500 repetitions.

k	h_{crit}^k	p_k
1	0.0458	0.032
2	0.0206	0.494
3	0.0201	0.108
4	0.0139	0.356

Table 5.3: Critical window widths and significance levels for the 73 observations from the $^{208}Pb/^{206}Pb$ ratio of the Larnaca field based on 500 repetitions.

5.3.3 Iron to manganese ratios in glass from Saxon Southampton

Figure 5.5 shows KDEs of 225 iron (Fe) manganese (Mn) ratios that were taken from glass samples excavated from Saxon Southampton (Hamwic), from Hunter and Heyworth (1998). Of the 225 samples, 163 are light blue in colour and the remaining 62 are light green. This structure is evident in figure 5.5 by the two main modes to the left of the KDEs. When h_{STE} is used, there also appears to be a number of outlying modes to the right. As previously, the 225 samples were subjected to sequential testing for $m = 1, \ldots, 10$ modes, the results for which are presented in table 5.4.

Examining the p-values generated by the test is inconclusive using Izenman and Sommer's (1988) suggested stopping rule as there is no obvious point at which the p-values exceed 0.4. However the p-value for $k = 4$ modes is 0.342, and for 9 modes is 0.392, which are the highest values of p. This does not provide conclusive evidence to support $k = 4$ or $k = 7$, so neither hypothesis could be accepted. The p-value for $k = 1$ is 0.192, the null hypothesis should not be rejected if we use a conservative stopping rule of $p \geq 0.1$.

Figure 5.6 shows KDEs for h_{crit}^k for selected k. Figure 5.6d shows the KDE with a critical window-width giving just 4 modes, which are not clearly defined, thus the hypothesis of $k = 4$ should be not be accepted. By similar reasoning, the hypothesis of $k = 9$ modes cannot be accepted (figure 5.6e). The hypothesis of $k = 3$ seems unlikely
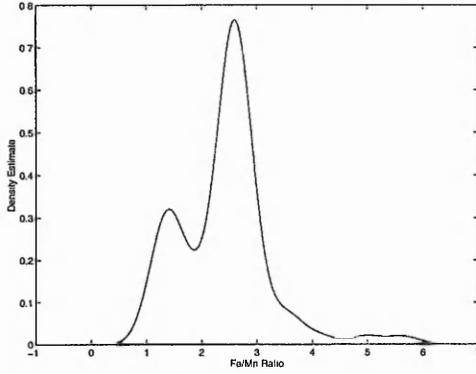
(a) Normal Scale
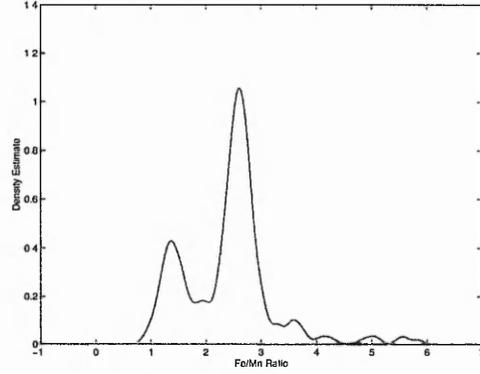
(b) Solve-the-Equation

Figure 5.5: KDEs showing the iron to manganese ratio of glass from Saxon Southampton suggesting 4 and 7 modes (constructed with $h_{NS} = 0.2404$ and $h_{STE} = 0.104$ respectively).

k	h_{crit}^k	p_k
1	0.4634	0.192
2	0.3538	0.01
3	0.2807	0.02
4	0.1749	0.342
5	0.1575	0.112
6	0.1200	0.212
7	0.1057	0.172
8	0.1035	0.034
9	0.0706	0.392
10	0.0673	0.282

Table 5.4: Critical window widths and estimated achieved significance for the Saxon Southampton glass data based on 500 repetitions.

given that $p_3 = 0.02$, however the KDE for h_{crit}^2 appears smooth, which is in line with earlier conclusions.

For this particular data set it is difficult to use this measure of modality to say with any certainty how many modes exist within the population. We know for certain that there are 2 groups, and would expect a bimodal KDE, however the p values do not support this hypothesis. A more conservative estimate of $k = 1$ seems more likely if we are unwilling to accept such a high probability of being wrong.
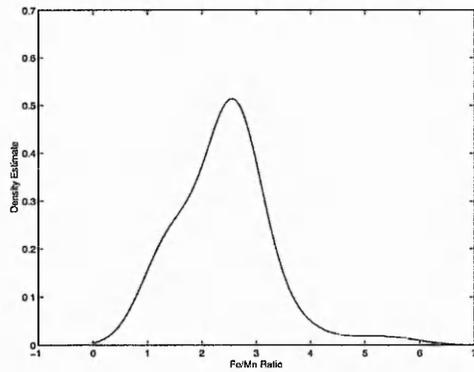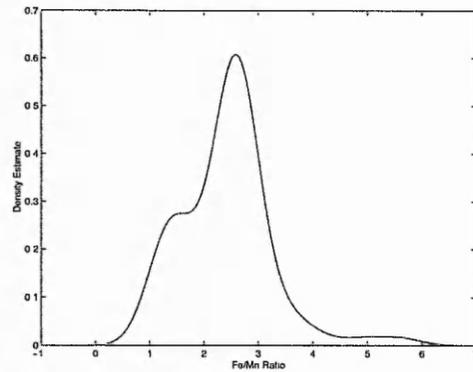
(a) $k = 1$ (b) $k = 2$
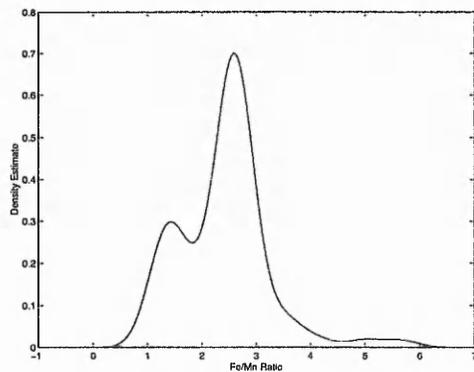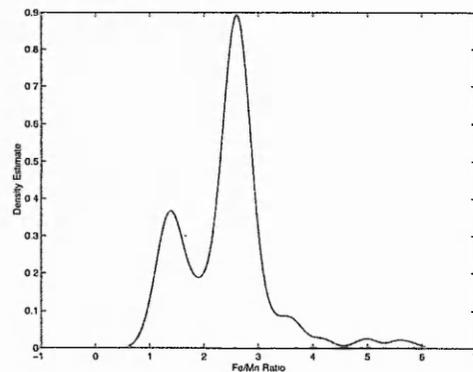
(c) $k = 3$ (d) $k = 4$
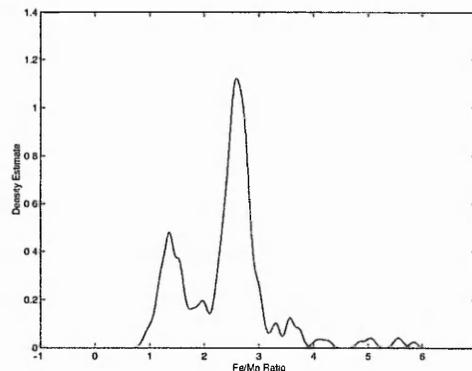
(e) $k = 9$

Figure 5.6: Non-adaptive KDEs showing the iron to manganese ratios in glass from Saxon Southampton data constructed using h_{crit}^k for $k = 1, 2, 3, 4$ and 9 for $n = 285$ observations.

5.4 Discussion of results

It has been demonstrated that interpretation of Silverman's (1981) test is subjective and often more difficult than might be anticipated. Although it has been shown that sensible conclusions can be drawn by accepting a hypothesis when a 'large' p value is reached, it has also been shown that interpretation based on p alone can be misleading. The 'large' values of p which are used to accept a hypothesis carry a high probability of being wrong. Using a stopping rule of $p = 0.1$ (10%) results in a conservative estimate of modality.

KDEs for h_{crit}^k prove useful in some cases, but in others can be misleading. Also, although Izenman and Sommer (1988) rely on this to reach their conclusions, there is no statistical justification, and interpretation is once again subjective.

Using a 'large' value of p, it is often the case that results are inconclusive and it is not possible to accept or reject a specific hypothesis, however for smaller values of p, the results seem conservative when compared to a visual inspection of the KDE. This may be related to the 'less than optimal' ability of non-adaptive KDEs to correctly produce accurate density estimates. In the following section, the possibility of extending the original test to make use of adaptive KDEs is discussed. As previous work has suggested that adaptive KDEs tend to reproduce more accurate density estimates than non-adaptive KDEs, the hope is that such a test will perform more satisfactorily and results will be more readily interpretable.

5.5 Adaptive version of Silverman's test

Izenman and Sommer (1988) suggest extending the work of Silverman (1981) to utilise adaptive kernel density estimates which, as illustrated in section 4.2.2, tend to produce more accurate estimates of the population's density from which the sample originates. Izenman and Sommer suggest using a vector of h values and smoothing bootstrap samples using some measure of average smoothness, but do not go into further detail. Below we present our suggestion for an adaptive version of Silverman's test of modality which has evolved from the original suggestion of Izenman and Sommer (1988).

Our adaptive test is very similar to that of Silverman (1981), discussed in section

5.2. Given a sample $X_1, X_2, \ldots, X_n$, estimate the sample variance, $\hat{\sigma}^2$.

Assuming the population is being tested for k-modality, construct an adaptive KDE

$$\tilde{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \lambda_i^{-1} K\left(\frac{x - X_i}{h\lambda_i}\right),$$

using a pilot bandwidth of h and vector λ of local bandwidth factors given by

$$\lambda_i = \{\tilde{f}(X_i)/g\}^{-\alpha} \qquad i = 1, 2, \ldots, n$$

as defined in section 3.2.3. In practice the pilot estimate of h can be any convenient estimate as the value will be scaled.

It is then necessary to find a vector of local bandwidth factors, λ which results in a KDE with just k modes, analogous to finding h_{crit}^k in Silverman's version of the test. Our approach to this is to scale λ by a scalar, s, which can be used to affect the modality of $\tilde{f}(x)$, leaving the pilot estimate of h constant throughout. Begin with λs where $s = 1$, generate an adaptive KDE based upon $X_1, X_2, \ldots, X_n$ and count the number of modes. If the number of modes is greater than k, multiply s by some value > 1 and create a new estimate of $\tilde{f}(x)$. If the number of modes is less than k, divide s by a value > 1.

We use the following algorithm to 'home' in on a value of λs to give just k modes which is a relatively quick way to find an appropriate value of s:

- Find a value of s such that an adaptive KDE, $\tilde{f}(x)$, created with local bandwidth factors λs, has less than k modes.

- Decrease s by small amounts until the number of modes changes from less than k to k.

We have found that scaling by small values initially can be very computationally expensive so prefer to begin by scaling s by larger values and continually decreasing the value by which s is scaled until the change needed in k to increase from $k - 1$ modes to k modes is suitably small. Now define

$$\mathbf{h}_{crit}^k = h\lambda s. \tag{5.2}$$

The actual test of modality can then be carried out by

1. Taking a bootstrap sample of size n from a vector $1, 2, \ldots, n$, that is sampled with replacement and label this vector $\mathbf{v}$. Now use $\mathbf{v}$ to create a bootstrap sample of $X_1, X_2, \ldots, X_n$. That is define $Y_i = X_{\mathbf{v}(i)}$. Also, define $\mathbf{h}_Y(i) = \mathbf{h}_{crit}^k(\mathbf{v}(i))$, that is use the $\mathbf{v}(i)^{th}$ value of $\mathbf{h}_{crit}$ to smooth Y_i. We take bootstrap samples from $1, 2, \ldots, n$ as it is necessary to select $\mathbf{h}_Y$ to apply the same amount of relative smoothing to each point throughout.

2. Smooth y using

$$Y_i^* = (1 + (\mathbf{h}_{Y(i)})^2 / \hat{\sigma}^2)^{-1/2}(Y_i + \mathbf{h}_{Y(i)}\epsilon_i) \qquad (5.3)$$

for $i = 1 \ldots n$ where ϵ_i is a random variable from the standard $N(0,1)$ normal pdf. This has the effect of scaling the sample so it has similar mean and variance to $X_1, X_2, \ldots, X_n$, and ensures that samples are independent.

3. Now construct an adaptive KDE based on $Y_1^*, Y_2^*, \ldots, Y_n^*$. We use $\bar{\mathbf{h}}_{crit}^k$ as a pilot estimate of h and define a new set of local bandwidth parameters, $\boldsymbol{\lambda}$. We then scale the new local bandwidth parameters, $h\boldsymbol{\lambda}$, such that $h\bar{\boldsymbol{\lambda}} = \bar{\mathbf{h}}_{crit}^k$. This results in a KDE with the same average smoothness as that used to calculate $\mathbf{h}_{crit}^k$.

This process is repeated, we use 500 repetitions. The effects of varying the number of repetitions has been investigated and 500 emerges as sufficient, with little difference to results obtained with 1000 repetitions. The estimated achieved significance level (or p value) is then the proportion of repetitions for which the number of modes in the adaptive KDE based $Y_1^*, Y_2^*, \ldots, Y_n^*$ on is found to be more than k.

As with the original test proposed by Silverman, the value of p at which the sequential testing of modality should cease is unclear. Our experiences, which are discussed in the following section, have been that the value is dependent on the data being tested.

The procedure described above has evolved with experimentation. Variations explored include using an average measure of smoothness instead of a scaled vector of local bandwidth parameters, as per the original suggestion of Izenman and Sommer (1988). A further variation investigated involved applying the same relative smoothing to each point in the creation of the KDE based on $Y_1^*, Y_2^*, \ldots, Y_n^*$. The method discussed above was selected for preference as it uses an adaptive procedure for estimating the modality of a sample.

5.6 Comparison

The following section discusses a comparison of the original test of Silverman (1981) and the adaptive version of the test proposed in section 5.5. Each version of the test is used to investigate the modality of a number of simulated populations with known structure, and thus known modality.

Populations are generated using the same technique detailed in section 4.2.1. Bimodal populations are of the form

$$0.5N(0,1) + 0.5N(\mu, 1) \tag{5.4}$$

for $2 \leq \mu \leq 4$ in increments of 0.5. Trimodal populations are generated using

$$\frac{1}{3}N(\mu_1, 1) + \frac{1}{3}N(0, 1) + \frac{1}{3}N(\mu_2, 1) \tag{5.5}$$

For μ_1 and μ_2 values between $-4 \leq \mu_1 \leq -2$ and $2 \leq \mu_2 \leq 4$. We have selected a set of 15 simulated populations with varying degrees of bimodality and trimodality.

Each of the 15 mixture distributions are used to create 100 samples of size $n = 40$, 60, 80 and 100. Each sample is then subjected to both Silverman's test and the adaptive test using 500 repetitions. Selected results are presented graphically within the text as cumulative frequency curves showing the number of replications for which p is greater than $0.1, 0.2, \ldots, 1$. A complete set of results is included in appendix C in the form of cumulative frequency tables.

As has been discussed in section 4.2.1 and in Everitt and Hand (1981), mixtures of the form defined in 5.4 with $\mu = 2$ are formed by mixing 2 populations but are unimodal in appearance. Table C.1 shows the resulting p values when Silverman's test and our adaptive test are applied to simulated data with this structure. The results show that Silverman's test results in the rejection of the correct hypothesis of $k = 1$ 35% of the time for samples of size 40 using a stopping rule of $p \geq 0.4$. The adaptive test results in rejection of the hypothesis of $k = 1$ 52% of the time, which suggests that the original test of Silverman is more successful at determining the correct modality. For larger n, the frequency with which the correct hypothesis of $k = 1$ is rejected increases (i.e. values of p are smaller) for both versions of the test.

For populations with just separated components, for example $\mu = 2.5$, results are shown in C.2 and are presented graphically as cumulative frequency curves in figure
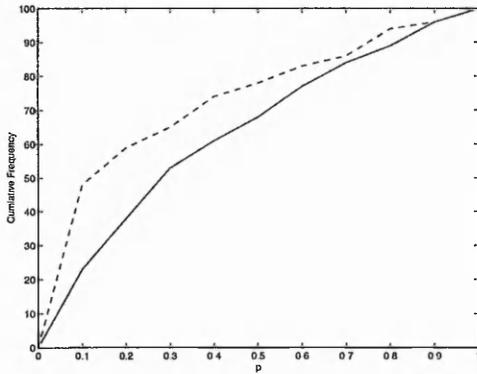
(a) $H_0 : k = 1$, $n = 40$
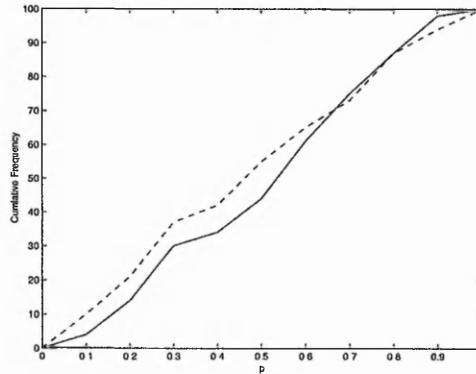
(b) $H_0 : k = 2$, $n = 40$
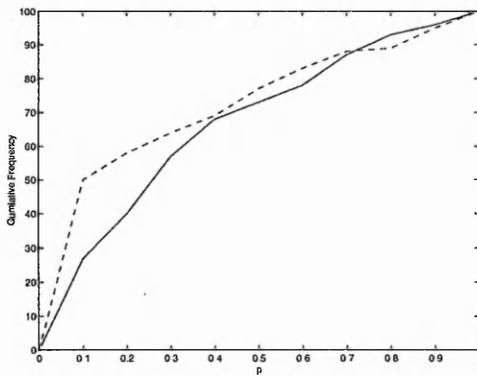
(c) $H_0 : k = 1$, $n = 100$
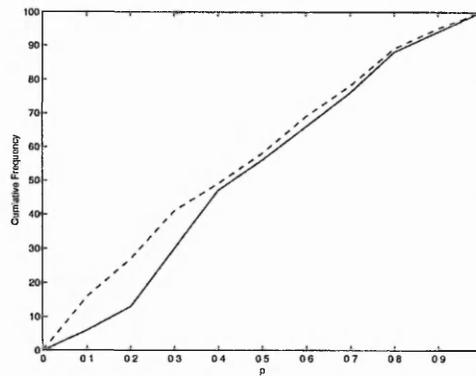
(d) $H_0 : k = 2$, $n = 100$

Figure 5.7: Cumulative frequency curves for a bimodal population with $\mu = 2.5$. p values resulting from Silverman's (1981) test of modality (solid line) and our adaptive test (dashed line) of modality (section 5.5). The curves shows the cumulative frequency with which $p = 0, \leq 0.1, \leq 0.2, \ldots, \leq 1.0$ for the hypothesis of $H_0 : k = 1$ against $H_1 : k > 1$ and $H_0 : m = 2$ against $H_1 : m > 2$.

5.7. The incorrect hypothesis of $k = 1$ is rejected more frequently when using the adaptive test; for $n = 40$, $k = 1$ is rejected 75% of the time with $p \geq 0.4$ used as a stopping rule and 61% of the time when using Silverman's test. However, the correct hypothesis of $k = 2$ is rejected 44% of the time by the adaptive test and only 34% of the time by Silverman's test. Although the p values for the hypothesis of $k = 1$ are smaller from the adaptive version of the test, they are also smaller for the hypothesis $k = 2$ which results in more frequent rejection of the correct hypothesis. Figure 5.7 does suggest that the difference between p values for the hypotheses of $k = 1$ and $k = 2$ are slightly greater for the adaptive test. As sample size increases, the frequency with which the hypotheses are rejected becomes similar for both versions of the test.

For populations with greater separation of components, i.e. $\mu \geq 3$, the adaptive test generally results in rejection of $k = 1$ more frequently that Silverman's test. As sample size increases, the frequency with which $k = 1$ is rejected increases for both tests, however for the hypothesis of $k = 2$ there is no obvious pattern.

Populations of the form detailed in equation (5.5) are trimodal if $|\mu_1|$ and $|\mu_2|$ are greater than 2. If $|\mu_1| = 2$ and $|\mu_2| > 2$, the population is bimodal. Table C.18 shows results for $\mu_1 = -2.5$ and $\mu_2 = 2.5$, a trimodal population with small separation between components. Figure 5.8 shows cumulative frequency curves of p values for this population. For $n = 40$ Silverman's test rejects the hypothesis of $k = 1$ 54% of the time with the adaptive test rejecting $k = 1$ 65% of the time, using $p \geq 0.4$ used as a stopping rule. For the hypothesis of $k = 2$, the adaptive test rejects this hypothesis 52% of the time, Silverman's test rejects this 53% of the time. For the correct hypothesis of trimodality the adaptive test would reject 44% of the time and Silverman's 49%. Thus the adaptive test is less likely to reject the hypothesis of $k = 3$ however Silverman's test is more likely to reject the incorrect hypothesis of $k = 1$ or 2. For other populations with small separation of components (i.e. $|\mu| \leq 3$) there is very little difference between the performance of the tests with both resulting in similar p values. The adaptive test is generally less likely to reject the correct hypothesis of $k = 3$ however it is also slightly less likely to reject the incorrect hypotheses of $k = 1$ or 2.

For trimodal populations with $|\mu| > 3$ there is little difference between results from
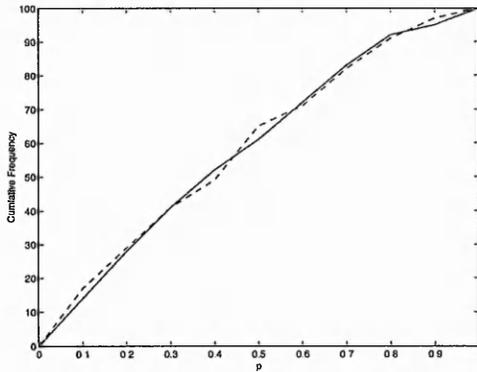
(a) $H_0 : k = 2$, $n = 60$
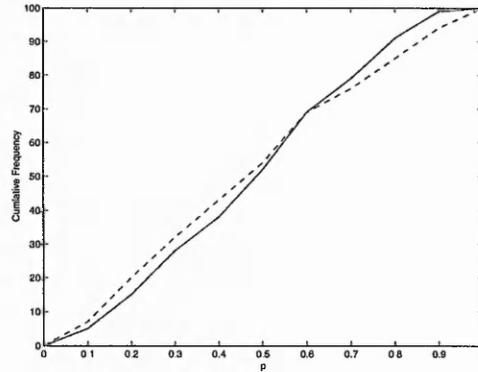
(b) $H_0 : k = 3$, $n = 60$
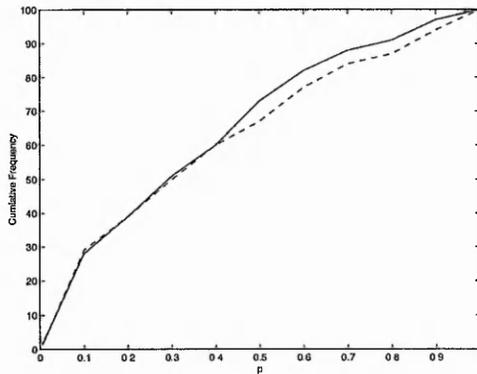
(c) $H_0 : k = 2$, $n = 100$
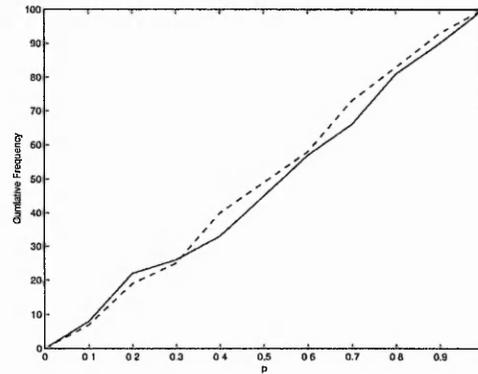
(d) $H_0 : k = 3$, $n = 100$

Figure 5.8: Cumulative frequency curves for a trimodal population with $\mu_1 = -2.5$ and $\mu_2 = 2.5$. p values resulting from Silverman's (1981) test of modality (solid line) and our adaptive test (dashed line) of modality (section 5.5). The curves shows the cumulative frequency with which $p = 0, \leq 0.1, \leq 0.2, \ldots, \leq 1.0$ for the hypothesis of $H_0 : k = 2$ against $H_1 : k > 2$ and $H_0 : m = 3$ against $H_1 : m > 3$.

82
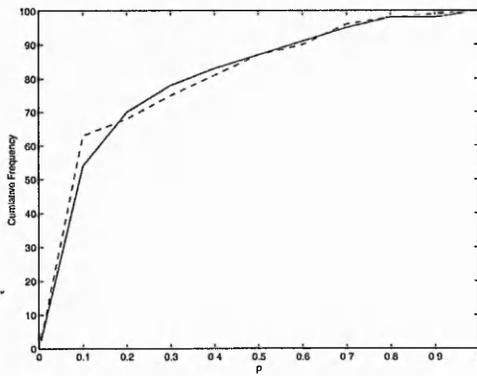
(a) $H_0 : k = 2$, $n = 60$
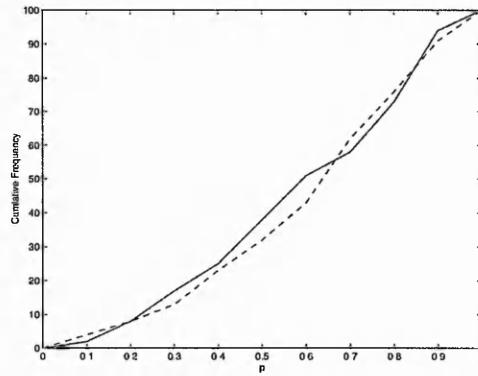
(b) $H_0 : k = 3$, $n = 60$
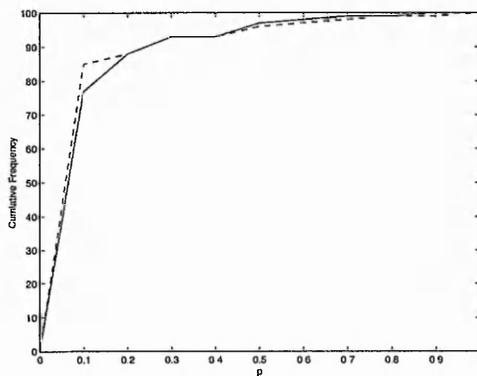
(c) $H_0 : k = 2$, $n = 100$
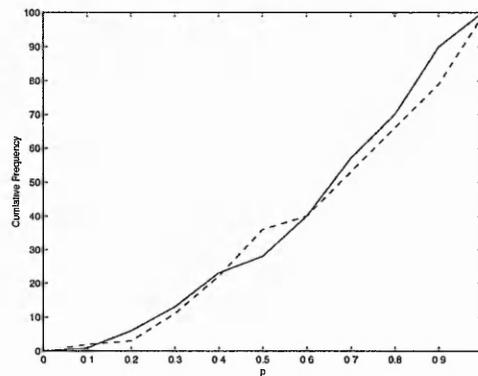
(d) $H_0 : k = 3$, $n = 100$

Figure 5.9: Cumulative frequency curves for a trimodal population with $\mu_1 = -3.5$ and $\mu_2 = 4$. p values resulting from Silverman's (1981) test of modality (solid line) and our adaptive test (dashed line) of modality (section 5.5). The curves shows the cumulative frequency with which $p = 0, \leq 0.1, \leq 0.2, \ldots, \leq 1.0$ for the hypothesis of $H_0 : k = 2$ against $H_1 : k > 2$ and $H_0 : m = 3$ against $H_1 : m > 3$.

either version of the test. Figure 5.9 shows a set of the cumulative frequency curves which is representative of all populations in this class.

5.7 Summary

Generally, it would appear that the adaptive version of the test provides p values that are easier to interpret than those arising from the original test proposed by Silverman.

For the bimodal populations considered, the adaptive test generally produces a lower p value associated with the incorrect hypothesis however it also results in a lower p value for the correct hypothesis which would result in the correct hypothesis

being rejected. There is a greater difference between p values for the correct and incorrect hypothesis for the adaptive test compared to Silverman's original test. This is of particular interest as a greater difference between p values for the correct and incorrect hypothesis means the chance of accepting the correct hypothesis increases. This can also give insight into selection of an appropriate stopping rule for particular forms of structure, however this value will be dependent on the, often unknown, form of structure within the data.

Both versions of the test fail to provide conclusive evidence to support any hypothesis for "intermediate" cases. Populations which are trimodal with component separation of less than 3 or bimodal with separation of less than 2.5 result in "poor" performance by both versions of the test. In some cases, larger sample sizes result in a clearer difference between p values for the correct and incorrect hypotheses. For trimodal populations with greater separation between components, the adaptive test tends reject the correct hypothesis less frequently that Silverman's original test, however there is little to choose between either version of the test.

Computationally, the adaptive test is far more "expensive" than the original proposed by Silverman. For example, the CPU time to calculate the p values for the hypothesis of $k = 1$ to 10 for the postal stamps data discussed in section 5.3.1 is approximately 40 seconds for Silverman's version of the test. For the adaptive version of the test, running under identical conditions, the CPU time is approximately 620 seconds. For our adaptive version of the test, the most computationally demanding exercise is fixing the smallest window-width vector to give just k modes.

Chapter 6

Projection Pursuit

6.1 Introduction

In previous chapters the emphasis has been on investigating structure in low dimensional data which can be directly visualised, specifically 1- to 3-dimensions. It is, however, more common to see analyses of artefact compositional data on an $n \times d$ data matrix, $\mathbf{x}$. In chapter 2, trends in the number of dimensions measured and used in statistical analyses were discussed. One of the findings was that d is typically in excess of 8. Clearly the techniques discussed in previous chapters cannot be directly applied to such data without first being transformed to a more manageable number of dimensions, then the previously discussed techniques can be applied. Dimension reduction techniques work on the assumption that much of the data are redundant and often attempt to display interesting structure in some graphical format. Multivariate techniques do exist which allow limited direct exploration of structure in high dimensional data, for example the ISE statistic of Bowman and Foster (1993) can be used to measure departures from normality in any number of dimensions, however the subjective manner in which structure is explored within plots is lost. Although direct measures of multivariate structure may give us, for example, an indication as to how non-normal data are, they do not replicate the human ability to visually interpret patterns within the high-dimensional data.

Of the dimension reduction methods available, cluster analysis and principal components analysis (PCA) are among the most widely used within artefact composition analyses. PCA is used to find d new variables, the principal components, which are uncorrelated linear combinations of the original variables. The first component has

maximal variance, subject to a normalising constraint on the coefficients, the second component has second greatest variance and so on. Often the first 5 principal components account for most of the variance within the data. Results are often presented in the form of bivariate plots based on the first few principal components in the hope of revealing interesting structure within the data.

It has been claimed that PCA is 'something of a blunt instrument' for detecting interesting structure since, for its success, it requires that large variation be interestingly structured (Jones and Sibson 1987, 2), an assumption which does not always hold in practice. Jones and Sibson (1987) go on to suggest that projection pursuit (PP) is a 'sharper tool' for the exploratory analysis of multivariate data. Whereas PCA seeks linear combinations of variables with maximal variance, PP seeks linear combinations of variables with maximal *interestingness*, as defined by an index of interestingness.

6.2 Background literature

The topic of projection pursuit has received much discussion in the literature over the last 15 years. Methodological discussions of PP have been available since Friedman and Tukey (1974). Further mathematical discussions are given in Huber (1985), Friedman (1987), Jones and Sibson (1987), Hall (1989), Cook *et al.* (1993), Sun (1991, 1993), Li and Cheng (1993), Eslava and Marriott (1994), Posse (1995a,b) and Nason (1995), with a useful overview being provided by Ripley (1996, 296-303).

Despite claims that applications of PP have 'flourished' (Posse, 1995a, 84) and have been 'promoted extensively in the literature and in implementation' (Nason, 1995, 413), published practical applications, as opposed to theoretical papers, are quite hard to find. Flenley and Olbricht (1993) and Wilhelm *et al.* (1999) are the only applications to archaeological data that that we are aware of, both of whom investigate the Oronsay particle data of Timmins (1981). Both make use of PP to investigate the possibility of a seaward shift of the beach-dune interface, with the latter making a comparison of high dimensional data display techniques. This data has undergone somewhat rigorous investigation in previous literature to determine if such a shift occurred.

Applications to data from other subject areas can be found in Friedman (1987), Jones and Sibson (1987), Nason (1995), Ripley (1996), Clements and Jones (1991),

Glover and Hopke (1992, 1994), Lendzionowski *et al.* (1990) and Walden (1994), which meet with varying degrees of success.

6.3 Implementing Projection Pursuit

A number of computer packages are freely available to implement projection pursuit, one of the most comprehensive is XGobi (Swayne *et al.*, 1991). XGobi implements 2-dimensional projection pursuit along with an array of other high dimensional visualisation tools and integrates with the S-Plus package (Venables and Ripley, 1999). We have used XGobi for the 2-dimensional examples presented in the following sections. XGobi can be obtained from

http://www.research.att.com/areas/stat/xgobi/

Westwood and Baxter (1999) list alternative software to implement projection pursuit and some newer packages more recently available include the following:

- FORTRAN software for Friedman's (1987) two-dimensional projection pursuit can be found at

  http://lib.stat.cmu.edu/general/projpurs

- A three-dimensional projection pursuit implementation from Nason (1995) can be found at

  http://www.stats.bris.ac.uk/~guy/Research/PP/PP.html

- FORTRAN code for two-dimensional PP from Jones and Sibson (1987) can be found at

  http://www.stats.bris.ac.uk/pub/software/pp2/mcj_pp.shar.gz

- The GGobi application has been developed as an updated version of XGobi and can be found here

  http://www.ggobi.org/

- IPP is a suite of S-Plus routines to implement projection pursuit

  `http://sun.cwru.edu/~jiayang/nsf/ipp.html`

- Commercial implementations of projection pursuit can be found in InforTEX's DataExplorer

  `http://www.itx.pl/`

6.4 Theory

The aim of projection pursuit is to find k linear functions (projections) of the original variables which show interesting structure when plotted. Typically $k = 1$ or $k = 2$ functions are sought which are easily visualised, although Nason (1995) discusses the case for $k = 3$. Many approaches begin by defining *uninteresting* structure as being normally distributed and aim to find projections which are as non-normal as possible, as measured by an *index of interestingness*, $I(\alpha)$, where α is a projection of the data in the direction $\alpha = \alpha_1, \alpha_2, \ldots, \alpha_d$.

We have found that there are no 'hard and fast' rules as to which measure of interestingness to use, and often find it most useful to repeat analyses using several different measures of interestingness. Each of the indices are designed to find projections of data which exhibit certain forms of structure, for example central mass or small interpoint distances. We have also found that under certain conditions, some indices tend to find projections of data with unique properties which are discussed below.

Our discussion below is limited to the case where α is a scalar, specifically 1-dimensional projection pursuit.

6.4.1 Measuring interestingness

The Friedman-Tukey Index

The Friedman-Tukey index (Friedman and Tukey, 1974) is composed of two parts

$$I_{FT}(\alpha) = s(\alpha)d(\alpha) \tag{6.1}$$

$s(\alpha)$ here is a 'spread' term, for example variance of the projection, and $d(\alpha)$ "captures" the density of the projection. Jones and Sibson (1987) note that the spread term is

used to "partially normalise the index against scale effects" which they accomplish by sphering data (section 6.4.2). In our examples in the following sections, we also sphere data prior to undertaking a projection pursuit analysis. Specifically we take principal components of scaled data and perform projection pursuit on this data, which has similar variances for each variable.

Jones and Sibson (1987) expanded the term d as a kernel density estimate

$$d(y) = \hat{I}_{FT}(y) = \frac{1}{n^2 h^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K\left(\frac{Y_j - Y_i}{h}\right) \tag{6.2}$$

which is an estimate of

$$I_{FT}(y) = \int_{\Re} f_y^2(y) dy = E(f_y(y)) \tag{6.3}$$

where $E(c)$ is the mean of c and y is the univariate projected data.

$I_{FT}(y)$ is minimised if y is spherical, by the parabolic density. Visually this appears similar to the normal density with smaller tails, thus departure from a parabolic density is also a departure from the standard normal density. The index was originally designed for use in a particle physics problem where they found that researchers preferred projections with very small interpoint distances while maintaining overall spread of the data. Our experience with the index is that it is quite effective at finding projections containing clusters, however it is one of the slowest indices to calculate, even with efficient approximations to kernel estimates (Ripley, 1996:298).

This index can easily be generalised to the multivariate case with the use of a multivariate kernel density estimate.

The Entropy Index

Huber (1985) used the fact that negative entropy is minimised by the multivariate normal distribution to develop the entropy index, which is defined as

$$I_E(y) = \int_{\Re} f_y(y) \log(f_y(y)) dy = E(\log(f_y(y))). \tag{6.4}$$

Again, this can be approximated using kernel density estimates

$$\hat{I}_E(y) = \frac{1}{n} \sum_{i=1}^{n} \log\left(\frac{1}{nh^2} \sum_{j=1}^{n} K\left(\frac{Y_j - Y_i}{h}\right)\right). \tag{6.5}$$

As with the Friedman-Tukey index, the entropy index is relatively slow to compute compared to the other indices.

89

The Legendre Index

Friedman (1987) developed the Legendre index in order to maximise the possibility of finding clusters in data. The index is motivated by first transforming normality to uniform $[-1, 1]$ using

$$z = 2\Phi(y) - 1$$

where $\Phi(y)$ is the standard normal cdf

$$\Phi(y) = (1/\sqrt{2\pi}) \int_{-\infty}^{y} e^{-1/2t^2} dt$$

such that z is uniformly distributed if y is normally distributed. The Legendre index is then defined as

$$I_L(y) = \int_{-1}^{1} (f_y(y) - 0.5)^2 dy. \tag{6.6}$$

This is then expanded using orthonormal series expansion, based upon Legendre polynomials. Cook *et al.* (1993) re-write the expanded Legendre index as

$$I_L(y) = \int_{\Re} (f_y(y) - \phi(y))^2 \frac{1}{2\phi(y)} dy \tag{6.7}$$

where $\phi(y)$ is the standard normal density, previously defined as the distribution which is un-interesting. This form of the index shows that the differences in the tails of the distribution are given more weight than differences in the centre of the distribution.

The Legendre index has a tendency to find skewed projections dominated by outliers. Ripley (1996:298) notes that the Legendre index "has the unfortunate effect of giving large weight to furcations in the density of f in its tails (where ϕ is small) and so will display sensitivity to outliers and the precise scaling used for the density".

Cook *et al.* (1993) note that the following indices are of a similar form to the Legendre index, specifically

$$I(y) = \int (f(y) - \phi(y))^2 w(y) \, dy \tag{6.8}$$

where $w(y)$ is a weight function which varies between the indices.

The Hermite Index

Hall (1989) was concerned about the Legendre index's tendency to find skewed projections so developed the Hermite index. Using Hermite polynomial series expansions,

the index approximates to

$$I_H(y) = \int_\Re (f_y(y) - \phi(y))^2 dy. \tag{6.9}$$

The Natural Hermite Index

Cook *et al.* (1993) develop the Natural Hermite index,

$$I_{NH}(y) = \int_\Re (f_y(y) - \phi(y))^2 \phi(y) dy. \tag{6.10}$$

Cook *et al.* (1993) constructed this new index to alleviate some of, what they describe as, the problems with the Legendre and Hermite indices. Their index returns to Friedman's original idea of giving more weight to the differences in the centre of the distribution. This has the effect of making the index less sensitive to outliers in the tails of the projection.

Other Indices

Simonoff (1996, 17) notes that it is in fact possible to use any reasonable test of normality as an index of interestingness for projection pursuit. Doing so results in the most non-normal projection being found, structure could be in the form of clusters, skewness or any other form of non-normality. As projection pursuit has developed, proposals have been made for numerous indices of interestingness which are designed to seek specific forms of structure.

6.4.2   Sphering

In PCA, data are typically standardised to eliminate scale effects which occur when variables are measured in different units or have widely differing variances. Similar effects are apparent in PP if the variance is not the same in all directions. Scale effects can be minimised by computing principal components and rescaling to zero mean and unit variance. As principal components are uncorrelated, the data are sphered.

Jones and Sibson (1987), Cook *et al.* (1993), Eslava and Marriott (1994) and Nason (1995) provide similar arguments for sphering data prior to performing projection pursuit. Eslava and Marriott (1994, 14) note that in addition to removing scale effects, sphering also results in projections fitting in roughly the same circle (with the

91

exception of outliers). This is convenient for graphical presentation as all projections can be shown on the same axis with no variable dominating any projection, as can sometimes happen with unscaled data. Reservations over sphering are given by Gower (discussion of Jones and Sibson 1987; 20). Equal variance in all directions implies that if there is symmetry about the centre, the result of sphering is a cloud of data points with spherical symmetry. However, it is also the case that structure will be distorted and will not have spherical symmetry (Eslava and Marriott 1994, 14).

In the examples in the following sections, data have been sphered (using the above procedure) prior to undertaking a projection pursuit analysis. Some analyses in archaeological applications routinely work with logarithmically transformed but unstandardised variables (for example Glascock, 1992). It would be equally possible to sphere such data, however this approach has not been adopted here.

6.5 Examples

The examples in the following section were originally written up by the author for inclusion in the published journal article Westwood and Baxter (2000). Technical aspects of the paper were written by Dr Baxter. The examples have been extended for inclusion in this thesis.

6.5.1 Example 1 - Lead isotope data

The following example, based on Baxter (1999), will be discussed in summary form only. Westwood and Baxter (2000) note that the technique discussed below has been presented routinely in the past, but not previously within the context of projection pursuit. In chapter 4 lead-isotope studies were discussed, in particular the assumption of normality (Sayre *et al.*, 1992) and the sample sizes which are required to demonstrate departures from normality were investigated. In section 4.3 we made use of the results to be presented below.

As discussed in chapter 4, lead isotope ratio data are three-dimensional and in their analysis it has sometimes been assumed that data from an ore source can be treated as a sample from a multivariate normal distribution (Sayre *et al.*, 1992). Recent work undertaken by Baxter and Gale (1998), Baxter (1999) and that discussed in chapter 4
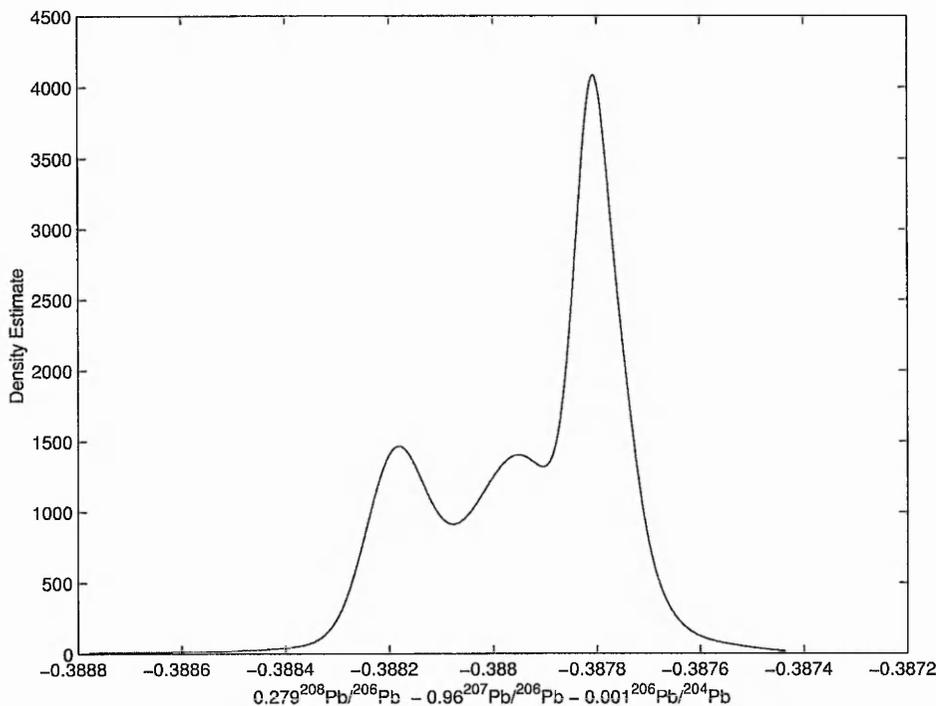
Figure 6.1: An adaptive kernel density estimate of the most non-normal linear combination of three-dimensional lead isotope ratio data for the Keos field (Stos-Gale *et al.* (1996). Window-width selected with h_{STE}).

has called into question the general validity of this assumption. The work presented in this thesis has extended that of Baxter and Gale (1998) and Baxter (1999). In particular, Baxter (1999) used a variety of tests of multivariate normality to demonstrate that many of the data sets in Stos-Gale *et al.* (1996) could not reasonably be regarded as samples from normally distributed data.

One test used was the multivariate extension of the Shapiro-Wilk test statistic for normality (Malkovich and Afifi, 1973). Malkovich and Afifi calculate the univariate Shapiro-Wilk statistic,

$$W = (\Sigma \alpha_i X_i)^2 / \Sigma (X_i - \bar{x})^2, \qquad (6.11)$$

for all possible univariate linear combinations of d variables, unlike projection pursuit they do not use an optimisation method and instead use 'brute force' methods. The most non-normal 'view' of the data is given by the linear combination of variables which results in the minimal value of W. This can be viewed as a form of projection pursuit with the Shapiro-Wilk statistic used as the index of interestingness, in the vein of the suggestion by Simonoff (1996) to use any test of normality as a measure of

93

interestingness. The result of applying this procedure to data is a global minimum of W.

In chapter 4 we used the above procedure to find a non-normal view of the Keos field (Stos-Gale *et al.*, 1996) which shows a clear departure from normality (figure 4.11), the procedure suggested the combination 0.1874^{208}Pb $/^{206}$Pb -0.9823^{207}Pb $/^{206}$Pb. A further example is shown in figure 6.1. Figure 6.1 shows a kernel density estimate of the most non-normal linear combination of all three ratios of the Keos field, 0.27889^{208}Pb $/^{206}$Pb -0.996029^{207}Pb $/^{206}$Pb -0.00876^{206}Pb $/^{204}$Pb.

Formal tests of normality suggest that this projection of the data is non-normal and this particular application of PP methodology suggests that the data are strongly multi-modal. Further illustrations of this kind of use can be found in Baxter and Gale (1998) and Baxter (1999).

6.5.2 Example 2 - Blue soda glass from York

Cox and Gillies (1986) published analyses of blue soda glass from the windows of York Minster and archaeological excavations. These specimens have been used elsewhere to illustrate a variety of methodologies (Baxter, 1989; Baxter and Buck, 1999; Bell and Croson, 1998). There are 27 specimens, measured with respect to the concentration of 12 oxides and elements. Most analyses clearly show three main groups in the data, with some analyses suggesting possible sub-groups or outliers.

Figure 6.2 shows four analyses of the data. The PCA analysis (of standardised data) in figure 6.2a shows three clusters, two of the groups are very tightly defined with the third (towards the top of the figure) more dispersed relative to the other two. This structure is readily found using projection pursuit, we used the Friedman-Tukey index to find the projection in figure 6.2b. The view found by projection pursuit shows the same three clusters with the third cluster being less dispersed in this plot.

The view illustrated in figure 6.2c, in which the structure is 'circular', occurs quite commonly in our experience with similar data sets, and has no useful practical interpretation. Similar examples can be found in Cook *et al.* (1993, 248) and Ripley (1996, 302). Figure 6.2d shows a further example of a problem encountered with projection pursuit, in this example the plot is dominated by outliers towards the top of the plot.
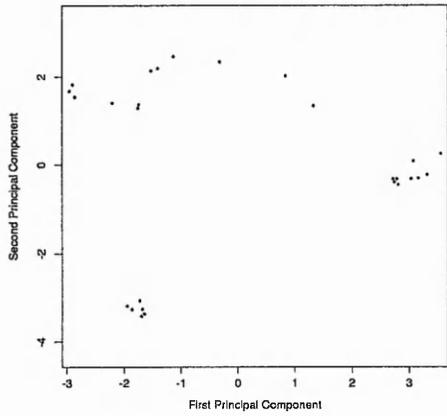
(a) PCA of standardised data
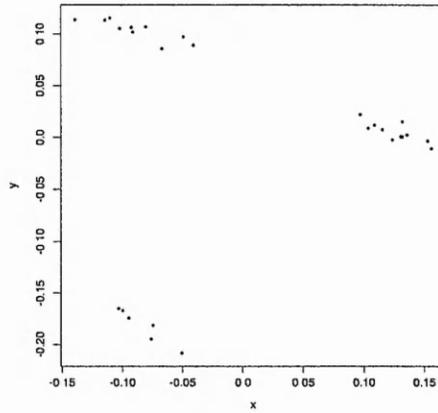
(b) PP using Friedman-Tukey index
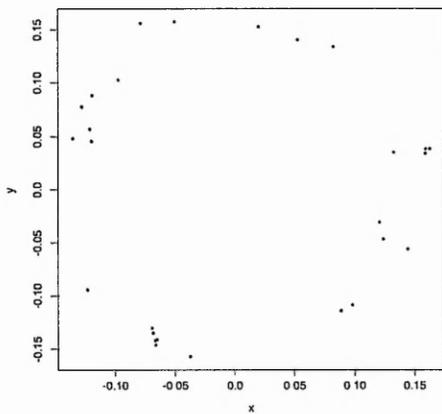
(c) Circular PP view found using the
Natural Hermite index
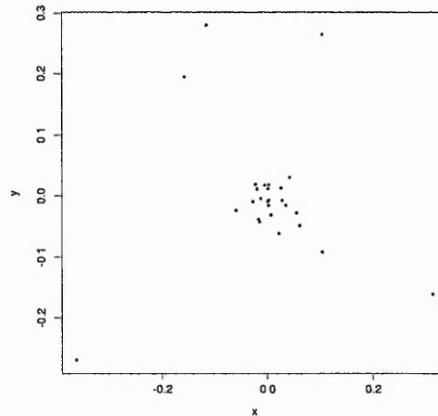
(d) PP using Natural Hermite index

Figure 6.2: Bivariate plots of York Minster data of Cox and Gillies (1986). Plots show the first two principal components of standardised data (a) and 'views' of the multivariate data suggested by projection pursuit (b-d).

(a) PCA of standardised data   (b) PP using Legendre index
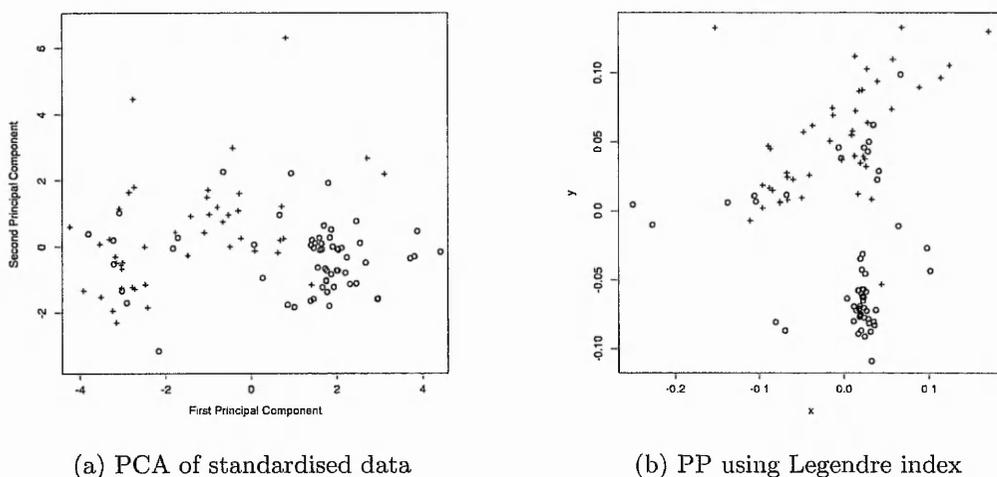
Figure 6.3: Plots showing the analyses of the waste glass compositions from Mancetter and Leicester (Jackson, 1992). Labelling is by site with "+" cases from Mancetter and "o" cases from Leicester.

Both of these views were suggested by the Natural Hermite index.

For this data set the structure is fairly obvious and found almost 'instantaneously' by projection pursuit. Other useful views (suggesting any other form of structure) were not found in the course of exploration using PP, although a number similar to those in figure 6.2c and d were found. Although the PP view in figure 6.2b is 'sharper' than the PCA view, it tells essentially the same story.

6.5.3 Example 3 - Waste glass from Leicester and Mancetter

The data used in this example consist of 105 specimens of waste glass found on furnace sites at Leicester and Mancetter and measured with respect to the concentration of 11 major and minor oxides. It is of interest to see if there are distinct chemical groups in the data, and if these correspond to the furnace sites. The data were collected and published by Jackson (1992) and are reproduced in Baxter (1994) where extensive analysis was undertaken using a variety of multivariate methods. These analyses suggest three concentrations in the data with some correspondence – by no means exact – to the furnace groups.

This is shown in the PCA plot in figure 6.3a, where labelling is by site. Without a knowledge of the sites it is possible, visually or with the aid of techniques such as

kernel density estimation (Baxter *et al.*, 1997) to detect three main concentrations in the data. There are no obviously distinct clusters. The densest concentration to the right consists mainly of glass from Leicester; the other two concentrations contain most of the Mancetter specimens, with 11 to 14 Leicester specimens mixed in (depending on how boundaries of concentrations are visualised).

The PP view in figure 6.3b quite clearly isolates a cluster of cases in the bottom half of the plot consisting, with one exception, of Leicester specimens. The remaining dispersed group, possibly sub-dividing into two, contains the Mancetter specimens with the same number of Leicester specimens mixed in as in the PCA. The cluster towards the bottom of figure 6.3b has observations with higher concentrations of Na, Mn and P than other samples, this is the chemical relationship PP has used to find this particular projection.

Arguably the PCA and PP analyses lead to similar conclusions, but the separation between material from the two sites, and the fact that it is less than perfect, is clearer in the latter analysis because of the clearer clustering revealed. We remark that we can be confident that PP is not revealing spurious structure in this case because information not used in the PP, concerning site of origin, allows us to interpret the revealed structure in a useful archaeological manner.

6.5.4   Example 4 - Oriental Greenwares

This example is based on a 133×9 data set published by Pollard and Hatcher (1986) showing the chemical composition of 133 oriental greenwares which are suspected to have originated from several areas of manufacture. We follow them in omitting three clear outliers and one variable, SiO_2, in our analysis.

There are two very obvious chemical groups in the data, as the PCA in figure 6.4a shows. The group to the left is associated with Northern Zhejiang Yue wares and that to the right with Longquan celadons. It is easy to get the same separation using PP and one such view, found with the Friedman-Tukey index, is shown in figure 6.4b. This additionally suggests a small group at the bottom of the plot that is a subset of the earlier wares, but we have been unable to interpret this as archaeologically distinct in any way.
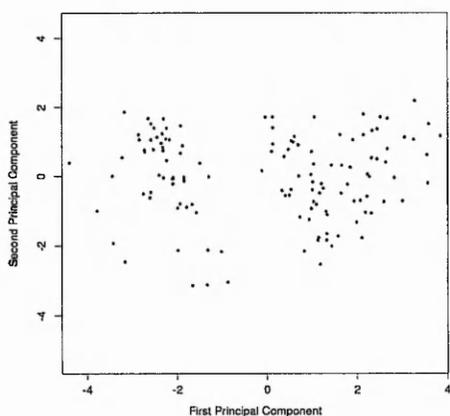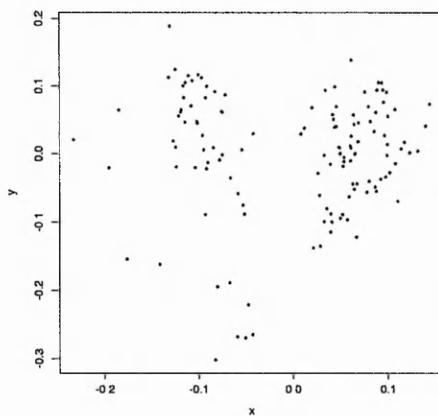
(a) PCA of standardised data   (b) PP using Friedman-Tukey index

Figure 6.4: Plots of the oriental greenware compositions of Pollard and Hatcher (1986) showing (a) the first two principal components and (b) view of the data suggested by projection pursuit.

Pollard and Hatcher (1986) applied cluster analysis to the 53 specimens in the earlier group and concluded there were three subgroups. After applying stepwise discriminant analysis to these, five outliers were removed and a discriminant analysis plot for the remaining 48 cases was shown on page 268 of their paper. A similar analysis is shown in figure 6.5a, the only difference being our use of all eight variables rather than the five selected in the original publication. Interpretation of the groups is not absolutely clear-cut, but they can be associated with regional differences in composition. Given a knowledge of this classification we have been unable to obtain a PP view that separates out the groups as well as the discriminant analysis. In figure 6.5b one PP view for the 48 cases is shown which separates out the smaller group but not the two larger ones. It may be noted that a PCA analysis of this subset (figure 6.5c) did as well as the PP in separating the groups.

Our PP analysis of this data set cannot be regarded as especially successful. Although the PP for the full data set did suggest structure additional to that revealed by PCA we were unable to interpret the results in an archaeologically useful fashion, so have no real way of determining whether the structure is spurious or not.

Similar remarks apply to PP analyses of the other subset to that used in figure 6.5 that are not shown here. For the subsets of data examined, projection pursuit was

98
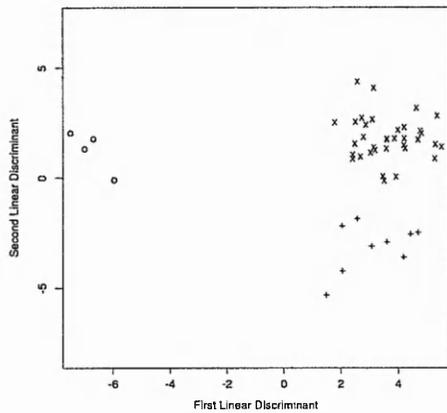
(a) Discriminant analysis plot for the smaller group from figure 6.4(a)
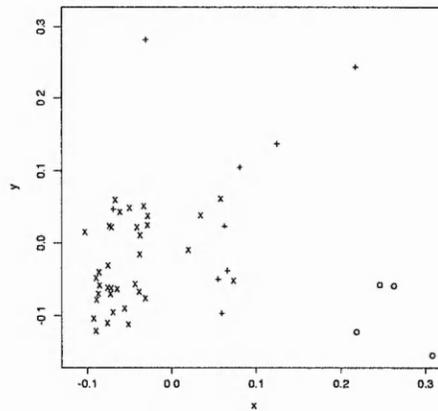
(b) PP using Legendre index for the smaller group from figure 6.4(a)
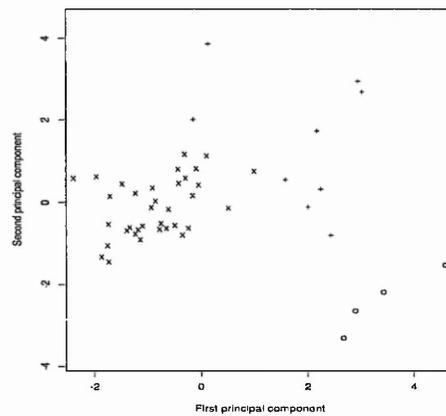
(c) PCA of standardised data for the smaller group from figure 6.4(a)

Figure 6.5: Plots based on analyses of the oriental greenware compositions. Figures use data from the left-hand group showing in figure 6.4a, after omitting five outliers, labelled according to the groups determined by Pollard and Hatcher (1986).

99

able to produce similar clusters to those suggested by principal components analysis, however neither was able to suggest the structure Pollard and Hatcher (1986) found using cluster analysis and discriminant analysis.

6.6   Discussion

For the specialised problems discussed in example 1, the application of one-dimensional projection pursuit, as illustrated, has a potentially useful role to play. In section 4.3 the use of 1D-PP was combined with tests of normality and mode counting to form a powerful tool to investigate structure in data.

The examples cited above show that two-dimensional PP can produce a sharper view of structure in data than that provided by PCA, however it was also the case that PP did not lead to an interpretation different from that achieved with PCA. We have used additional data to compare the views of data found by both methods, for example we make use of information relating to the known origin of a sample or the date of manufacture. If we can relate this archaeological knowledge to clusters in the projection, and make sense of this, we view the projection as a successful one.

In example 4 the PP analysis did not lead to any new insights into the data; although some additional structure was suggested, there is no obvious way of determining if this has an archaeological meaning or whether it is spurious. There is a tendency for some 'optimal' projections to be dominated by outliers, as shown in figure 6.2d. These types of projection generally have no obvious meaning and the outlying observations are not suggested by other multivariate techniques. We have also seen that PP can mislead if the sample sizes are small, as suggested by the circular plots in figure 6.2c. This is discussed further below.

Our original motivation for investigating PP was the suggestion of Jones and Sibson (1987) that PP is possibly a sharper method of identifying structure than the commonly used PCA. We had hoped that if PP had been 'sharper' it might be better able to identify structure with smaller sample sizes (compared to PCA), however it turned out that this was not the case. We have seen that PP is very good at detecting spurious structure in multivariate data (as have Cook *et al.* 1993:248 and Ripley, 1996:302) however we have not seen examples of PP providing anything beyond that which can

be found by examining the first few principal components.

In the wider literature there undoubtedly exist examples where PP does produce informative views of the data that PCA does not reveal. This sometimes occurs when the structure in the data is 'unusual' (see, for example, the structures used in Posse's (1995a, 91) simulation study), and of a kind that we suspect would often be regarded as uninterpretable in the context of the type of data used here. The model, often implicit, in studies that produce data similar to those used in examples 2, 3 and 4 is that the data may be viewed as a sample from a mixture of distributions which, in those studies that make statistical assumptions, are multivariate normal. In d-dimensional space the expectation is either that there will be distinct point clouds, or that there will be overlapping point clouds with distinct high-density regions. We suspect that methods such as PCA or cluster analysis will often be adequate to detect this, and that a PP view showing a marked departure from the underlying model might be difficult to interpret (we also recognise that this is not a good argument for not using PP).

Where PP has been contrasted with PCA and judged to be superior (e.g. Glover and Hopke, 1992) the judgement is sometimes a fine one. It is also the case that in order to select a PP view and judge that it is superior to PCA it may be necessary to use additional information (e.g. a prior classification of the data) to confirm that the PP view is a useful one. Given the ease with which PP can suggest spurious structure with 'small' data sets we have found it very difficult to interpret results where such prior knowledge has not been available. The superiority of PP compared with PCA has sometimes been exaggerated. Posse (1995a, 83-84) analyses data on five measurements for 200 Australian crabs, most belonging to four groups. He claims that PP is able to reveal a 'clustered projection' that was 'not found by principal component analysis'. In fact the first component has an obvious size interpretation, and any of several standard approaches to PCA that aim to remove size effects (including a plot of the second and third components) will reveal a clustered projection similar to that found by PP.

Our experiences suggest that the use of PP as a tool for the *routine* analysis of data of the kind discussed in examples 2-4 is not always practical. In addition to the reasons discussed so far, there are two pragmatic reasons that give rise to this concern. The first concerns the size of the data sets typically available. Most examples of two-

dimensional PP that we have seen use $d < 10$; our examples used $d = 8$, 11 and 12; as has been seen in the review of archaeometric analyses in chapter 2, it is now quite common to see analyses based on data sets for which $d > 20$. There has not been a commensurate increase in the number of samples typically collected, so that $n < 100$ is quite usual. In the context of a technique that can easily suggest spurious structure in small data sets, and where 100×5 is considered to be small (Cook *et al.*, 1993), many archaeometric data sets are small and subject to the problems that this entails.

The second reason concerns the time required to carry out PP. A large number of local optima arise in analysis, and the views they are associated with need to be inspected to see if they are 'interesting' and have a useful archaeological interpretation. In XGobi the plots produced in the course of pursuit can be viewed in real time and visually 'interesting' projections, including some used here, do not necessarily even correspond to local optima. These also need to be assessed and this is very demanding of time and has not, in the many analyses that we have undertaken, led to much extra insight into the data being gained, beyond that provided by PCA and cluster analysis – the tools most usually deployed in the literature.

The time required to undertake a projection pursuit is extended because of the need to undertake pursuits with a number of indices. Although each index was designed to detect a particular form of structure, for example the Friedman-Tukey index was designed to detect projections with very small inter-point distances where as the Legendre index was designed to maximise the possibility of finding clusters, we do not know in advance which index will provide the most useful results. For the examples cited above, the Friedman-Tukey index has tended to find projections which show the most archaeologically interesting view of data, however we have applied each index to each data set as each index has its own merits.

Our experience with the indices is that optimal projections suggested by the Friedman-Tukey index tend to find clusters in the data (be they meaningful or not) relatively quickly. Although it has been noted that the Legendre index has a tendency to find projections dominated by outliers or skewness (Ripley, 1996), we have not found this to be a major drawback of the index, although it is very slow to seek optimal projections. The entropy index is also very slow to calculate, we have found the other indices

are able to find similar projections to those found by Entropy. The Natural Hermite index was developed by Cook *et al.* (1993) to, as they describe, alleviate some of the problems with the Legendre index. Our experience of this index is that it tends to be more dominated by outliers than any of the others. Given that each index was designed to be optimal under slightly different circumstances, we have found it advisable to undertake a projection pursuit with at least two of the indices, we prefer to make use of the Friedman-Tukey and Legendre indices.

Given the reasons discussed above, we do not view PP, in its current state of development, as a tool we would recommend for routine archaeometric data analysis. We have have shown that PP is at least able to produce results comparable to PCA, however we have not seen an example in which PP shows something 'new'.

Chapter 7

Conclusions

The thesis investigates a number of sample size related issues which arise in the use of multivariate statistics applied to compositional data. Our ultimate goal was to give some form of guidance on the types of sample size which should be used for multivariate analysis, however this depends on the unknown structure of the population from which a sample originates. As such our conclusions are based on specific areas of archaeometry.

In chapter 2, a review was undertaken of the typical sample sizes currently used in published articles in the journal *Archaeometry* between 1975 and 1999. The advent of new methods of chemical analysis, such as ICP-OES, have increased the number of variables available for statistical analysis. Sample sizes available have generally increased at a similar level, as suggested by the n/d ratio throughout the 25 year period. The review highlighted the fact that it is uncommon to see an analysis of less than 8 variables, however as many as 20 variables have been used in statistical analyses in published articles, with sample sizes typically between 30 and 100. It is apparent that the use of multivariate statistical analysis has increased in popularity. We suspect this is due to the availability of larger sample sizes and the increased number of variables accessible through the use of newer methods of chemical analysis. The increased availability of accessible tools to perform multivariate statistical analysis has also aided the popularity of multivariate statistical analysis.

The assumption of normality of trivariate lead-isotope fields, and the sample sizes required to detect departures from normality are addressed in chapter 4. Our approach was to examine the univariate, bivariate and trivariate cases in isolation and determine sample size requirements to detect structure in both simulated and real data using

direct testing of normality and mode counting techniques. The commonly accepted minimum number of samples to define a lead-isotope field, $n = 20$ (Pollard and Heron, 1996), is questioned. We demonstrate that this is inadequate to detect departures from normality and that in some cases a minimum of 40 samples is required to detect quite clear departures from normality. We also suggest that the normality of lead-isotope fields cannot be assumed, as is required for the construction of confidence ellipsoids and in probability calculations which are sometimes used in lead-isotope ratio studies.

The method adopted for mode counting in chapter 4 was extremely sensitive to spurious structure. In section 4.2.2 we discuss possible modifications to the process we use. An alternative approach, investigated in chapter 5, is to use more formal methods of mode counting. Silverman (1981) proposed a test of univariate modality based on the use of non-adaptive kernel density estimates. We make use of this test to assess the modality of a number of data sets including lead-isotope data. Izenman and Sommer (1988) suggest that Silverman's test should be extended to make use of adaptive KDEs. Our work on univariate lead-isotope ratios confirmed that adaptive KDEs provide more accurate density estimates than non-adaptive KDEs. As such we proposed a version of the test making use of adaptive KDEs and investigated its performance in relation to the original test of Silverman (1981). Results suggest that the adaptive version of the test shows a greater difference between results for a correct and incorrect hypothesis which should allow for easier interpretation of results.

Techniques such as principal components analysis allow multivariate data to be reduced to lower dimensional views (typically 1-, 2- or 3-dimensional views) which allow for direct investigation into structure present in data, often using graphical methods such as plots of the first few principal components. It is possible for the techniques discussed in chapters 4 and 5 to be applied to multivariate data which have been projected to a lower number of dimensions using such techniques.

Jones and Sibson (1987) describe principal components analysis as 'something of a blunt instrument' for the analysis of structure in multivariate data, they make the suggestion that projection pursuit is a 'sharper' tool. In chapter 6 we investigated this claim and the possibility of using projection pursuit to find low dimensional views of data given the relatively small sample sizes available. We note that projection pursuit

is good at locating spurious structure with the number of samples typically available. We conclude that the routine use of projection pursuit in the analysis of archaeometric data is not advisable due to the amount of time taken to review 'interesting' views of the data found by projection pursuit, and the frequency with which spurious structure is identified.

7.1   Further Work

An obvious extension to the work presented is discussed in section 4.2.3. Our approach to mode counting has raised a number of issues and has lead to the conclusion that formal methods of testing of modality should be investigated. A number of tests have been identified, including that of Silverman (1981) which we discuss and extend in chapter 5. Others include Excess Mass methods and the DIP, RUNT and MAP tests. Mixture modelling approaches may also be of use. These work by determining components which can be used to simulate a population with similar structure.

Despite the number of tests proposed in the literature, there are few published applications of their use. A comparative study of these techniques could be undertaken. This could be in the form of extending the work on lead-isotope data in chapter 4. For example investigating the performance of tests when applied to simulated data with a known structure and subsequently applied to real data.

The work to date has focused primarily on working with low dimensional data, this is either through working directly with univariate, bivariate or trivariate data or by using principal components analysis or projection pursuit to project the data into a lower number of dimensions. An obvious drawback of projecting data into a lower number of dimensions is the possibility of losing, or hiding, structure present in other dimensions. As such, working with the multivariate data directly is of interest.

The ISE statistic (Bowman and Foster, 1993) discussed in 4.4.1 can be directly applied to multivariate data to test for multivariate normality, as can the extension of the Shapiro-Wilk statistic of Malkovich and Afifi (1973). Further measures of multivariate structure are presented by Huffer and Park (2000) and could form the basis for a simulation study of higher dimensional data and a review of measures of multivariate structure. The review of published analyses presented in chapter 2 could form an initial

point from which to research varying forms of structure present in compositional data and be used to investigate sample size requirements. In a sense, building a catalogue of the forms of structure commonly found in archaeometry.

.

References

[1] Baxter, M.J. (1989) Multivariate analysis of data on glass compositions: a methodological note. *Archaeometry*, **31**, 45-53.

[2] Baxter, M.J. (1994) *Exploratory Multivariate Analysis in Archaeology*, Edinburgh: Edinburgh University Press.

[3] Baxter, M.J. (1999) On the multivariate normality of lead isotope fields. *Journal of Archaeological Science*, **26**, 117-124.

[4] Beardah, C.C. and Baxter, M.J. (1995) MATLAB Routines for Kernel Density Estimation and the Graphical Presentation of Archaeological Data, Analecta Prehistorica Leidensia 28, Interfacing the Past, Computer Applications and Quantitative Methods in Archaeology eds H. Kammermans and K. Fennema, Leiden.

[5] Baxter, M.J., Beardah, C.C. and Wright, R.V.S. (1997) Kernel density estimation for archaeology. *Journal of Archaeological Science*, **24**, 347-354.

[6] Baxter, M.J., Beardah, C.C. and Westwood, S. (2000) Sample size and related issues in the analysis of lead isotope data. *Journal of Archaeological Sciences*, 973-980.

[7] Baxter, M.J. and Gale, N.H. (1998) Testing for multivariate normality via univariate tests: a case study using lead isotope ratio data. *Journal of Applied Statistics*, **25**, 679-691.

[8] Baxter, M.J. and Buck, C.E. (1999) Data handling and statistical analysis. In (E Ciliberto and G Spoto, eds.) *Modern Analytical Methods in Art and Archaeology*, Wiley: New York.

[9] Bell, S. and Croson, C. (1998) Artificial neural networks as a tool for archaeological data analysis. *Archaeometry*, **40**, 139-151.

[10] Bowman, A.W. (1992) Density based tests for goodness-of-fit. *Journal of Statistical Computation and Simulation*, **40**, 1-13.

[11] Bowman A.W. and Azzalini A. (1997) *Applied Smoothing Techniques for Data Analysis - The Kernel Approach with S-Plus Illustrations* Oxford : Clarendom Press.

[12] Bowman, A.W. and Foster, P.J. (1993) Adaptive smoothing and density-based tests of multivariate normality. *Journal of the American Statistical Association*, **422**, 529-537.

[13] Breiman, L., Mesiel, W. and Purcell, E. (1977) Variable kernel estimates of multivariate densities. *Technometrics*, **19**, 135-144.

[14] Brill, R.H. and Wampler, J.M. (1967) Isotope studies of ancient lead. *American Journal of Archeology*, **71**, 63-77.

[15] Budd, P., Gale, D., Pollard, A.M., Thomas, R.G. and Williams, P.A. (1993) Evaluating lead isotope data: further observations. *Archaeometry*, **35**, 241-265.

[16] Budd, P., Pollard A.M., Scaife, B and Thomas R.G. (1995). Oxide ingots, recycling and the Mediterranean metal trade. *Journal of Mediterranean Archaeology*, **8**, 1-32.

[17] Budd, P., Haggerty, R., Pollard A.M., Scaife, B and Thomas R.G. (1996). Rethinking the quest for provenance. *Antiquity*, **70**, 168-174.

[18] Clements, A.M. and Jones, M.C. (1991) An ecological example of the application of projection pursuit to compositional data. *Vegetatio*, **95**, 101-117.

[19] Cook, D., Buja, A. and Cabrera, J. (1993) Projection pursuit indexes based on orthonormal function expansions. *Journal of Computational and Graphical Statistics*, **2**, 225-250.

[20] Cox, G.A. and Gillies, K.J.S. (1986) The X-ray fluorescence analysis of medieval durable blue soda glass from York Minster. *Archaeometry*, **28**, 57-68.

[21] Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap.* London: Chapman and Hall.

[22] Efron, B. and Tibshirani, R.J. (1997) The problem of regions. Technical report, 1997, Stanford University.

[23] Eslava, G. and Marriott, F.H.C. (1994) Some criteria for projection pursuit. *Statistics and Computing*, **4**, 13-20.

[24] Everitt, B.S. and Hand, D.J. (1981) *Finite Mixture Distributions*. London: Chapman and Hall.

[25] Flenley, E. C., and Olbricht, W. (1993) Classification of archaeological sands by particle size analysis. In (O. Opitz, B. Lausen, and R. Klar, eds.), *Information and Classification. Concepts, Methods and Applications. Proceedings of the 16th Annual Conference of the Gesellschaft fr Klassifikation e. V.*, Springer, Berlin, Heidelberg, 478-489.

[26] Foy D. (1985) Essai de typologie des verres medievaux d'aprs les fouilles provenales et languedociennes. *Journal of Glass Studies*, **27**, 18-71.

[27] Friedman, J.H. (1987) Exploratory projection pursuit. *Journal of the American Statistical Association*, **82**, 249-66.

[28] Friedman, J.H. and Tukey, J.W. (1974) A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, **23**, 881-889.

[29] Gale, N.H. and Stos-Gale, Z.A. (1992) Evaluating lead isotope data: comments on E.V. Sayre, K.A. Yener, E.C. Joel and I.L. Barnes, 'Statistical evaluation of the presently accumulated lead isotope data from Anatolia and surrounding regions', Archaeometry **34**(1)(1992), 73-105, and reply. *Archaeometry*, **34**, 311-336.

[30] Gale, N.H. and Stos-Gale, Z.A. (1993) Evaluation lead-isotope data - further observations - comments II. *Archaeometry*, **35**, 252-259.

[31] Gale, N.H., Stos-Gale, Z.A., Maliotis, G. and Annetts, N. (1997) Lead isotope data from the Isotrace Laboratory, Oxford: *Archaeometry* data base 4, ores from Cyrpus. *Archaeometry*, **39**, 237-246.

[32] Glascock, M.D. (1992) Characterization of archaeological ceramics at MURR by neutron activation analysis and multivariate statistics. In (Neff H., ed.) *Chemical Characterization of Ceramic Pastes in Archaeology*, Madison, Wisconsin: Prehistory Press, 11-26.

[33] Glover D.M. and Hopke P.K. (1992) Exploration of multivariate chemical data by projection pursuit. *Chemometrics and Intelligent Laboratory Systems*, **16**, 45-59.

[34] Glover, D.M. and Hopke, P.K. (1994) Exploration of multivariate atmospheric particulate compositional data by projection pursuit. *Atmospheric Environment*, **28**, 1411-1424.

[35] Good, I.J. and Gaskins, R.A. (1980) Density estimation and bump-hunting by penalized likelihood method exemplified by scattering and meteorite data. *Journal of the American Statistical Association*, **75**, 42-56.

[36] Hall, P. (1989) Polynomial projection pursuit. *The Annals of Statistics*, **17**, 589-605.

[37] Harbottle, G. (1976) Activation analysis in archaeology. *Radiochemistry*, **3**, 3372.

[38] Hartigan, J.A. and Hartigan, P.M. (1985) The DIP test of unimodality. *Annals of Statistics*, **13**, 70-84.

[39] Hartigan, J.A. and Mohanty, S. (1992) The RUNT test for multimodality. *Journal of Classification*, **9**, 63-70.

[40] Huber, P.J. (1985) Projection pursuit (with discussion), *Annals of Statistics*, **13**, 435-525.

[41] Huffer, F.W. and Park C. (2000) A test for multivariate structure. *Journal of the American Statistical Association*, **27**, 633-350.

[42] Hunter, J.R. and Heyworth, M.P. (1998) *The Hamwic Glass: CBA Research Report 116*. York: Council for British Archaeology.

[43] Izenman, A.J. and Sommer, C.J. (1988) Philatelic mixtures and multi-modal densities. *Journal of the American Statistical Association*, **83**, 941-953.

[44] Jackson, C.M. (1992) *A Compositional Analysis of Roman and Early Post-Roman Glass and Glass-working Waste from Selected British Sites*, unpublished Ph.D thesis, University of Bradford, UK.

[45] Jones, M.C., and Sibson, R. (1987) What is projection pursuit? (with discussion) *Journal of the Royal Statistical Society A*, **150**, 1-36.

[46] Lendzionowski V., Walden A.T. and White R.E. (1990) Seismic character mapping over resevoir intervals. *Geographical Prospecting*, **38**, 951-969.

[47] Li, G.Y. and Cheng, P. (1993) Some recent developments in projection pursuit in China. *Statistica Sinica*, **3**, 35-51.

[48] Malkovich, J.F. and Afifi, A.A. (1973) On tests for multivariate normality *Journal of the American Statistical Association*, **68**, 176-179.

[49] Manly, B.F.J. (1994) *Multivariate Statistical Methods: A Primer*. London: Chapman and Hall.

[50] Mardia, K.V. (1987) Discussion of 'What is projection pursuit?' by Jones and Sibson (1987). *Journal of the Royal Statistical Society A*, **150**, 22-23.

[51] McLachlan, G.J., Peel, D., Basford, K.E., and Adams, P. (1999) The EMMIX software for the fitting of mixtures of normal and t-components. *Journal of Statistical Software* **4**, No. 2.

[52] Muller, D.W. and Sawitzki, G. (1991) Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association*, **86** 738-746.

[53] Nason, G.P. (1995) Three-dimensional projection pursuit. *Applied Statistics*, **44**, 411-430.

[54] Pollard, A.M. and Hatcher, H. (1986) The chemical analysis of oriental ceramic body compositions: part 2 - greenwares. *Journal of Archaeological Science*, **13**, 261-287.

[55] Pollard, A.M. and Heron, C. (1996) *Archaeological Chemistry*. Cambridge: Royal Society of Chemistry.

[56] Posse, C. (1995a) Tools for two-dimensional exploratory projection pursuit. *Journal of Computational and Graphical Statistics*, **4**, 83-100.

[57] Posse, C. (1995b) Projection pursuit exploratory data analysis. *Computational Statistics and Data Analysis*, **20**, 669-687.

[58] Reedy, T.J. and Reedy, C.L. (1992) Evaluating lead isotope data: comments on E.V. Sayre, K.A. Yener, E.C. Joel and I.L. Barnes, I.L. 'Statistical evaluation of the presently accumulated lead isotope data from Anatolia and surrounding regions'. *Archaeometry*, **34 (1)** (1992), 73-105, and reply. Comments IV *Archaeometry*, **34**, 327-329.

[59] Ripley, B.D. (1996) *Pattern Recognition and Neural Networks*, Cambridge: Cambridge University Press.

[60] Rozal, G.P.M. and Hartigan, J.A. (1994) The MAP test for multimodality. *Journal of Classification*, **11**, 5-36.

[61] Sayre, E.V., Yener, K.A., Joel, E.C. and Barnes, I.L. (1992) Statistical evaluation of the presently accumulated lead isotope data from Anatolia and surrounding regions. *Archaeometry*, **34**, 73-105.

[62] Sayre, E. V., Yener, K. A. and Joel, E. C. (1995). Comments on "Oxhide ingots, recycling and the Mediterranean metal trade". *Journal of Mediterranean Archaeology*, **8**, 4553.

[63] Sayre, E.V., Joel, E.C., Blackman, M.J., Yener, K.A., Ozbal, H. (2001) Stable lead isotope studies of Black Sea Anatolian ore sources and related Bronze Age and Phrygian artefacts from nearby archaeological sites. *Archaeometry*, **43**, 77-115.

[64] Scaife, B., Budd, P., McDonnell, J.G., Pollard, A.M. and Thomas, R.G. (1996) A reappraisal of statistical techniques used in lead isotope analysis in (Demirci S., Ozer A.M. and Summers G.D., eds.) *Archaeometry 94: Proceedings of the 29th International Symposium on Archaeometry, Ankara 9-14 May 1994*, Ankarac Tubitak, 301-307.

[65] Scott, D.W. (1992) *Multivariate Density Estimation - Theory, Practice and Visualization*. New York:Wiley and Sons.

[66] Sheather, J.S. and Jones, M.C. (1991) A reliable data-based bandwidth selection method for kernel density estimates. *Journal of the Royal Statistical Society B*, **53**, 683-90.

[67] Silverman, B.W. (1981) Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society B*, **43**, 97-99.

[68] Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

[69] Simonoff, J.S. (1996) *Smoothing Methods in Statistics*, New York: Springer.

[70] Stos-Gale, Z.A. and Gale, N.H. (1994) Metals in (Knapp A.B. and Chery J.F.) *Provenience Studies and Bronze Age Cyprus*, Madison: Prehistory Press, 92-121.

[71] Stos-Gale, Z.A., Gale, N.H. and Annetts, N. (1996) Lead isotope data from the Isotrace Laboratory, Oxford: *Archaeometry* data base 3, ores from the Aegean, Part 1. *Archaeometry*, **38**, 381-390.

[72] 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**, 82-123.

[73] Sun J. (1991) Significance levels in exploratory projection pursuit. *Biometrika*, **78**, 759-769.

[74] Sun J. (1993) Some practical aspects of exploratory projection pursuit. *SIAM Journal on Scientific Computing*, **14**, 68-80.

[75] Swayne, D.F., Cook, D. and Buja, A. (1991) User's manual for XGobi, a dynamic graphics program for data analysis implemented in the X Windows system (Release 2). Bellcore Technical Memorandum.

[76] Rozal, G.P.M. and Hartigan, J.A. (1994) The MAP test for multimodality. *Journal of Classification*, **11**, 5-36.

[77] Timmins, D.A.Y. (1981) *Study of Sediment in Mesolithic Middens on Oronsay*, MA Thesis, University of Sheffield, Sheffield, UK.

[78] Venables, W.N. and Ripley, B.D. (1999) *Modern Applied Statistics with S-Plus: Third Edition*, Springer: New York.

[79] Walden, A.T. (1994) Spatial clustering: using simple summaries of seismic data to find the edge of an oil-field. *Applied Statistics*, **43**, 385-398.

[80] Wand, M.P. and Jones, M.C. (1993) Comparison of smoothing parameterizations in bivariate kernel density estimates. *Journal of the American Statistical Association*, **422**, 520-528.

[81] Wand, M.P. and Jones, M.C. (1995) *Kernel Smoothing*. London: Chapman and Hall.

[82] Wilhelm, A.F.X., Wegman, E.J. and Symanzik, J. (1999) Visual Clustering and Classification: The Oronsay Particle Size Data Set Revisited. *Computational Statistics*, **14**, 109-146.

[83] Westwood S., Baxter M.J. and Beardah C. (1998) Sample size problems with archaeometric data. Computer Applications and Quantitative Methods in Archaeology 1998 (eds. J. Barcelo, I. Briz and A. Vila), 155-7 (figures on CD-ROM). BAR International Series 757, Oxford: Archaeopress (1999).

[84] Westwood S. and Baxter, M.J. (2000) Exploring archaeometric data using projection pursuit methodology, UK Chapter of Computer Applications and Quantitative Methods in Archaeology 1999 (eds. C. Buck et al.), 81-90. BAR International Series 844, Oxford: Archaeopress (2000).

References - Review of Statistical Analysis in Archaeometry

[85] Adan-Bayewitz, D., Asaro, F. and Giauque, R.D. (1999), Determining pottery provenance: application of a new high-precision x-ray fluorescence method and comparison with instrumental neutron activation analysis, *Archaeometry*, **41**, 1-24.

[86] Adriaens, A., Veny, P., Adams, F., Sporken, R., Louette, P., Earl, B., Özbal, H. and Yener, K.A. (1999), Analytical investigation of archaeological powders from Göltepe (Turkey), *Archaeometry*, **41**, 81-89.

[87] Alvey, R.C. and Laxton, R.R. (1978), A note on the chemical analysis of some Nottingham clay tobacco pipes, *Archaeometry*, **20**, 189-196.

[88] Argyropoulous, V. (1995), A characterization of the compositional variations of Roman Samian pottery manufactured at the Lezaux production centre, *Archaeometry*, **37**, 271-285.

[89] Attas, M., Yaffe, L. and Fossey, J.M. (1977), Neutron activation analysis of Early Bronze Age pottery from Lake Vouliagméni, Perakhóra, Central Greece, *Archaeometry*, **19**, 33-43.

[90] Baxter, M.J. (1989), Multivariate-analysis of data on glass compositions - A methodological note, *Archaeometry*, **31**, 45-53.

[91] Baxter, M.J. (1991), Principal component and correspondence analyses of glass compositions - An empirical-study, *Archaeometry*, **33**, 29-41.

[92] Baxter, M.J. (1992), Statistical analysis of chemical compositional data and the comparison of analyses, *Archaeometry*, **34**, 267-277.

[93] Baxter, M.J., Cool, H.E.M., Heyworth, M.P. and Jackson, C. (1995), Compositional variability in colourless Roman vessel glass, *Archaeometry*, **37**, 129-141.

[94] Bell, S. and Croson, C. (1998), Artificial neural networks as a tool for archaeological data analysis, *Archaeometry*, **40**, 139-151.

[95] Bello, M.A. and Martin, A. (1992), Microchemical characterization of building stone from Seville Cathedral, Spain, *Archaeometry*, **34**, 21-29.

[96] Bieber Jr., A.M., Brookes, D.W., Harbottle, G. and Sayre, E.V. (1976), Application of multivariate techniques to analytical data on Aegean ceramics, *Archaeometry*, **18**, 59-74.

[97] Bimson, M., Laneice, S. and Leese, M. (1982), The characterization of mounted garnets, *Archaeometry*, **24**, 51-58.

[98] Birgül, O., Dukšić, M. and Yaffe, L. (1979), X-ray flourescence analysis of Turkish clays and pottery, *Archaeometry*, **21**, 203-210.

[99] Blasius, E., Wagner, H., Braun, H., Krumbholz, R. and Schwartz, B. (1983), Tile fragments as characteristic evidence for ancient-roman settlements - scientific investigations, *Archaeometry*, **25**, 165-178.

[100] Brodie, N.J. and Steel, L. (1996), Cypriot Black-on-Red wares: towards a classification, *Archaeometry*, **38**, 263-278.

[101] Burmester, A. (1983), Far-eastern lacquers - classification by pyrolysis mass-spectrometry, *Archaeometry*, **25**, 45-58.

[102] Capannesi, G., Seccaroni, C., Sedda, A.F., Majerini, V. and Musco, S (1991), Classification of 15th-century to 19th-century mortars from Gabii using instrumental neutron-activation analysis, *Archaeometry*, **33**, 255-266.

[103] Carter, G.F. and Frurip, D.J. (1985), Discriminant-analysis of the chemical compositions and physical measurements of 245 Augustan Quadrantes, *Archaeometry*, **27**, 117-126.

[104] Carter, G.F. and Powell, R.R. (1994), The chronology of groups of dies for large issues of ancient coins using Mahalanobis distances, *Archaeometry*, **36**, 277-286.

[105] Castellano, A., D'Innocenzo, A., Pagliara, C. and Raho, F. (1996), Composition and origin of Iapygian pottery from Roca Vecchia, Italy, *Archaeometry*, **38**, 59-65.

[106] Christie, O.H.J., Brenna, J.A. and Straume, E. (1979), Multivariate classification of Roman glasses found in Norway, *Archaeometry*, **21**, 233-241.

[107] Cox, G.A. and Gillies, K.J.S. (1986), The X-ray-fluorescence analysis of Medieval durable blue soda glass from York-Minster, *Archaeometry*, **28**, 57-68.

[108] Craddock, P.T., Cowell, M.R., Leese, M.N. and Hughes, M.J. (1983), The trace-element composition of polished flint axes as an indicator of source, *Archaeometry*, **25**, 135-163.

[109] Djingova, R. and Kuleff, I. (1992), An archaeometric study of medieval glass from the 1st Bulgarian capital, Pliska (9th to 10th-century A.D.), *Archaeometry*, **34**, 53-61.

[110] Doherty, C.J. and Maske, A.L. (1998), Characterization of Takatori stoneware from Chikuzen province, Japan, *Archaeometry*, **40**, 71-95.

[111] Ericson, J.E. and Kimberlin, J. (1977), Obsidian sources; chemical characterization and hydration routes in west Mexico, *Archaeometry*, **19**, 157-166.

[112] Glinsman, L.A. and Hayek, L.C. (1993), A multivariate-analysis of renaissance portrait medals - An expanded nomenclature for defining alloy composition, *Archaeometry*, **35**, 49-67.

[113] Gunneweg, J., Beier, T., Diehl, U., Lambrecht, D. and Mommsen, H. (1991), "Edomite", "Negbite" and "'Midianite" - pottery from the Negev desert and Jordan - Instrumental neutron-activation analysis results, *Archaeometry*, **33**, 239-253.

[114] Hall, M. and Yablonsky, L. (1997), Chemical analyses of glass beads found in two Sarmatian burials, *Archaeometry*, **39**, 369-377.

[115] Hammond, N., Harbottle, G. and Gazard, T. (1975), Neutron activation and statistical analysis of Maya ceramics and clays from Lubaantun, Belize, *Archaeometry*, **18**, 147-168.

[116] Hatcher, H., Hedges, R.E.M., Pollard, A.M. and Kenrick P.M. (1980), Analysis of Hellenistic and Roman fine pottery from Benghazi, *Archaeometry*, **22**, 133-151.

[117] Hedges, R.E.M. and Salter, C.J. (1979), Source determination of iron currency bars through the analysis of slag inclusions, *Archaeometry*, **21**, 161-175.

[118] Holmes, L.L. and Harbottle, G. with Blanc, A. (1994), Compositional characterization of French limestone: a new tool for art historians, *Archaeometry*, **36**, 25-39.

[119] Kallithrakas-Kontos, N., Katsanos, A.A., Aravantinos, A., Oeconomides, M., and Touratsoglou, I. (1993), Study of ancient-greek copper coins from Nikopolis (Epirus) and Thessaloniki (Macedonia), *Archaeometry*, **35**, 265-278.

[120] Kerschner, M., M.ommsen, H., Beier, T., Heimermann, D. and Hein, A. (1993), Neutron-activation analysis of bird bowls and related archaic ceramics from Miletus, *Archaeometry*, **35**, 197-210.

[121] Kilikoglou, V., Maniatis, Y. and Grimanis, A.P. (1988), The effect of purification and firing of clays on trace-element provenance studies, *Archaeometry*, **30**, 37-46.

[122] Koh Choo, C.K., Kim, S., Kang, H.T., Do, J.Y., Loo, Y.E. and Kim, G.H. (1999), A comparative scientific study of the earliest kiln sites of Koryo celadon, *Archaeometry*, **41**, 51-69.

[123] Krywonos, W., Newton, G.W.A., Robinson, V.J. and Riley, J.A. (1980), Neutron activation analysis of Roman coarse ware from Cyrenaica, *Archaeometry*, **22**, 189-196.

[124] Kuleff, I., Djingova, R. and djingov, G. (1985), Provenance study of medieval bulgarian glasses by NAA and cluster-analysis, *Archaeometry*, **27**, 185-193.

[125] Lambert, J.B., Szpunar, C.B. and Buikstra, J.E. (1979), Chemical analysis of excavated human bone from middle and late woodland sites, *Archaeometry*, **21**, 115-129.

[126] McLachlan, G.J., Peel, D., Basford, K.E., and Adams, P. (1999), The EMMIX software for the fitting of mixtures of normal and t-components, *Journal of Statistical Software 4*, **No. 2**.

[127] Lambert, J.R., McLoughlin, C.D. and Leonard Jr., A. (1978), X-ray photoelection spectroscopic analysis of the Mycenaean pottery from Megiddo, *Archaeometry*, **20**, 107-122.

[128] Mello, E., Monna, D. and Oddone, M. (1988), Discriminating sources of mediterranean marbles - a pattern-recognition approach, *Archaeometry*, **30**, 102-108.

[129] Mirti, P., Aceto, M. and Preacco Ancona, M.C. (1998), Campanian pottery from ancient Bruttium (Southern Italy): Scientific analysis of local and imported products, *Archaeometry*, **40**, 311-329.

[130] Mirti, P., Casoli, A. and Appolonia, L. (1993), Scientific analysis of roman glass from Augusta-Praetoria, *Archaeometry*, **35**, 225-240.

[131] Mirti, P., Casoli, A., Barra Bagnasco, M. and Preacco Ancona (1995), Fine wares from Locri Epizephiri : a provenance study by inductively coupled plasma emission spectroscopy, *Archaeometry*, **37**, 41-51.

[132] Mirti, P., Zelano, V., Aruga, R., Ferrara, E. and Appolonia, L (1990), Roman pottery from Augusta-Praetoria (Aosta, Italy) - A provenance study, *Archaeometry*, **32**, 163-175.

[133] Molera, J., Garíco-Vallés, M., Pradell, T. and Vendrell-Saz, M. (1996), Hispano-Moresque pottery production of the fourteenth-centuary workshop of Testar de Molí (Paterno, Spain), *Archaeometry*, **38**, 67-80.

[134] Mommsen, H., Kreuser, A. and Weber, J. (1988), A method for grouping pottery by chemical-composition, *Archaeometry*, **30**, 47-57.

[135] Neff, H. (1994), RQ-mode principal components analysis of ceramic compositional data, *Archaeometry*, **36**, 115-130.

[136] Newman, J.R. and Nielsen, R.L. (1987), Initial notes on the X-ray-fluorescence characterization of the Rhyodacite sources of the Taos Plateau, New-Mexico, *Archaeometry*, **29**, 262-274.

[137] Philip, G. and Ottaway, B.S. (1983), Mixed data cluster-analysis - an illustration using Cypriot hooked-tang weapons, *Archaeometry*, **25**, 119-133.

[138] Pike, H.H.M. and Fulford, M.G. (1983), Neutron-activation-analysis of black-glazed pottery from Carthage, *Archaeometry*, **25**, 77-86.

[139] Pollard, A.M. and Hatcher, H. (1994), The chemical analysis of oriental ceramic body compositions: part 1: wares from north China, *Archaeometry*, **36**, 41-62.

[140] Rauret, G., Casassas, E., Rius, F.X. and Muñoz, M. (1987), Cluster-analysis applied to spectrochemical data of european medieval stained glasses, *Archaeometry*, **29**, 240-249.

[141] Rotunno, T., Sabbatini, L. and Corrente, M. (1997), A provenance study of pottery from archaeological sites near Canosa, Puglia (Italy), *Archaeometry*, **39**, 343-354.

[142] Sayre, E.V., Yener, K.A., Joel, E.C. and Barnes, I.L. (1992), Statistical evaluation of the presently accumulated lead isotope data from Anatolia and surrounding regions, *Archaeometry*, **34**, 73-105.

[143] Schubert, P. (1986), Petrographic modal-analysis - a necessary complement to chemical-analysis of ceramic coarse ware, *Archaeometry*, **28**, 163-178.

[144] Tangri, D. and Wright, R.V.S. (1993), Multivariate-analysis of compositional data - Applied comparisons favor standard principal components-analysis over Aitchison loglinear contrast method, *Archaeometry*, **35**, 103-112.

[145] Taylor, R.J. and Robinson, V.J. (1996), Neutron activation analysis of Roman African red slip ware kilns, *Archaeometry*, **38**, 231-243.

[146] Taylor, R.J. and Robinson, V.J. (1996), Provenance studies of Roman African Red Slip ware using neutron activation analysis, *Archaeometry*, **38**, 245-255.

[147] Taylor, R.J., Robinson, V.J. and Gibbins, D.J.L. (1997), An investigation of the provenance of the Roman amphora cargo from the Plemmirio B shipwreck, *Archaeometry*, **39**, 9-21.

[148] Topping, P.G. and Mackenzie, A.B. (1988), A test of the use of neutron-activation analysis for clay source characterization, *Archaeometry*, **30**, 92-101.

[149] Troja, S.O., Cro, A., Gueli, A.M., La Rosa, V., Mazzoleni, P., Pezzino, A. and Romeo, M. (1996), Characterization and thermoluminescence dating of prehistoric pottery shreds from Milena, *Archaeometry*, **38**, 113-128.

[150] Truncer, J., Glascock, M.D. and Neff, H. (1998), Steatite source characterization in eastern North America: New results using instrumental neutron activation analysis, *Archaeometry*, **40**, 23-44.

[151] Tubb, A., Parker, A.J. and Nickless, G. (1980), The analysis of Romano-British pottery by atomic absorption spectrophotometry, *Archaeometry*, **22**, 153-171.

[152] Vednrell-Saz, M., Alarcón, S., Molera, J. and Garíco-Vallés, M. (1996), Dating ancient lime mortars by geochemical and mineralogical analysis, *Archaeometry*, **38**, 143-149.

[153] Yap, C.T. and Younan Hua (1994), A study of Chinese porcelain raw materials for Ding. Xing, Gangxian and Dehuo wares, *Archaeometry*, **36**, 63-76.

[154] Yu, K.N. and Miao, J.M. (1998), Multivariate analysis of the energy dispersive x-ray fluorescence results from blue and white Chinese porcelains, *Archaeometry*, **40**, 331-339.

Appendix A

Bivariate Mixtures

The following table details the 12 bivariate mixture distributions suggested by Wand and Jones (1993). The distributions are selected to show a sub-set of possible forms of bivariate structure. We use these distributions in chapter 4 to investigate sample size requirements necessary to detect varying forms of structure. The distributions are defined using the form $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ as done so by Wand and Jones (1993).

Density	Parameters
(A) Uncorrelated normal	$N(0, 0, \frac{1}{4}, 1, 0)$
(B) Correlated normal	$N(0, 0, 1, 1, \frac{7}{10})$
(C) Skewed	$\frac{1}{5}N(0,0,1,1,0) + \frac{1}{5}N(\frac{1}{2},\frac{1}{2},(\frac{2}{3})^2,(\frac{2}{3})^2,0) + \frac{3}{5}N(\frac{13}{12},\frac{13}{12},(\frac{5}{9})^2,(\frac{5}{9})^2,0)$
(D) Kurtotic	$\frac{2}{3}N(0,0,1,4,\frac{1}{2}) + \frac{1}{3}N(0,0,(\frac{2}{3})^2,(\frac{1}{3})^2,-\frac{1}{2})$
(E) Bimodal I	$\frac{1}{2}N(-1,0,(\frac{2}{3})^2,(\frac{2}{3})^2,0) + \frac{1}{2}N(1,0,(\frac{2}{3})^2,(\frac{2}{3})^2,0)$
(F) Bimodal II	$\frac{1}{2}N(-\frac{3}{2},0,(\frac{1}{4})^2,1,0) + \frac{1}{2}N(\frac{3}{2},0,(\frac{1}{4})^2,1,0)$
(G) Bimodal III	$\frac{1}{2}N(-1,1,(\frac{2}{3})^2,(\frac{2}{3})^2,\frac{3}{5}) + \frac{1}{2}N(1,-1,(\frac{2}{3})^2,(\frac{2}{3})^2,\frac{3}{5})$
(H) Bimodal IV	$\frac{1}{2}N(-1,1,(\frac{2}{3})^2,(\frac{2}{3})^2,\frac{7}{10}) + \frac{1}{2}N(1,-1,(\frac{2}{3})^2,(\frac{2}{3})^2,0)$
(I) Trimodal I	$\frac{5}{20}N(-\frac{6}{5},\frac{6}{5},(\frac{3}{5})^2,(\frac{3}{5})^2,\frac{3}{10}) + \frac{5}{20}N(\frac{6}{5},-\frac{6}{5},(\frac{3}{5})^2,(\frac{3}{5})^2,-\frac{3}{5}) + \frac{1}{10}N(0,0,(\frac{1}{4})^2,(\frac{1}{4})^2,\frac{1}{5})$
(J) Trimodal II [1]	$\frac{1}{3}N(-\frac{6}{5},0,(\frac{3}{5})^2,(\frac{3}{5})^2,\frac{7}{10}) + \frac{1}{3}N(\frac{6}{5},0,(\frac{3}{5})^2,(\frac{3}{5})^2,\frac{7}{10}) + \frac{1}{3}N(0,0,(\frac{3}{5})^2,(\frac{3}{5})^2,-\frac{7}{10})$
(K) Trimodal III	$\frac{3}{7}N(-1,0,(\frac{3}{5})^2,(\frac{7}{10})^2,\frac{3}{5}) + \frac{3}{7}N(1,\frac{2\sqrt{3}}{3},(\frac{3}{5})^2,(\frac{7}{10})^2,0) + \frac{1}{7}N(1,-\frac{2\sqrt{3}}{3},(\frac{3}{5})^2,(\frac{7}{10})^2,0)$
(L) Quadimodal	$\frac{1}{8}N(-1,1,(\frac{2}{3})^2,(\frac{2}{3})^2,\frac{2}{5}) + \frac{3}{8}N(-1,-1,(\frac{2}{3})^2,(\frac{2}{3})^2,\frac{3}{5}) + \frac{1}{8}N(1,-1,(\frac{2}{3})^2,(\frac{2}{3})^2,\frac{7}{10}) + \frac{3}{8}N(1,1,(\frac{2}{3})^2,(\frac{2}{3})^2,-\frac{1}{2})$

[1] Although Wand and Jones (1993) list this mixture as being trimodal, figure A.1j clearly shows it to be bimodal.

Table A.1: Details of the 12 bivariate mixture distributions used by Wand and Jones (1993).
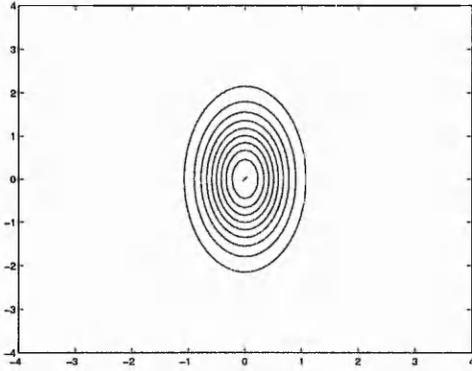
(a) Uncorrelated normal
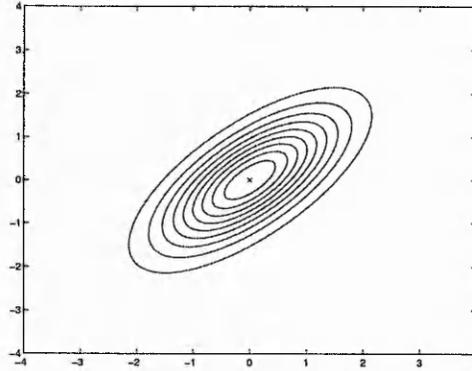
(b) Correlated normal
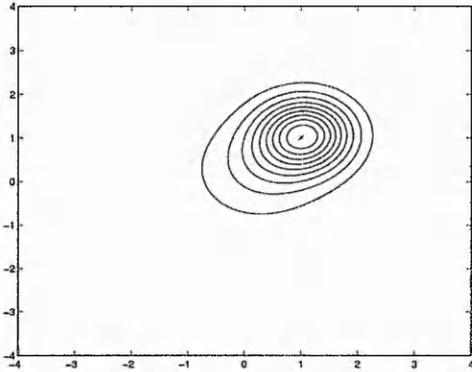
(c) Skewed
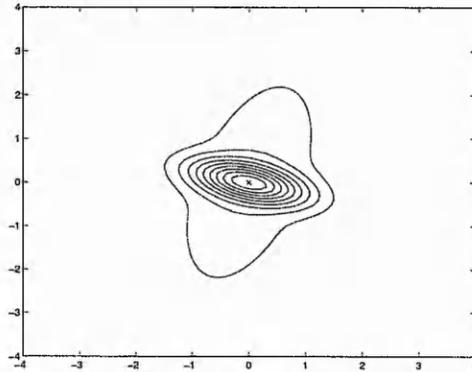
(d) Kurtotic
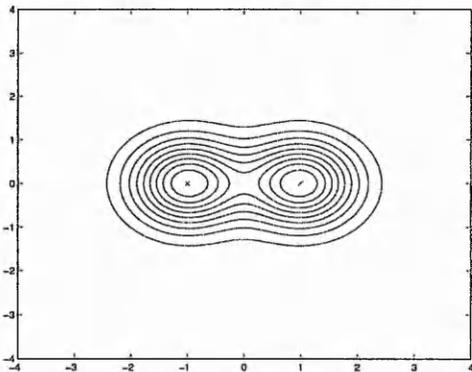
(e) Bimodal I

(f) Bimodal II

125
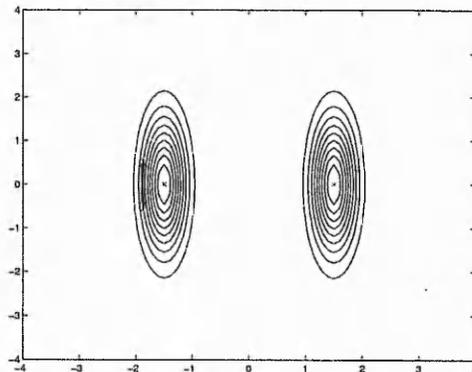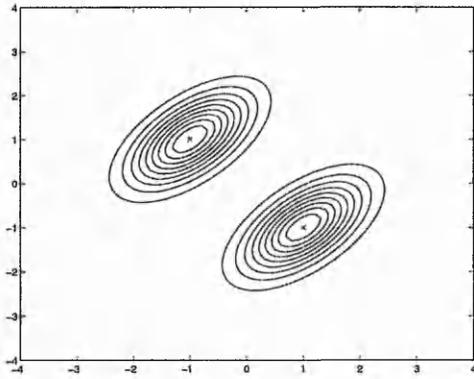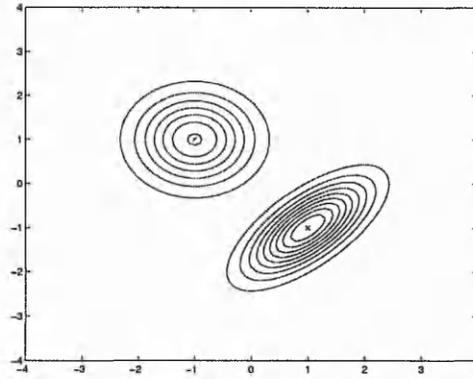
(g) Bimodal III

(h) Bimodal IV
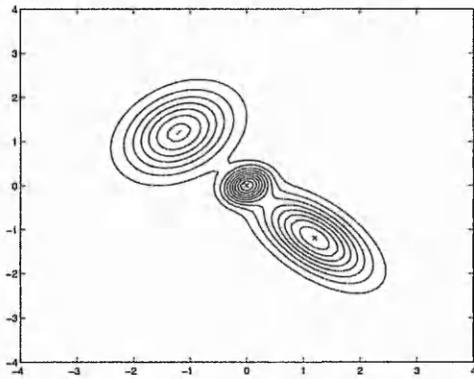
(i) Trimodal I
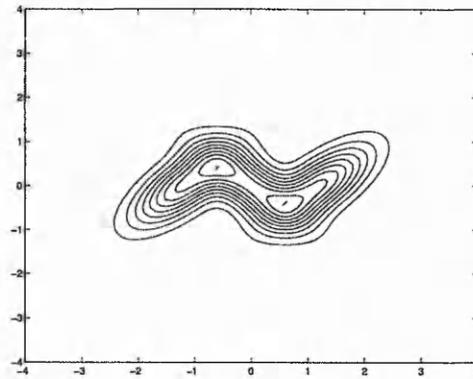
(j) Trimodal II
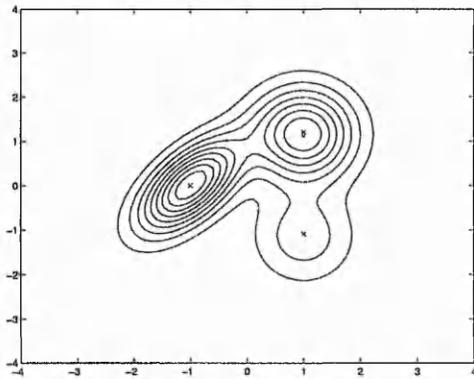
(k) Trimodal III
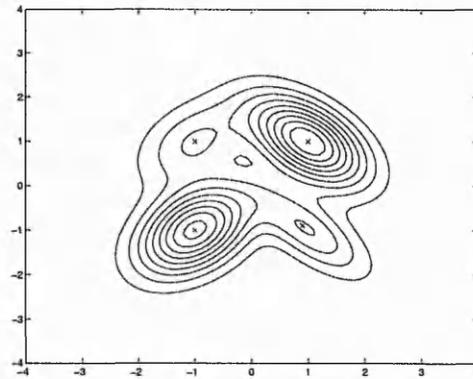
(l) Quadimodal

Figure A.1: Contour plots of the 12 mixtures used in chapter 4 to investigate sample size requirements for varying forms of bivariate structure. Modes are indicated with a 'x'.

Appendix B

Results from Simulation Studies Undertaken in Chapter 4

The tables below detail results from the simulation studies undertaken in chapter 4. Table B.1 shows the performance of the ISE statistic (section 4.2.1). The table shows the proportion of 1000 replications for which a departure from normality was detected at the 5% level for samples from 14 univariate populations of the form $(1 - p)N(0,1) + pN(\mu, \sigma^2)$ for varying sample sizes.

Tables B.2 to B.5 show results from the univariate mode counting undertaken in section 4.2.2. The table details the number of 1000 replications for which 1, 2 or more than 2 modes are detected in 14 simulated univariate distributions of the form $(1 - p)N(0,1) + pN(\mu, \sigma^2)$ for varying sample sizes. The four tables detail results for mode counting for non-adaptive KDEs with h selected by h_{NS}, adaptive KDEs with h selected by h_{NS}, non-adaptive KDEs with h selected by h_{STE} and adaptive KDEs with h selected by h_{STE}.

Tables B.6 to B.11 detail results from the bivariate mode counting procedure undertaken in section 4.4.2. The tables detail the number of 1000 replications for which 1, 2, 3, 4, 5 or more modes were detected in KDEs of samples taken from the 12 bivariate mixture distributions, discussed in appendix A, for varying sample sizes. The KDEs are both non-adaptive and adaptive with window width selected by the univariate h_{NS} and h_{STE} procedures and also using the bivariate window-width selection routine, h_{DPI2}.

p	σ^2	n	Proportion of replications which are significant at the 5% level						
			$\mu = 0$	$\mu = 1$	$\mu = 2$	$\mu = 3$	$\mu = 4$	$\mu = 5$	$\mu = 6$
0.5	1	20	0.0520	0.0630	0.0500	0.1090	0.3780	0.7480	0.9560
0.5	1	30	0.0530	0.0590	0.0660	0.2650	0.6950	0.9700	1.0
0.5	1	40	0.0530	0.0590	0.0660	0.2650	0.6950	0.9700	1.0
0.5	1	50	0.0620	0.0450	0.1060	0.4690	0.9530	1.0	1.0
0.5	1	100	0.0450	0.0580	0.1510	0.8560	0.9990	1.0	1.0
0.1	1	50	0.0580	0.0400	0.0980	0.5050	0.8650	0.9560	0.9790
0.2	1	50	0.0590	0.0520	0.1590	0.6530	0.9760	1.0	1.0
0.3	1	50	0.0510	0.0680	0.1530	0.6270	0.9710	1.0	1.0
0.4	1	50	0.0360	0.0420	0.0940	0.5190	0.9620	1.0	1.0
0.5	1	50	0.0520	0.0450	0.0910	0.5130	0.9550	1.0	1.0
0.1	1	25	0.0600	0.0500	0.1100	0.3420	0.6110	0.8530	0.9020
0.2	1	25	0.0480	0.0560	0.1190	0.3840	0.7680	0.9620	0.9880
0.3	1	25	0.0440	0.0620	0.1000	0.3210	0.7370	0.9530	0.9970
0.4	1	25	0.0470	0.0520	0.0600	0.2520	0.6590	0.9350	0.9940
0.5	1	25	0.0560	0.0470	0.0520	0.2120	0.5710	0.9270	0.9960
0.1	1	100	0.0610	0.0470	0.1980	0.7690	0.9860	0.9980	1.0
0.2	1	100	0.0510	0.0500	0.2750	0.9430	1.0	1.0	1.0
0.3	1	100	0.0560	0.0570	0.2780	0.9280	1.0	1.0	1.0
0.4	1	100	0.0680	0.0620	0.2110	0.8750	1.0	1.0	1.0
0.5	1	100	0.0600	0.0480	0.1540	0.8610	0.9990	1.0	1.0
0.5	0.25	50	0.8450	0.9730	0.9940	1.0	1.0	1.0	1.0
0.5	0.5	50	0.2060	0.5560	0.8260	0.9870	1.0	1.0	1.0
0.5	1	50	0.0400	0.0580	0.0890	0.4920	0.9460	1.0	1.0
0.5	2	50	0.2100	0.3200	0.5610	0.7000	0.8250	0.9350	0.9860
0.5	4	50	0.8240	0.8700	0.9080	0.9580	0.9730	0.9830	0.9950
0.5	0.25	25	0.5220	0.7960	0.8380	0.9110	0.9800	0.9990	1.0
0.5	0.5	25	0.1490	0.3270	0.4600	0.7510	0.9380	0.9980	1.0
0.5	1	25	0.0600	0.0610	0.0600	0.2180	0.5990	0.9240	0.9960
0.5	2	25	0.1420	0.1910	0.3230	0.4040	0.4510	0.5520	0.7210
0.5	4	25	0.5460	0.5840	0.6460	0.7160	0.8000	0.8220	0.8250
0.5	0.25	100	0.9910	1.0	1.0	1.0	1.0	1.0	1.0
0.5	0.5	100	0.2890	0.8410	0.9890	1.0	1.0	1.0	1.0
0.5	1	100	0.0540	0.0620	0.1790	0.8790	1.0	1.0	1.0
0.5	2	100	0.3430	0.5430	0.8560	0.9500	0.9950	0.9990	1.0
0.5	4	100	0.9930	0.9900	0.9990	0.9980	1.0	1.0	1.0

p	σ^2	n	Proportion of replications which are significant at the 5% level						
			$\mu = 0$	$\mu = 1$	$\mu = 2$	$\mu = 3$	$\mu = 4$	$\mu = 5$	$\mu = 6$
0.1	0.25	50	0.0890	0.0980	0.1220	0.5190	0.8910	0.9650	0.9900
0.1	0.5	50	0.0660	0.0660	0.0750	0.4850	0.8800	0.9670	0.9900
0.1	1	50	0.0520	0.0480	0.1340	0.5090	0.8410	0.9610	0.9830
0.1	2	50	0.1220	0.2140	0.4090	0.6480	0.8470	0.9320	0.9780
0.1	4	50	0.6850	0.6840	0.7480	0.8340	0.8940	0.9250	0.9580
0.3	0.25	50	0.3920	0.6410	0.6720	0.9420	1.0	1.0	1.0
0.3	0.5	50	0.1210	0.2380	0.3130	0.8120	0.9990	1.0	1.0
0.3	1	50	0.0500	0.0500	0.1440	0.6370	0.9780	1.0	1.0
0.3	2	50	0.2460	0.3710	0.6930	0.9090	0.9840	1.0	1.0
0.3	4	50	0.9260	0.9480	0.9740	0.9920	0.9980	0.9990	1.0

Table B.1: Results from the power study undertaken on univariate mixture distributions of the form $(1 - p)N(0, 1) + pN(\mu, \sigma^2)$ for varying values of μ, σ^2, p and sample size. The table details the proportion of 1000 replications for which the ISE statistic is significant at the 5% level for a sample of size n from the mixture distribution.

σ^2	p	n	$\mu=1$ Reps with m			$\mu=2$ Count with m			$\mu=3$ Count with m			$\mu=4$ Count with m			$\mu=5$ Count with m			$\mu=6$ Count with m		
			$=2$	$=1$	>2	$=2$	$=1$	>2	$=2$	$=1$	>2	$=2$	$=1$	>2	$=2$	$=1$	>2	$=2$	$=1$	>2
1	0.5	20	142	845	13	174	817	9	328	671	1	692	308	0	898	100	2	985	14	1
1	0.5	25	142	845	13	170	828	2	361	639	0	759	241	0	947	52	1	997	3	0
1	0.5	30	109	885	6	152	847	1	421	579	0	842	158	0	972	27	1	1000	0	0
1	0.5	40	79	918	3	163	836	1	484	516	0	899	101	0	997	3	0	1000	0	0
1	0.5	50	71	926	3	135	865	0	571	429	0	953	47	0	997	3	0	1000	0	0
1	0.5	100	33	967	0	172	828	0	767	233	0	999	1	0	1000	0	0	1000	0	0
1	0.1	25	156	833	11	163	824	13	309	663	28	487	477	36	589	372	39	615	329	56
1	0.1	50	78	921	1	113	885	2	308	679	13	563	409	28	631	330	39	631	337	32
1	0.1	100	48	952	0	49	950	1	286	711	3	663	314	23	664	291	45	684	274	42
1	0.3	25	152	834	14	162	827	11	328	663	9	679	305	16	888	72	40	950	13	37
1	0.3	50	85	914	1	110	888	2	315	683	2	783	206	11	969	16	15	987	0	13
1	0.3	100	43	957	0	54	946	0	301	698	1	908	91	1	997	1	2	997	0	3
0.25	0.5	25	437	496	67	552	432	16	857	137	6	976	19	5	998	0	2	997	0	3
0.25	0.5	50	391	572	37	615	379	6	974	26	0	1000	0	0	1000	0	0	1000	0	0
0.25	0.5	100	354	626	20	787	213	0	998	2	0	1000	0	0	1000	0	0	1000	0	0
0.5	0.5	25	251	731	18	379	610	11	744	252	4	966	33	1	999	1	0	993	0	7
0.5	0.5	50	160	838	2	403	597	0	919	80	1	999	1	0	1000	0	0	1000	0	0
0.5	0.5	100	105	895	0	451	549	0	989	11	0	1000	0	0	1000	0	0	1000	0	0
2	0.5	25	194	790	16	224	762	14	235	763	2	332	664	4	510	488	2	745	254	1
2	0.5	50	414	458	128	389	536	75	399	562	39	337	644	19	357	633	10	373	624	3
2	0.5	100	103	894	3	101	897	2	155	845	0	474	526	0	858	142	0	990	10	0
4	0.5	25	413	472	115	391	531	78	371	581	48	365	609	26	367	619	14	389	595	16
4	0.5	50	414	458	128	389	536	75	399	562	39	337	644	19	357	633	10	373	624	3
4	0.5	100	430	469	101	418	519	63	382	574	44	366	617	17	373	622	5	416	583	1

Table B.2: Results from the simulation study undertaken on univariate mixture distributions of the form $(1-p)N(0,1) + pN(\mu, \sigma^2)$ for varying values of μ, σ^2, p and sample size. The table details the number of 1000 replications for which the the univariate mode counting process detected 1, 2 or more than 2 modes in sample of size n from the mixture distribution. KDEs which modes are counted in are non-adaptive with window-width selected by h_{NS}.

130

σ^2	p	n	$\mu=1$ Reps with m			$\mu=2$ Count with m			$\mu=3$ Count with m			$\mu=4$ Count with m			$\mu=5$ Count with m			$\mu=6$ Count with m		
			$=2$	$=1$	>2	$=2$	$=1$	>2	$=2$	$=1$	>2	$=2$	$=1$	>2	$=2$	$=1$	>2	$=2$	$=1$	>2
1	0.5	20	52	946	2	117	881	2	276	723	1	560	440	0	802	198	0	930	70	0
1	0.5	25	99	899	2	142	855	3	356	644	0	687	313	0	907	93	0	967	33	0
1	0.5	30	74	924	2	149	851	0	417	583	0	804	196	0	951	49	0	994	6	0
1	0.5	40	84	916	0	215	785	0	531	469	0	884	116	0	983	17	0	998	2	0
1	0.5	50	127	873	0	208	792	0	636	364	0	952	48	0	996	4	0	1000	0	0
1	0.5	100	162	838	0	348	652	0	874	126	0	999	1	0	1000	0	0	1000	0	0
1	0.1	25	101	896	3	91	908	1	94	903	3	113	884	3	135	862	3	147	846	7
1	0.1	50	103	897	0	94	905	1	70	929	1	91	907	2	118	880	2	120	875	5
1	0.1	100	158	842	0	107	893	0	82	918	0	137	863	0	151	848	1	147	850	3
1	0.3	25	107	891	2	111	888	1	182	817	1	429	563	8	660	334	6	834	160	6
1	0.3	50	120	880	0	122	878	0	174	825	1	518	480	2	846	151	3	946	50	4
1	0.3	100	151	849	0	167	833	0	220	780	0	739	261	0	980	18	2	994	6	0
0.25	0.5	25	190	809	1	467	533	0	850	150	0	977	23	0	990	10	0	993	7	0
0.25	0.5	50	110	890	0	644	356	0	989	11	0	998	2	0	999	1	0	999	1	0
0.25	0.5	100	93	907	0	841	159	0	1000	0	0	1000	0	0	1000	0	0	1000	0	0
0.5	0.5	25	95	902	3	266	734	0	703	297	0	938	62	0	981	19	0	989	11	0
0.5	0.5	50	63	937	0	378	622	0	917	83	0	999	1	0	1000	0	0	999	1	0
0.5	0.5	100	62	937	1	529	471	0	995	5	0	1000	0	0	1000	0	0	1000	0	0
2	0.5	25	77	918	5	85	915	0	112	888	0	244	756	0	446	554	0	688	312	0
2	0.5	50	77	920	3	57	941	2	51	949	0	78	921	1	138	862	0	274	726	0
2	0.5	100	83	917	0	54	946	0	142	858	0	537	463	0	904	96	0	997	3	0
4	0.5	25	119	876	5	84	911	5	81	919	0	85	914	1	121	879	0	209	791	0
4	0.5	50	77	920	3	57	941	2	51	949	0	78	921	1	138	862	0	274	726	0
4	0.5	100	94	901	5	87	911	2	72	927	1	101	898	1	217	783	0	367	633	0

Table B.3: Results from the simulation study undertaken on univariate mixture distributions of the form $(1-p)N(0,1) + pN(\mu, \sigma^2)$ for varying values of μ, σ^2, p and sample size. The table details the number of 1000 replications for which the the univariate mode counting process detected 1, 2 or more than 2 modes in sample of size n from the mixture distribution. KDEs which modes are counted in are adaptive with window-width selected by h_{NS}.

σ^2	p	n	$\mu=1$ Reps with m =2	=1	>2	$\mu=2$ Count with m =2	=1	>2	$\mu=3$ Count with m =2	=1	>2	$\mu=4$ Count with m =2	=1	>2	$\mu=5$ Count with m =2	=1	>2	$\mu=6$ Count with m =2	=1	>2
1	0.5	20	235	641	124	303	584	113	501	397	102	783	112	105	882	17	101	917	2	81
1	0.5	25	227	649	124	281	636	83	520	410	70	795	91	114	884	3	113	939	0	61
1	0.5	30	220	706	74	277	667	56	557	375	68	856	53	91	927	3	70	956	0	44
1	0.5	40	185	751	64	265	690	45	614	328	58	883	39	78	940	1	59	954	0	46
1	0.5	50	167	786	47	241	728	31	663	286	51	914	16	70	944	0	56	961	0	39
1	0.5	100	105	880	15	246	735	19	804	147	49	946	1	53	974	0	26	972	0	28
1	0.1	25	211	695	94	235	644	121	334	512	154	449	370	181	509	305	186	519	276	205
1	0.1	50	178	765	57	174	762	64	349	544	107	555	321	124	573	280	147	574	287	139
1	0.1	100	106	869	25	126	853	21	376	589	35	640	257	103	635	257	108	667	239	94
1	0.3	25	237	653	110	240	645	115	433	421	146	747	107	146	820	16	164	847	4	149
1	0.3	50	177	779	44	258	703	39	552	382	66	834	44	122	874	1	125	892	0	108
1	0.3	100	104	874	22	161	821	18	588	361	51	875	10	115	900	0	100	940	0	60
0.25	0.5	25	504	137	359	358	30	612	326	2	672	396	1	603	516	0	484	639	0	361
0.25	0.5	50	424	68	508	203	3	794	162	0	838	264	0	736	384	0	616	468	0	532
0.25	0.5	100	407	28	565	111	0	889	95	0	905	163	0	837	248	0	752	366	0	634
0.5	0.5	25	348	523	129	546	289	165	689	46	265	711	1	288	790	0	210	822	0	178
0.5	0.5	50	335	588	77	618	205	177	687	5	308	702	0	298	735	0	265	832	0	168
0.5	0.5	100	317	630	53	640	144	216	635	0	365	691	0	309	751	0	249	821	0	179
2	0.5	25	262	608	130	344	517	139	441	446	113	535	299	166	614	136	250	690	50	260
2	0.5	50	235	676	89	338	582	80	424	456	120	590	222	188	661	57	282	655	3	342
2	0.5	100	200	756	44	303	649	48	471	430	99	670	125	205	654	9	337	634	0	366
4	0.5	25	407	279	314	435	250	315	496	219	285	505	173	322	486	127	387	467	92	441
4	0.5	50	401	226	373	465	194	341	460	133	407	465	95	440	408	43	549	339	21	640
4	0.5	100	471	200	329	493	159	348	481	97	422	379	41	580	284	11	705	205	3	792

Table B.4: Results from the simulation study undertaken on univariate mixture distributions of the form $(1-p)N(0,1) + pN(\mu,\sigma^2)$ for varying values of μ, σ^2, p and sample size. The table details the number of 1000 replications for which the the univariate mode counting process detected 1, 2 or more than 2 modes in sample of size n from the mixture distribution. KDEs which modes are counted in are non-adaptive with window-width selected by h_{STE}.

σ^2	p	n	$\mu=1$ Reps with m			$\mu=2$ Count with m			$\mu=3$ Count with m			$\mu=4$ Count with m			$\mu=5$ Count with m			$\mu=6$ Count with m		
			$=2$	$=1$	>2	$=2$	$=1$	>2	$=2$	$=1$	>2	$=2$	$=1$	>2	$=2$	$=1$	>2	$=2$	$=1$	>2
1	0.5	20	232	737	31	320	654	26	547	430	23	844	133	23	964	23	13	989	3	8
1	0.5	25	255	715	30	312	661	27	564	417	19	873	97	30	969	7	24	993	1	6
1	0.5	30	211	756	33	292	678	30	602	369	29	928	51	21	983	3	14	992	1	7
1	0.5	40	193	774	33	304	678	18	670	300	30	930	32	38	987	2	11	986	0	14
1	0.5	50	205	777	18	314	677	9	737	231	32	949	14	37	982	0	18	984	0	16
1	0.5	100	209	779	12	350	637	13	859	83	58	925	0	75	959	0	41	976	0	24
1	0.1	25	216	748	36	241	719	40	242	716	42	245	698	57	282	671	47	271	681	48
1	0.1	50	224	760	16	190	789	21	203	773	24	245	725	30	249	730	21	256	714	30
1	0.1	100	216	767	17	197	799	4	207	786	7	278	700	22	281	697	22	268	711	21
1	0.3	25	257	706	37	276	686	38	436	516	48	732	209	59	853	104	43	894	71	35
1	0.3	50	209	773	18	242	737	21	483	475	42	839	90	71	899	29	72	925	21	54
1	0.3	100	198	788	14	262	723	15	524	405	71	845	20	135	885	1	114	902	0	98
0.25	0.5	25	577	358	65	777	83	140	789	30	181	872	23	105	921	16	63	955	18	27
0.25	0.5	50	661	260	79	696	19	285	658	11	331	764	8	228	873	4	123	899	6	95
0.25	0.5	100	700	156	144	533	0	467	487	1	512	581	0	419	671	1	328	780	0	220
0.5	0.5	25	281	671	48	624	340	36	881	56	63	931	7	62	968	1	31	980	6	14
0.5	0.5	50	267	697	36	737	205	58	878	7	115	907	0	93	954	0	46	979	1	20
0.5	0.5	100	275	696	29	770	117	113	795	0	205	849	0	151	882	0	118	939	0	61
2	0.5	25	219	738	43	290	675	35	421	550	29	609	357	34	778	163	59	897	57	46
2	0.5	50	198	773	29	263	706	31	433	532	35	719	219	62	854	54	92	890	5	105
2	0.5	100	201	777	22	289	685	26	495	446	59	811	90	99	820	4	176	822	0	178
4	0.5	25	326	597	77	397	544	59	474	492	34	526	436	38	627	316	57	699	230	71
4	0.5	50	316	617	67	383	562	55	498	438	64	612	313	75	724	175	101	748	94	158
4	0.5	100	348	579	73	429	467	104	530	335	135	666	176	158	706	68	226	638	22	340

Table B.5: Results from the simulation study undertaken on univariate mixture distributions of the form $(1-p)N(0,1) + pN(\mu,\sigma^2)$ for varying values of μ, σ^2, p and sample size. The table details the number of 1000 replications for which the the univariate mode counting process detected 1, 2 or more than 2 modes in sample of size n from the mixture distribution. KDEs which modes are counted in are adaptive with window-width selected by h_{STE}.

133

Mixture	n	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m > 5$
A	15	227	431	255	68	14	5
A	20	151	373	286	133	39	18
A	25	135	330	282	152	69	32
A	50	269	368	225	96	21	21
A	100	404	410	148	33	4	1
B	15	277	526	160	29	7	1
B	20	204	525	190	59	17	5
B	25	161	463	259	85	21	11
B	50	391	443	142	22	2	0
B	100	494	411	87	8	0	0
C	15	131	393	332	105	31	8
C	20	74	307	356	168	62	33
C	25	61	199	306	216	132	86
C	50	212	393	237	88	45	25
C	100	259	438	211	83	9	0
D	15	117	345	309	149	58	22
D	20	58	189	258	224	147	124
D	25	35	137	192	201	179	256
D	50	93	244	260	193	126	84
D	100	69	218	291	224	135	63
E	15	172	496	276	44	11	1
E	20	155	456	277	86	19	7
E	25	102	383	321	132	47	15
E	50	123	410	312	122	30	3
E	100	88	465	361	79	7	0
F	15	0	435	465	85	13	2
F	20	0	367	461	129	37	6
F	25	0	333	465	133	51	18
F	50	0	533	394	66	6	1
F	100	0	665	298	36	1	0
G	15	1	851	138	10	0	0
G	20	0	877	110	10	2	1
G	25	1	878	111	7	3	0
G	50	0	968	32	0	0	0
G	100	0	986	14	0	0	0
H	15	11	874	106	8	1	0
H	20	6	891	91	11	0	1
H	25	3	905	84	6	2	0
H	50	0	978	22	0	0	0
H	100	0	989	11	0	0	0

Mixture	n	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$	$m>5$
I	15	91	865	38	3	3	0
I	20	57	891	49	2	0	1
I	25	33	893	72	2	0	0
I	50	7	877	115	0	1	0
I	100	1	782	217	0	0	0
J	15	154	447	324	68	6	1
J	20	90	383	390	108	24	5
J	25	78	334	350	162	67	9
J	50	86	347	372	164	28	3
J	100	92	414	373	107	13	1
K	15	104	497	349	45	5	0
K	20	151	373	286	133	39	18
K	25	41	402	439	89	23	6
K	50	37	397	509	52	4	1
K	100	8	278	642	70	2	0
L	15	108	617	247	25	3	0
L	20	74	588	302	33	2	1
L	25	40	583	324	46	5	2
L	50	22	587	343	44	4	0
L	100	4	531	385	77	3	0

Table B.6: Results from the simulation study undertaken on bivariate mixture distributions detailed in appendix A. The table details the number of 1000 replications for which the the bivariate mode counting process detected 1, 2, 3, 4, 5 or more modes sample of size n from the mixture distribution. KDEs are non-adaptive with window-width selected by h_{NS}.

Mixture	n	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m > 5$
A	15	468	442	82	6	2	0
A	20	378	465	139	17	1	0
A	25	287	469	209	32	0	3
A	50	310	471	196	21	2	0
A	100	218	453	277	49	3	0
B	15	521	458	20	1	0	0
B	20	413	531	54	2	0	0
B	25	334	569	89	8	0	0
B	50	390	532	76	2	0	0
B	100	342	563	91	4	0	0
C	15	541	393	64	2	0	0
C	20	451	440	101	8	0	0
C	25	351	445	173	29	0	2
C	50	438	427	120	14	1	0
C	100	384	458	141	16	1	0
D	15	556	367	73	3	1	0
D	20	385	451	139	20	5	0
D	25	295	432	215	51	6	1
D	50	320	439	191	43	6	1
D	100	232	435	248	79	6	0
E	15	324	516	152	7	1	0
E	20	252	545	180	22	1	0
E	25	157	560	257	22	4	0
E	50	101	526	329	44	0	0
E	100	30	404	438	114	14	0
F	15	10	776	204	10	0	0
F	20	2	735	243	18	2	0
F	25	3	661	313	21	2	0
F	50	0	726	262	12	0	0
F	100	0	712	260	28	0	0
G	15	11	960	29	0	0	0
G	20	4	960	34	2	0	0
G	25	2	963	33	2	0	0
G	50	0	982	17	1	0	0
G	100	0	982	18	0	0	0
H	15	34	940	25	1	0	0
H	20	9	967	23	1	0	0
H	25	3	973	22	2	0	0
H	50	0	982	18	0	0	0
H	100	0	986	14	0	0	0

Mixture	n	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m > 5$
I	15	160	825	15	0	0	0
I	20	92	873	35	0	0	0
I	25	57	894	48	1	0	0
I	50	7	865	128	0	0	0
I	100	0	747	253	0	0	0
J	15	347	517	136	0	0	0
J	20	240	538	199	22	0	1
J	25	165	535	253	44	3	0
J	50	89	513	334	61	3	0
J	100	31	415	426	117	10	1
K	15	213	619	162	6	0	0
K	20	378	465	139	17	1	0
K	25	77	632	263	27	1	0
K	50	38	613	328	20	1	0
K	100	3	484	446	64	3	0
L	15	177	708	111	4	0	0
L	20	115	717	162	6	0	0
L	25	58	743	190	9	0	0
L	50	18	685	279	17	1	0
L	100	0	622	326	45	7	0

Table B.7: Results from the simulation study undertaken on bivariate mixture distributions detailed in appendix A. The table details the number of 1000 replications for which the the bivariate mode counting process detected 1, 2, 3, 4, 5 or more modes sample of size n from the mixture distribution. KDEs are adaptive with window-width selected by h_{NS}.

Mixture	n	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m > 5$
A	15	158	209	220	184	115	114
A	20	153	191	222	170	95	169
A	25	151	197	189	165	119	179
A	50	221	221	185	128	86	159
A	100	338	310	177	91	48	36
B	15	196	319	228	146	68	43
B	20	169	309	214	160	80	68
B	25	155	291	220	149	93	92
B	50	314	294	223	83	43	43
B	100	412	351	163	51	13	10
C	15	78	208	247	211	127	129
C	20	55	170	211	219	130	215
C	25	47	131	187	181	162	292
C	50	138	184	173	144	107	254
C	100	146	287	246	166	78	77
D	15	61	114	166	204	183	272
D	20	47	67	126	162	146	452
D	25	26	62	92	112	134	574
D	50	18	55	92	86	132	617
D	100	6	20	63	135	191	585
E	15	106	241	242	176	111	124
E	20	107	214	236	173	119	151
E	25	70	195	224	179	129	203
E	50	76	174	198	214	127	211
E	100	54	232	288	194	110	122
F	15	0	256	293	209	138	104
F	20	0	227	272	199	149	153
F	25	0	216	281	187	133	183
F	50	0	204	255	209	143	189
F	100	0	326	341	174	84	75
G	15	1	463	336	136	50	14
G	20	0	486	300	129	61	24
G	25	0	470	314	140	49	27
G	50	0	590	268	89	42	11
G	100	0	683	250	53	13	1
H	15	14	505	283	137	47	14
H	20	9	520	288	126	34	23
H	25	5	530	256	134	53	22
H	50	0	603	257	91	31	18
H	100	0	698	232	50	19	1

Mixture	n	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m > 5$
I	15	73	521	302	75	25	4
I	20	49	573	274	75	16	13
I	25	37	518	310	97	27	11
I	50	4	433	464	79	15	5
I	100	1	225	662	97	11	4
J	15	123	230	253	217	114	63
J	20	97	217	258	190	127	111
J	25	84	207	249	166	141	153
J	50	96	237	250	196	95	126
J	100	99	315	315	164	66	41
K	15	56	249	326	194	108	67
K	20	153	191	222	170	95	169
K	25	30	187	295	230	127	131
K	50	27	148	325	255	126	119
K	100	3	125	415	282	113	62
L	15	72	258	269	202	112	87
L	20	45	230	276	208	139	102
L	25	37	231	252	224	135	121
L	50	11	189	225	227	153	195
L	100	4	124	230	250	196	196

Table B.8: Results from the simulation study undertaken on bivariate mixture distributions detailed in appendix A. The table details the number of 1000 replications for which the the bivariate mode counting process detected 1, 2, 3, 4, 5 or more modes sample of size n from the mixture distribution. KDEs are non-adaptive with window-width selected by h_{STE}.

Mixture	n	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m > 5$
A	15	308	333	172	90	48	49
A	20	286	331	214	99	34	36
A	25	254	324	241	111	41	29
A	50	257	358	237	90	43	15
A	100	196	358	279	109	44	14
B	15	344	448	150	33	10	15
B	20	275	472	186	47	11	9
B	25	278	420	208	64	20	10
B	50	339	433	169	46	11	2
B	100	301	513	147	33	6	0
C	15	310	371	175	67	36	41
C	20	266	382	207	89	26	30
C	25	238	352	248	91	51	20
C	50	240	401	218	93	33	15
C	100	199	415	228	118	32	8
D	15	277	345	175	89	47	67
D	20	212	346	261	107	41	33
D	25	139	294	281	170	70	46
D	50	97	257	296	192	97	61
D	100	40	159	277	275	158	91
E	15	172	369	239	111	68	41
E	20	167	347	276	128	47	35
E	25	97	365	282	166	58	32
E	50	66	273	336	196	74	55
E	100	21	184	321	265	133	76
F	15	3	497	284	134	56	26
F	20	2	455	300	167	48	28
F	25	1	415	327	163	70	24
F	50	0	368	328	190	81	33
F	100	0	245	400	232	83	40
G	15	6	781	189	22	2	0
G	20	2	779	190	26	2	1
G	25	2	769	189	39	1	0
G	50	0	769	213	18	0	0
G	100	0	680	281	37	2	0
H	15	24	774	160	35	5	2
H	20	8	775	194	21	2	0
H	25	5	798	168	26	3	0
H	50	0	776	200	21	2	1
H	100	0	689	268	40	3	0

Mixture	n	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m > 5$
I	15	111	716	163	9	1	0
I	20	60	730	199	9	2	0
I	25	43	664	265	24	3	1
I	50	2	497	470	30	1	0
I	100	0	198	695	103	4	0
J	15	230	359	251	100	42	18
J	20	182	370	278	116	37	17
J	25	161	351	266	153	51	18
J	50	89	366	322	160	50	13
J	100	51	319	371	185	54	20
K	15	105	465	288	89	27	26
K	20	286	331	214	99	34	36
K	25	54	408	334	151	41	12
K	50	23	342	397	175	44	19
K	100	1	200	422	277	79	21
L	15	99	444	272	114	44	27
L	20	63	410	309	146	57	15
L	25	43	397	309	170	63	18
L	50	9	294	315	219	117	46
L	100	0	152	265	267	182	134

Table B.9: Results from the simulation study undertaken on bivariate mixture distributions detailed in appendix A. The table details the number of 1000 replications for which the the bivariate mode counting process detected 1, 2, 3, 4, 5 or more modes sample of size n from the mixture distribution. KDEs are adaptive with window-width selected by h_{STE}.

Mixture	n	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m > 5$
A	15	56	312	399	198	32	3
A	20	97	338	362	178	23	2
A	25	154	380	321	118	25	2
A	50	571	336	81	11	0	1
A	100	849	138	13	0	0	0
B	15	71	431	397	90	11	0
B	20	98	478	330	83	11	0
B	25	130	479	317	65	6	3
B	50	437	439	115	9	0	0
B	100	646	317	33	4	0	0
C	15	35	276	440	206	38	5
C	20	51	286	407	201	48	7
C	25	61	303	404	182	42	8
C	50	445	426	113	13	3	0
C	100	638	312	48	2	0	0
D	15	29	239	452	214	58	8
D	20	29	235	399	254	69	14
D	25	49	234	378	222	97	20
D	50	302	432	222	41	3	0
D	100	478	394	111	14	3	0
E	15	51	252	464	205	28	0
E	20	91	340	411	135	21	2
E	25	86	384	371	134	24	1
E	50	259	468	229	40	4	0
E	100	287	624	82	7	0	0
F	15	0	371	445	175	7	2
F	20	0	418	436	137	9	0
F	25	0	485	399	105	11	0
F	50	0	800	184	14	2	0
F	100	0	940	56	4	0	0
G	15	0	540	400	55	5	0
G	20	0	601	350	44	5	0
G	25	0	634	310	54	2	0
G	50	0	875	121	4	0	0
G	100	0	957	42	1	0	0
H	15	1	565	369	58	7	0
H	20	1	646	315	35	3	0
H	25	1	716	247	35	1	0
H	50	0	920	79	1	0	0
H	100	0	973	26	1	0	0

Mixture	n	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m > 5$
I	15	9	689	270	29	3	0
I	20	4	719	261	14	2	0
I	25	8	709	260	23	0	0
I	50	2	717	278	2	1	0
I	100	0	613	385	2	0	0
J	15	29	210	462	255	42	2
J	20	33	268	425	231	38	5
J	25	51	286	384	224	52	3
J	50	126	413	344	101	16	0
J	100	192	548	226	32	2	0
K	15	20	251	515	195	18	1
K	20	97	338	362	178	23	2
K	25	22	275	500	174	26	3
K	50	34	369	519	73	5	0
K	100	19	335	598	47	0	1
L	15	16	310	472	179	23	0
L	20	27	323	450	184	13	3
L	25	15	379	436	151	19	0
L	50	26	533	372	65	2	2
L	100	11	631	318	39	1	0

Table B.10: Results from the simulation study undertaken on bivariate mixture distributions detailed in appendix A. The table details the number of 1000 replications for which the the bivariate mode counting process detected 1, 2, 3, 4, 5 or more modes sample of size n from the mixture distribution. KDEs are non-adaptive with window-width selected by h_{DPI2}.

Mixture	n	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m > 5$
A	15	271	487	218	23	1	0
A	20	337	480	162	21	0	0
A	25	383	452	159	6	0	0
A	50	530	396	73	1	0	0
A	100	629	328	42	1	0	0
B	15	279	610	109	2	0	0
B	20	291	585	117	7	0	0
B	25	323	573	103	1	0	0
B	50	438	502	58	2	0	0
B	100	483	480	36	1	0	0
C	15	401	440	147	11	1	0
C	20	448	431	114	7	0	0
C	25	505	415	74	6	0	0
C	50	694	285	19	2	0	0
C	100	779	211	8	2	0	0
D	15	420	427	141	12	0	0
D	20	481	405	106	7	1	0
D	25	521	391	80	8	0	0
D	50	662	299	39	0	0	0
D	100	749	225	26	0	0	0
E	15	145	492	330	32	1	0
E	20	178	531	260	29	2	0
E	25	158	565	243	34	0	0
E	50	176	616	199	8	1	0
E	100	131	692	166	11	0	0
F	15	8	695	269	28	0	0
F	20	5	763	213	19	0	0
F	25	5	791	194	10	0	0
F	50	0	891	107	2	0	0
F	100	0	952	45	3	0	0
G	15	7	814	173	6	0	0
G	20	5	858	131	6	0	0
G	25	3	862	129	6	0	0
G	50	0	927	69	4	0	0
G	100	0	944	54	1	1	0
H	15	15	825	155	5	0	0
H	20	4	882	111	3	0	0
H	25	4	917	78	1	0	0
H	50	0	947	53	0	0	0
H	100	0	959	41	0	0	0

Mixture	n	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m > 5$
I	15	33	815	148	4	0	0
I	20	23	809	166	2	0	0
I	25	10	790	198	2	0	0
I	50	1	729	270	0	0	0
I	100	0	598	400	2	0	0
J	15	127	497	320	55	1	0
J	20	128	483	330	58	1	0
J	25	141	499	310	50	0	0
J	50	134	538	280	48	0	0
J	100	96	624	235	42	3	0
K	15	88	556	337	19	0	0
K	20	337	480	162	21	0	0
K	25	49	572	347	32	0	0
K	50	37	588	353	21	1	0
K	100	12	546	396	46	0	0
L	15	55	573	329	40	3	0
L	20	47	586	324	42	1	0
L	25	31	600	331	37	1	0
L	50	20	651	307	21	1	0
L	100	5	706	267	22	0	0

Table B.11: Results from the simulation study undertaken on bivariate mixture distributions detailed in appendix A. The table details the number of 1000 replications for which the the bivariate mode counting process detected 1, 2, 3, 4, 5 or more modes sample of size n from the mixture distribution. KDEs are adaptive with window-width selected by h_{DPI2}.

Appendix C

Results from Chapter 5

The following tables show results from the comparison of Silverman's (1981) test of modality with our adaptive modification to the original test (section 5.5). Results are presented as cumulative frequency tables which show the cumulative frequency of the resulting value of the test being $p = 0, \leq 0.1, \leq 0.2, \cdots \leq 1.0$ for the hypothesis of $k = 1$ and $k = 2$ when the data are simulated from a bimodal population and $k = 1$, $k = 2$ and $k = 3$ when the data are simulated from a trimodal population.

Test	Hypothesis	n	Count of p										
			$= 0$	≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	≤ 1.0
Silvermans	$H_0 : k = 1$	40	0	5	14	24	35	43	54	71	87	97	100
Silvermans	$H_0 : k = 2$	40	0	9	16	24	40	56	67	80	91	99	100
Adaptive	$H_0 : k = 1$	40	3	20	30	43	52	65	71	80	92	95	100
Adaptive	$H_0 : k = 2$	40	0	9	23	38	52	62	74	83	93	98	100
Silvermans	$H_0 : k = 1$	60	0	7	19	35	49	57	72	84	91	97	100
Silvermans	$H_0 : k = 2$	60	0	7	20	34	50	61	79	89	94	97	100
Adaptive	$H_0 : k = 1$	60	9	33	43	50	60	67	72	80	90	96	100
Adaptive	$H_0 : k = 2$	60	0	15	28	40	51	63	72	82	85	92	100
Silvermans	$H_0 : k = 1$	80	0	4	18	31	43	62	71	80	89	97	100
Silvermans	$H_0 : k = 2$	80	0	5	15	26	35	48	61	73	85	96	100
Adaptive	$H_0 : k = 1$	80	6	27	41	51	64	74	80	82	91	96	100
Adaptive	$H_0 : k = 2$	80	0	11	25	34	49	58	66	78	86	92	100
Silvermans	$H_0 : k = 1$	100	0	9	24	37	50	63	76	86	92	98	100
Silvermans	$H_0 : k = 2$	100	0	7	17	26	42	55	68	73	84	96	100
Adaptive	$H_0 : k = 1$	100	9	37	51	65	72	81	84	88	93	96	100
Adaptive	$H_0 : k = 2$	100	1	17	29	37	46	63	72	80	87	95	100

Table C.1: Cumulative frequency of the p value resulting from Silverman's (1981) test of modality and our adaptive modification to the original test (section 5.5). Results are from testing the hypothesis of unimodality and bimodality in samples of size n from the population $0.5N(0, 1) + 0.5N(2, 1)$.

147

Test	Hypothesis	n	$= 0$	≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	≤ 1.0
								Count of p					
Silvermans	$H_0 : k = 1$	40	0	23	39	53	61	68	77	84	89	96	100
Silvermans	$H_0 : k = 2$	40	0	4	15	30	34	45	61	75	87	98	100
Adaptive	$H_0 : k = 1$	40	11	48	59	65	75	78	83	86	94	97	100
Adaptive	$H_0 : k = 2$	40	0	10	21	37	44	56	65	73	87	94	100
Silvermans	$H_0 : k = 1$	60	0	18	32	49	65	75	80	89	95	99	100
Silvermans	$H_0 : k = 2$	60	0	4	11	26	35	49	62	75	86	92	100
Adaptive	$H_0 : k = 1$	60	15	37	51	60	68	73	79	88	93	97	100
Adaptive	$H_0 : k = 2$	60	0	12	24	32	44	60	72	80	90	97	100
Silvermans	$H_0 : k = 1$	80	0	19	35	46	62	70	80	91	95	99	100
Silvermans	$H_0 : k = 2$	80	0	9	18	27	42	61	70	83	89	96	100
Adaptive	$H_0 : k = 1$	80	16	49	56	63	73	76	83	88	94	98	100
Adaptive	$H_0 : k = 2$	80	0	11	23	35	49	61	73	82	90	95	100
Silvermans	$H_0 : k = 1$	100	0	27	40	57	68	73	78	88	93	96	100
Silvermans	$H_0 : k = 2$	100	0	6	14	30	47	56	66	76	88	95	100
Adaptive	$H_0 : k = 1$	100	20	50	59	64	69	77	83	88	89	95	100
Adaptive	$H_0 : k = 2$	100	3	16	27	41	49	59	70	78	89	95	100

Table C.2: Cumulative frequency of the p value resulting from Silverman's (1981) test of modality and our adaptive modification to the original test (section 5.5). Results are from testing the hypothesis of unimodality and bimodality in samples of size n from the population $0.5N(0,1) + 0.5N(2.5,1)$.

Test	Hypothesis	n	Count of p										
			$= 0$	≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	≤ 1.0
Silvermans	$H_0 : k = 1$	40	0	33	49	61	72	77	84	88	96	99	100
Silvermans	$H_0 : k = 2$	40	0	10	22	30	46	54	66	76	85	99	100
Adaptive	$H_0 : k = 1$	40	22	54	64	73	79	82	88	92	96	99	100
Adaptive	$H_0 : k = 2$	40	0	14	22	32	44	63	74	83	91	97	100
Silvermans	$H_0 : k = 1$	60	2	51	64	72	78	81	90	92	97	100	100
Silvermans	$H_0 : k = 2$	60	0	0	7	18	34	44	63	73	85	97	100
Adaptive	$H_0 : k = 1$	60	45	71	77	80	87	89	90	93	95	99	100
Adaptive	$H_0 : k = 2$	60	0	9	16	30	43	54	69	81	91	95	100
Silvermans	$H_0 : k = 1$	80	6	57	72	81	89	93	94	96	98	100	100
Silvermans	$H_0 : k = 2$	80	0	7	13	22	33	44	54	65	76	85	100
Adaptive	$H_0 : k = 1$	80	47	76	83	86	88	93	93	95	98	100	100
Adaptive	$H_0 : k = 2$	80	1	8	18	27	42	51	63	74	83	90	100
Silvermans	$H_0 : k = 1$	100	3	62	74	80	84	92	93	96	100	100	100
Silvermans	$H_0 : k = 2$	100	0	6	13	21	30	40	52	67	82	92	100
Adaptive	$H_0 : k = 1$	100	54	84	85	90	93	96	96	98	98	99	100
Adaptive	$H_0 : k = 2$	100	0	11	21	32	40	52	62	71	78	91	100

Table C.3: Cumulative frequency of the p value resulting from Silverman's (1981) test of modality and our adaptive modification to the original test (section 5.5). Results are from testing the hypothesis of unimodality and bimodality in samples of size n from the population $0.5N(0,1) + 0.5N(3,1)$.

Test	Hypothesis	n	Count of p										
			$=0$	≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	≤ 1.0
Silvermans	$H_0 : k = 1$	40	2	62	82	87	92	96	97	98	98	100	100
Silvermans	$H_0 : k = 2$	40	0	4	7	16	29	47	56	67	87	96	100
Adaptive	$H_0 : k = 1$	40	47	85	87	91	93	95	96	97	98	100	100
Adaptive	$H_0 : k = 2$	40	0	3	10	18	34	48	55	64	79	95	100
Silvermans	$H_0 : k = 1$	60	6	75	83	89	92	94	96	99	100	100	100
Silvermans	$H_0 : k = 2$	60	1	4	11	26	36	47	56	65	78	89	100
Adaptive	$H_0 : k = 1$	60	62	86	93	93	95	96	96	97	99	100	100
Adaptive	$H_0 : k = 2$	60	0	6	12	23	33	46	61	76	85	95	100
Silvermans	$H_0 : k = 1$	80	12	82	91	95	96	98	99	100	100	100	100
Silvermans	$H_0 : k = 2$	80	0	2	9	18	32	43	46	59	70	92	100
Adaptive	$H_0 : k = 1$	80	67	93	96	96	98	99	100	100	100	100	100
Adaptive	$H_0 : k = 2$	80	0	3	8	24	29	35	45	64	81	92	100
Silvermans	$H_0 : k = 1$	100	25	86	92	99	99	100	100	100	100	100	100
Silvermans	$H_0 : k = 2$	100	0	0	3	12	23	34	44	58	76	89	100
Adaptive	$H_0 : k = 1$	100	80	98	98	98	99	100	100	100	100	100	100
Adaptive	$H_0 : k = 2$	100	0	4	16	30	42	51	61	69	83	89	100

Table C.4: Cumulative frequency of the p value resulting from Silverman's (1981) test of modality and our adaptive modification to the original test (section 5.5). Results are from testing the hypothesis of unimodality and bimodality in samples of size n from the population $0.5N(0, 1) + 0.5N(3.5, 1)$.

Test	Hypothesis	n	$= 0$	≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	≤ 1.0
							Count of p						
Silvermans	$H_0 : k = 1$	40	2	71	88	93	95	97	100	100	100	100	100
Silvermans	$H_0 : k = 2$	40	0	2	5	13	23	36	44	61	81	95	100
Adaptive	$H_0 : k = 1$	40	54	91	95	96	99	99	100	100	100	100	100
Adaptive	$H_0 : k = 2$	40	0	3	10	15	31	43	56	72	89	95	100
Silvermans	$H_0 : k = 1$	60	18	86	95	97	99	99	100	100	100	100	100
Silvermans	$H_0 : k = 2$	60	0	0	6	12	20	29	43	57	78	90	100
Adaptive	$H_0 : k = 1$	60	80	96	98	98	99	99	99	99	100	100	100
Adaptive	$H_0 : k = 2$	60	0	3	9	19	29	40	55	66	78	91	100
Silvermans	$H_0 : k = 1$	80	34	98	99	100	100	100	100	100	100	100	100
Silvermans	$H_0 : k = 2$	80	0	1	4	12	21	32	45	57	74	88	100
Adaptive	$H_0 : k = 1$	80	93	100	100	100	100	100	100	100	100	100	100
Adaptive	$H_0 : k = 2$	80	0	3	14	17	24	39	48	62	73	86	100
Silvermans	$H_0 : k = 1$	100	45	99	100	100	100	100	100	100	100	100	100
Silvermans	$H_0 : k = 2$	100	0	0	5	9	13	26	42	52	68	86	100
Adaptive	$H_0 : k = 1$	100	96	100	100	100	100	100	100	100	100	100	100
Adaptive	$H_0 : k = 2$	100	0	5	11	24	34	40	56	65	77	92	100

Table C.5: Cumulative frequency of the p value resulting from Silverman's (1981) test of modality and our adaptive modification to the original test (section 5.5). Results are from testing the hypothesis of unimodality and bimodality in samples of size n from the population $0.5N(0, 1) + 0.5N(4, 1)$.

| Test | Hypothesis | n | Count of p |
			$= 0$	≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	≤ 1.0
Silvermans	$H_0 : k = 1$	40	1	35	70	88	93	95	98	98	99	100	100
Silvermans	$H_0 : k = 2$	40	1	61	74	82	85	89	92	97	100	100	100
Silvermans	$H_0 : k = 3$	40	0	4	12	22	31	38	54	66	81	93	100
Adaptive	$H_0 : k = 1$	40	32	76	88	94	95	97	97	97	97	99	100
Adaptive	$H_0 : k = 2$	40	15	69	77	87	90	93	98	98	100	100	100
Adaptive	$H_0 : k = 3$	40	0	0	5	19	24	29	39	62	76	93	100
Silvermans	$H_0 : k = 1$	60	2	43	72	90	96	99	100	100	100	100	100
Silvermans	$H_0 : k = 2$	60	12	70	88	92	96	97	97	99	100	100	100
Silvermans	$H_0 : k = 3$	60	0	3	7	13	23	32	42	64	75	93	100
Adaptive	$H_0 : k = 1$	60	46	85	91	94	95	98	99	99	100	100	100
Adaptive	$H_0 : k = 2$	60	25	76	86	93	95	96	99	99	99	100	100
Adaptive	$H_0 : k = 3$	60	0	1	6	12	24	36	47	61	75	87	100
Silvermans	$H_0 : k = 1$	80	4	49	81	96	99	100	100	100	100	100	100
Silvermans	$H_0 : k = 2$	80	22	85	92	95	98	99	100	100	100	100	100
Silvermans	$H_0 : k = 3$	80	0	3	9	16	22	26	38	49	69	87	100
Adaptive	$H_0 : k = 1$	80	52	94	98	100	100	100	100	100	100	100	100
Adaptive	$H_0 : k = 2$	80	52	87	94	97	98	99	99	100	100	100	100
Adaptive	$H_0 : k = 3$	80	0	1	5	9	19	32	41	56	70	86	100
Silvermans	$H_0 : k = 1$	100	7	62	95	100	100	100	100	100	100	100	100
Silvermans	$H_0 : k = 2$	100	32	94	97	97	99	99	99	99	99	100	100
Silvermans	$H_0 : k = 3$	100	0	0	7	12	15	27	38	53	76	92	100
Adaptive	$H_0 : k = 1$	100	63	93	100	100	100	100	100	100	100	100	100
Adaptive	$H_0 : k = 2$	100	61	94	97	99	99	99	99	99	100	100	100
Adaptive	$H_0 : k = 3$	100	0	0	2	8	16	27	41	58	74	84	100

Table C.6: Cumulative frequency of the p value resulting from Silverman's (1981) test of modality and our adaptive modification to the original test (section 5.5). Results are from testing the hypothesis of unimodality, bimodality and trimodality in samples of size n from the population $\frac{1}{3}N(4, 1) + \frac{1}{3}N(0, 1) + \frac{1}{3}N(4, 1)$.

Test	Hypothesis	n	Count of p										
			$= 0$	≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	≤ 1.0
Silvermans	$H_0 : k = 1$	40	2	40	66	83	91	97	100	100	100	100	100
Silvermans	$H_0 : k = 2$	40	3	46	63	73	81	86	90	95	98	100	100
Silvermans	$H_0 : k = 3$	40	0	5	13	22	26	40	55	71	85	96	100
Adaptive	$H_0 : k = 1$	40	31	69	76	84	88	93	97	100	100	100	100
Adaptive	$H_0 : k = 2$	40	8	50	69	76	84	89	93	95	97	99	100
Adaptive	$H_0 : k = 3$	40	0	7	11	21	31	41	54	70	81	94	100
Silvermans	$H_0 : k = 1$	60	3	42	70	91	95	99	100	100	100	100	100
Silvermans	$H_0 : k = 2$	60	4	54	70	78	83	87	91	95	98	98	100
Silvermans	$H_0 : k = 3$	60	0	3	8	17	25	39	51	58	73	95	100
Adaptive	$H_0 : k = 1$	60	31	75	80	89	91	95	98	98	99	100	100
Adaptive	$H_0 : k = 2$	60	12	63	69	75	81	87	90	96	98	99	100
Adaptive	$H_0 : k = 3$	60	0	4	8	13	23	32	43	62	76	91	100
Silvermans	$H_0 : k = 1$	80	0	51	86	97	100	100	100	100	100	100	100
Silvermans	$H_0 : k = 2$	80	9	67	80	86	88	93	98	98	100	100	100
Silvermans	$H_0 : k = 3$	80	0	2	7	15	26	35	48	55	70	83	100
Adaptive	$H_0 : k = 1$	80	46	83	90	97	99	100	100	100	100	100	100
Adaptive	$H_0 : k = 2$	80	37	77	79	86	94	96	98	99	100	100	100
Adaptive	$H_0 : k = 3$	80	0	1	12	12	19	33	41	50	64	80	100
Silvermans	$H_0 : k = 1$	100	2	55	88	95	99	100	100	100	100	100	100
Silvermans	$H_0 : k = 2$	100	19	77	88	93	93	97	98	99	99	100	100
Silvermans	$H_0 : k = 3$	100	0	1	7	13	23	29	40	57	71	90	100
Adaptive	$H_0 : k = 1$	100	48	92	95	98	99	99	100	100	100	100	100
Adaptive	$H_0 : k = 2$	100	57	85	88	93	93	96	97	98	99	99	100
Adaptive	$H_0 : k = 3$	100	0	2	3	11	22	36	40	53	66	80	100

Table C.7: Cumulative frequency of the p value resulting from Silverman's (1981) test of modality and our adaptive modification to the original test (section 5.5). Results are from testing the hypothesis of unimodality, bimodality and trimodality in samples of size n from the population $\frac{1}{3}N(3.5, 1) + \frac{1}{3}N(0, 1) + \frac{1}{3}N(4, 1)$.

Test	Hypothesis	n	$= 0$	≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	≤ 1.0
Silvermans	$H_0 : k = 1$	40	0	40	67	78	85	93	98	99	100	100	100
Silvermans	$H_0 : k = 2$	40	1	26	39	53	61	71	79	87	96	99	100
Silvermans	$H_0 : k = 3$	40	0	2	11	22	30	44	58	72	85	96	100
Adaptive	$H_0 : k = 1$	40	22	61	73	80	84	88	92	96	99	100	100
Adaptive	$H_0 : k = 2$	40	4	30	45	64	72	82	85	90	95	99	100
Adaptive	$H_0 : k = 3$	40	0	6	8	16	26	37	49	61	76	93	100
Silvermans	$H_0 : k = 1$	60	2	46	80	94	96	98	98	98	99	100	100
Silvermans	$H_0 : k = 2$	60	4	34	45	54	77	80	85	92	95	97	100
Silvermans	$H_0 : k = 3$	60	0	6	13	21	31	42	58	69	82	94	100
Adaptive	$H_0 : k = 1$	60	30	73	85	89	93	94	95	97	98	100	100
Adaptive	$H_0 : k = 2$	60	17	41	53	65	75	78	86	91	96	98	100
Adaptive	$H_0 : k = 3$	60	0	2	11	19	30	38	55	65	79	90	100
Silvermans	$H_0 : k = 1$	80	8	69	85	95	99	99	100	100	100	100	100
Silvermans	$H_0 : k = 2$	80	1	42	60	68	77	84	85	93	95	97	100
Silvermans	$H_0 : k = 3$	80	0	3	9	17	28	39	52	63	75	90	100
Adaptive	$H_0 : k = 1$	80	48	85	91	93	94	97	99	99	100	100	100
Adaptive	$H_0 : k = 2$	80	10	57	63	72	79	87	89	91	93	95	100
Adaptive	$H_0 : k = 3$	80	0	2	4	9	16	25	35	59	72	84	100
Silvermans	$H_0 : k = 1$	100	5	73	93	98	99	100	100	100	100	100	100
Silvermans	$H_0 : k = 2$	100	3	39	63	71	79	86	91	92	94	98	100
Silvermans	$H_0 : k = 3$	100	0	4	15	23	32	41	50	65	70	87	100
Adaptive	$H_0 : k = 1$	100	56	90	95	97	100	100	100	100	100	100	100
Adaptive	$H_0 : k = 2$	100	17	58	72	79	84	88	90	93	97	97	100
Adaptive	$H_0 : k = 3$	100	0	4	10	17	27	39	50	67	81	90	100

Table C.8: Cumulative frequency of the p value resulting from Silverman's (1981) test of modality and our adaptive modification to the original test (section 5.5). Results are from testing the hypothesis of unimodality, bimodality and trimodality in samples of size n from the population $\frac{1}{3}N(3, 1) + \frac{1}{3}N(0, 1) + \frac{1}{3}N(4, 1)$.

154

Test	Hypothesis	n	Count of p										
			$= 0$	≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	≤ 1.0
Silvermans	$H_0 : k = 1$	40	2	60	76	87	93	96	97	98	99	99	100
Silvermans	$H_0 : k = 2$	40	0	12	17	29	35	49	66	78	83	93	100
Silvermans	$H_0 : k = 3$	40	0	3	9	19	31	43	64	74	90	96	100
Adaptive	$H_0 : k = 1$	40	34	69	77	82	87	95	95	96	98	100	100
Adaptive	$H_0 : k = 2$	40	1	14	30	44	54	65	75	84	93	98	100
Adaptive	$H_0 : k = 3$	40	0	5	13	24	31	40	60	71	84	94	100
Silvermans	$H_0 : k = 1$	60	4	63	85	92	96	97	99	99	99	99	100
Silvermans	$H_0 : k = 2$	60	0	13	23	34	45	60	67	73	85	96	100
Silvermans	$H_0 : k = 3$	60	0	3	8	17	32	46	61	70	85	95	100
Adaptive	$H_0 : k = 1$	60	40	82	85	93	95	97	97	98	99	100	100
Adaptive	$H_0 : k = 2$	60	3	24	37	47	60	65	76	80	86	94	100
Adaptive	$H_0 : k = 3$	60	0	6	10	17	30	40	54	72	80	91	100
Silvermans	$H_0 : k = 1$	80	4	78	93	97	100	100	100	100	100	100	100
Silvermans	$H_0 : k = 2$	80	0	14	26	36	50	58	69	81	86	93	100
Silvermans	$H_0 : k = 3$	80	0	4	11	22	34	43	54	67	83	94	100
Adaptive	$H_0 : k = 1$	80	54	89	95	97	97	100	100	100	100	100	100
Adaptive	$H_0 : k = 2$	80	7	27	44	47	60	71	80	88	94	95	100
Adaptive	$H_0 : k = 3$	80	0	4	14	22	31	44	54	64	81	90	100
Silvermans	$H_0 : k = 1$	100	6	87	97	100	100	100	100	100	100	100	100
Silvermans	$H_0 : k = 2$	100	2	16	29	38	52	68	76	79	87	95	100
Silvermans	$H_0 : k = 3$	100	0	5	15	19	29	46	57	73	85	94	100
Adaptive	$H_0 : k = 1$	100	61	93	97	98	99	100	100	100	100	100	100
Adaptive	$H_0 : k = 2$	100	5	33	44	58	65	72	75	82	89	96	100
Adaptive	$H_0 : k = 3$	100	0	11	18	24	30	45	56	65	73	93	100

Table C.9: Cumulative frequency of the p value resulting from Silverman's (1981) test of modality and our adaptive modification to the original test (section 5.5). Results are from testing the hypothesis of unimodality, bimodality and trimodality in samples of size n from the population $\frac{1}{3}N(2.5, 1) + \frac{1}{3}N(0, 1) + \frac{1}{3}N(4, 1)$.

Test	Hypothesis	n	Count of p										
			$= 0$	≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	≤ 1.0
Silvermans	$H_0 : k = 1$	40	0	63	82	93	96	96	97	97	98	100	100
Silvermans	$H_0 : k = 2$	40	0	2	3	13	24	37	51	67	84	95	100
Silvermans	$H_0 : k = 3$	40	0	3	11	20	42	55	66	75	88	98	100
Adaptive	$H_0 : k = 1$	40	35	73	85	87	91	94	95	98	100	100	100
Adaptive	$H_0 : k = 2$	40	0	9	23	34	45	60	72	82	87	95	100
Adaptive	$H_0 : k = 3$	40	0	2	9	17	31	48	66	79	88	93	100
Silvermans	$H_0 : k = 1$	60	1	70	88	95	97	99	100	100	100	100	100
Silvermans	$H_0 : k = 2$	60	0	2	13	20	25	38	49	63	82	91	100
Silvermans	$H_0 : k = 3$	60	0	6	19	28	44	53	63	74	87	94	100
Adaptive	$H_0 : k = 1$	60	42	85	92	96	96	97	98	99	100	100	100
Adaptive	$H_0 : k = 2$	60	1	14	26	36	43	56	64	71	81	95	100
Adaptive	$H_0 : k = 3$	60	0	5	13	21	26	36	44	59	77	92	100
Silvermans	$H_0 : k = 1$	80	5	88	96	97	100	100	100	100	100	100	100
Silvermans	$H_0 : k = 2$	80	0	2	8	19	28	44	58	65	77	94	100
Silvermans	$H_0 : k = 3$	80	0	3	13	20	29	43	60	70	82	93	100
Adaptive	$H_0 : k = 1$	80	52	96	97	98	99	99	100	100	100	100	100
Adaptive	$H_0 : k = 2$	80	1	17	28	39	54	61	69	79	82	91	100
Adaptive	$H_0 : k = 3$	80	0	6	19	24	30	41	54	64	80	89	100
Silvermans	$H_0 : k = 1$	100	8	89	98	98	98	100	100	100	100	100	100
Silvermans	$H_0 : k = 2$	100	0	5	11	21	31	44	49	62	72	91	100
Silvermans	$H_0 : k = 3$	100	0	3	8	15	27	41	50	65	80	94	100
Adaptive	$H_0 : k = 1$	100	69	96	97	98	98	98	100	100	100	100	100
Adaptive	$H_0 : k = 2$	100	0	16	33	37	50	58	66	72	81	93	100
Adaptive	$H_0 : k = 3$	100	0	3	8	13	25	35	44	60	72	78	100

Table C.10: Cumulative frequency of the p value resulting from Silverman's (1981) test of modality and our adaptive modification to the original test (section 5.5). Results are from testing the hypothesis of unimodality, bimodality and trimodality in samples of size n from the population $\frac{1}{3}N(2,1) + \frac{1}{3}N(0,1) + \frac{1}{3}N(4,1)$.

Test	Hypothesis	n	$= 0$	≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	≤ 1.0
									Count of p				
Silvermans	$H_0 : k = 1$	40	2	28	51	69	87	91	97	100	100	100	100
Silvermans	$H_0 : k = 2$	40	0	41	60	72	76	80	86	92	94	99	100
Silvermans	$H_0 : k = 3$	40	0	9	17	25	37	44	58	65	80	97	100
Adaptive	$H_0 : k = 1$	40	20	59	67	75	80	87	91	94	95	99	100
Adaptive	$H_0 : k = 2$	40	9	39	55	67	74	77	85	88	91	98	100
Adaptive	$H_0 : k = 3$	40	0	5	15	21	31	43	56	68	77	90	100
Silvermans	$H_0 : k = 1$	60	3	38	61	84	92	96	99	100	100	100	100
Silvermans	$H_0 : k = 2$	60	3	50	65	78	85	89	91	96	97	99	100
Silvermans	$H_0 : k = 3$	60	0	4	14	23	36	51	59	66	83	95	100
Adaptive	$H_0 : k = 1$	60	31	65	77	87	90	94	95	97	98	99	100
Adaptive	$H_0 : k = 2$	60	13	55	67	78	82	86	90	94	96	97	100
Adaptive	$H_0 : k = 3$	60	0	4	11	22	35	52	58	70	81	93	100
Silvermans	$H_0 : k = 1$	80	5	44	68	89	93	94	96	98	99	99	100
Silvermans	$H_0 : k = 2$	80	12	59	76	81	87	91	94	96	97	99	100
Silvermans	$H_0 : k = 3$	80	0	6	19	26	38	49	57	68	79	92	100
Adaptive	$H_0 : k = 1$	80	37	76	86	90	92	94	95	95	98	99	100
Adaptive	$H_0 : k = 2$	80	23	64	71	80	85	89	89	90	94	98	100
Adaptive	$H_0 : k = 3$	80	0	10	15	23	29	38	58	69	77	92	100
Silvermans	$H_0 : k = 1$	100	5	46	76	91	98	100	100	100	100	100	100
Silvermans	$H_0 : k = 2$	100	11	69	84	90	93	94	97	97	98	99	100
Silvermans	$H_0 : k = 3$	100	0	3	10	16	23	35	52	61	74	86	100
Adaptive	$H_0 : k = 1$	100	48	85	90	96	99	99	99	100	100	100	100
Adaptive	$H_0 : k = 2$	100	36	75	85	87	91	94	96	97	99	100	100
Adaptive	$H_0 : k = 3$	100	0	2	5	13	20	30	43	56	73	88	100

Table C.11: Cumulative frequency of the p value resulting from Silverman's (1981) test of modality and our adaptive modification to the original test (section 5.5). Results are from testing the hypothesis of unimodality, bimodality and trimodality in samples of size n from the population $\frac{1}{3}N(3.5, 1) + \frac{1}{3}N(0, 1) + \frac{1}{3}N(3.5, 1)$.

Test	Hypothesis	n	Count of p										
			$=0$	≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	≤ 1.0
Silvermans	$H_0 : k = 1$	40	0	37	64	77	86	91	93	95	97	100	100
Silvermans	$H_0 : k = 2$	40	0	29	42	48	59	68	79	87	94	98	100
Silvermans	$H_0 : k = 3$	40	0	2	8	14	27	41	56	65	81	96	100
Adaptive	$H_0 : k = 1$	40	22	62	67	78	82	87	90	92	93	99	100
Adaptive	$H_0 : k = 2$	40	1	30	42	52	64	72	78	91	95	97	100
Adaptive	$H_0 : k = 3$	40	0	1	8	18	27	40	50	61	79	91	100
Silvermans	$H_0 : k = 1$	60	0	32	56	71	82	93	96	98	99	99	100
Silvermans	$H_0 : k = 2$	60	2	44	56	71	77	83	87	92	96	99	100
Silvermans	$H_0 : k = 3$	60	0	5	11	16	25	43	53	72	81	95	100
Adaptive	$H_0 : k = 1$	60	23	54	66	75	80	86	94	97	99	100	100
Adaptive	$H_0 : k = 2$	60	7	51	62	70	81	87	91	92	94	97	100
Adaptive	$H_0 : k = 3$	60	0	3	12	21	31	43	60	71	78	89	100
Silvermans	$H_0 : k = 1$	80	4	44	69	84	94	96	99	99	99	99	100
Silvermans	$H_0 : k = 2$	80	2	47	63	69	74	82	85	89	95	99	100
Silvermans	$H_0 : k = 3$	80	0	4	16	25	34	45	61	71	82	89	100
Adaptive	$H_0 : k = 1$	80	29	76	80	86	92	96	98	98	100	100	100
Adaptive	$H_0 : k = 2$	80	14	53	64	72	78	83	87	91	94	98	100
Adaptive	$H_0 : k = 3$	80	0	5	9	23	32	41	54	69	80	89	100
Silvermans	$H_0 : k = 1$	100	4	50	75	89	91	96	98	99	100	100	100
Silvermans	$H_0 : k = 2$	100	4	43	57	66	73	81	88	93	96	99	100
Silvermans	$H_0 : k = 3$	100	0	6	14	23	27	39	49	62	79	92	100
Adaptive	$H_0 : k = 1$	100	33	72	84	85	90	93	96	97	99	100	100
Adaptive	$H_0 : k = 2$	100	12	54	60	67	75	79	83	87	93	96	100
Adaptive	$H_0 : k = 3$	100	0	3	10	22	30	36	47	61	72	91	100

Table C.12: Cumulative frequency of the p value resulting from Silverman's (1981) test of modality and our adaptive modification to the original test (section 5.5). Results are from testing the hypothesis of unimodality, bimodality and trimodality in samples of size n from the population $\frac{1}{3}N(3,1) + \frac{1}{3}N(0,1) + \frac{1}{3}N(3.5,1)$.

158

Test	Hypothesis	n	Count of p										
			$= 0$	≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	≤ 1.0
Silvermans	$H_0 : k = 1$	40	4	30	49	67	77	85	89	93	98	99	100
Silvermans	$H_0 : k = 2$	40	0	10	23	39	52	66	75	84	92	98	100
Silvermans	$H_0 : k = 3$	40	0	5	15	25	44	55	67	80	88	96	100
Adaptive	$H_0 : k = 1$	40	18	49	62	71	76	82	90	92	96	98	100
Adaptive	$H_0 : k = 2$	40	1	13	29	44	53	63	73	82	93	97	100
Adaptive	$H_0 : k = 3$	40	0	5	15	26	36	48	56	70	88	97	100
Silvermans	$H_0 : k = 1$	60	2	47	65	79	87	92	96	98	100	100	100
Silvermans	$H_0 : k = 2$	60	1	23	32	44	53	61	72	79	89	95	100
Silvermans	$H_0 : k = 3$	60	0	7	15	22	30	44	59	71	79	92	100
Adaptive	$H_0 : k = 1$	60	29	67	72	78	84	89	92	95	99	99	100
Adaptive	$H_0 : k = 2$	60	4	27	41	54	62	69	73	82	90	94	100
Adaptive	$H_0 : k = 3$	60	0	4	9	21	30	42	54	71	79	91	100
Silvermans	$H_0 : k = 1$	80	2	59	79	89	95	97	99	100	100	100	100
Silvermans	$H_0 : k = 2$	80	2	14	28	43	53	63	75	82	90	98	100
Silvermans	$H_0 : k = 3$	80	0	3	12	24	38	51	65	75	84	96	100
Adaptive	$H_0 : k = 1$	80	38	74	79	82	88	92	95	97	99	99	100
Adaptive	$H_0 : k = 2$	80	4	26	40	54	69	74	81	86	91	94	100
Adaptive	$H_0 : k = 3$	80	0	4	12	21	29	44	61	73	91	91	100
Silvermans	$H_0 : k = 1$	100	4	57	80	91	95	98	98	98	99	99	100
Silvermans	$H_0 : k = 2$	100	1	21	31	44	52	58	65	74	79	94	100
Silvermans	$H_0 : k = 3$	100	0	2	9	20	32	42	53	63	80	93	100
Adaptive	$H_0 : k = 1$	100	31	72	84	88	91	94	96	98	99	99	100
Adaptive	$H_0 : k = 2$	100	9	33	44	52	57	63	69	78	84	92	100
Adaptive	$H_0 : k = 3$	100	0	1	5	18	28	38	49	66	75	85	100

Table C.13: Cumulative frequency of the p value resulting from Silverman's (1981) test of modality and our adaptive modification to the original test (section 5.5). Results are from testing the hypothesis of unimodality, bimodality and trimodality in samples of size n from the population $\frac{1}{3}N(2.5, 1) + \frac{1}{3}N(0, 1) + \frac{1}{3}N(3.5, 1)$.

Table C.14: Cumulative frequency of the p value resulting from Silverman's (1981) test of modality and our adaptive modification to the original test (section 5.5). Results are from testing the hypothesis of unimodality, bimodality and trimodality in samples of size n from the population $\frac{1}{3}N(2,1) + \frac{1}{3}N(0,1) + \frac{1}{3}N(3.5,1)$.

Test	Hypothesis	n	Count of p										
			$= 0$	≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	≤ 1.0
Silvermans	$H_0 : k = 1$	40	3	40	61	76	84	87	88	93	97	100	100
Silvermans	$H_0 : k = 2$	40	0	6	16	33	45	62	72	77	89	94	100
Silvermans	$H_0 : k = 3$	40	0	7	23	35	43	53	65	80	89	96	100
Adaptive	$H_0 : k = 1$	40	23	57	65	77	82	82	86	88	92	96	100
Adaptive	$H_0 : k = 2$	40	1	20	28	43	55	65	73	81	93	98	100
Adaptive	$H_0 : k = 3$	40	0	7	22	31	42	51	63	72	83	95	100
Silvermans	$H_0 : k = 1$	60	3	57	75	84	90	93	94	96	99	99	100
Silvermans	$H_0 : k = 2$	60	0	5	11	15	28	35	40	54	68	86	100
Silvermans	$H_0 : k = 3$	60	0	0	7	16	24	37	49	67	83	91	100
Adaptive	$H_0 : k = 1$	60	32	74	82	86	91	95	97	97	98	100	100
Adaptive	$H_0 : k = 2$	60	0	7	14	29	40	52	59	71	79	89	100
Adaptive	$H_0 : k = 3$	60	0	3	19	29	42	49	62	71	82	95	100
Silvermans	$H_0 : k = 1$	80	1	60	81	87	92	97	97	98	99	99	100
Silvermans	$H_0 : k = 2$	80	0	6	12	22	34	48	56	72	84	91	100
Silvermans	$H_0 : k = 3$	80	0	5	13	19	26	35	50	65	79	91	100
Adaptive	$H_0 : k = 1$	80	30	75	82	85	89	91	96	96	98	100	100
Adaptive	$H_0 : k = 2$	80	2	16	25	34	45	54	61	69	76	91	100
Adaptive	$H_0 : k = 3$	80	0	5	11	20	27	41	49	63	77	87	100
Silvermans	$H_0 : k = 1$	100	4	68	87	93	96	96	96	96	99	99	100
Silvermans	$H_0 : k = 2$	100	0	10	18	27	37	51	65	71	80	94	100
Silvermans	$H_0 : k = 3$	100	0	1	10	18	30	38	54	64	77	91	100
Adaptive	$H_0 : k = 1$	100	46	80	86	90	92	93	97	97	98	98	100
Adaptive	$H_0 : k = 2$	100	5	21	30	42	53	59	68	76	84	91	100
Adaptive	$H_0 : k = 3$	100	0	5	10	18	33	41	51	58	70	87	100

Test	Hypothesis	n	Count of p										
			$= 0$	≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	≤ 1.0
Silvermans	$H_0 : k = 1$	40	1	31	46	57	67	77	87	94	97	99	100
Silvermans	$H_0 : k = 2$	40	0	17	29	41	56	63	77	85	92	98	100
Silvermans	$H_0 : k = 3$	40	0	5	22	32	44	51	66	75	88	96	100
Adaptive	$H_0 : k = 1$	40	19	41	51	58	70	79	85	89	93	100	100
Adaptive	$H_0 : k = 2$	40	0	19	31	43	57	65	75	81	89	98	100
Adaptive	$H_0 : k = 3$	40	0	4	19	29	41	57	67	76	89	97	100
Silvermans	$H_0 : k = 1$	60	1	31	48	67	75	79	83	91	94	97	100
Silvermans	$H_0 : k = 2$	60	0	25	42	53	59	69	75	84	90	98	100
Silvermans	$H_0 : k = 3$	60	0	4	18	28	40	54	64	79	88	97	100
Adaptive	$H_0 : k = 1$	60	21	52	62	65	74	82	87	90	97	98	100
Adaptive	$H_0 : k = 2$	60	3	25	39	48	60	69	72	80	87	94	100
Adaptive	$H_0 : k = 3$	60	0	6	14	26	42	56	65	79	88	96	100
Silvermans	$H_0 : k = 1$	80	2	31	56	70	80	92	92	97	98	100	100
Silvermans	$H_0 : k = 2$	80	3	41	59	68	72	81	90	95	98	99	100
Silvermans	$H_0 : k = 3$	80	0	4	11	21	30	42	55	69	86	95	100
Adaptive	$H_0 : k = 1$	80	24	59	70	78	84	89	91	94	95	99	100
Adaptive	$H_0 : k = 2$	80	5	35	49	62	70	81	85	92	96	98	100
Adaptive	$H_0 : k = 3$	80	0	4	11	22	34	44	51	68	85	95	100
Silvermans	$H_0 : k = 1$	100	1	27	54	67	78	88	92	97	99	100	100
Silvermans	$H_0 : k = 2$	100	4	46	61	72	79	85	87	92	96	99	100
Silvermans	$H_0 : k = 3$	100	0	6	10	21	36	44	59	74	83	91	100
Adaptive	$H_0 : k = 1$	100	19	58	65	69	74	79	83	91	95	96	100
Adaptive	$H_0 : k = 2$	100	14	43	58	63	70	78	85	92	96	99	100
Adaptive	$H_0 : k = 3$	100	0	3	10	23	34	52	63	70	79	92	100

Table C.15: Cumulative frequency of the p value resulting from Silverman's (1981) test of modality and our adaptive modification to the original test (section 5.5). Results are from testing the hypothesis of unimodality, bimodality and trimodality in samples of size n from the population $\frac{1}{3}N(3,1) + \frac{1}{3}N(0,1) + \frac{1}{3}N(3,1)$.

| | | | | Count of p | | | | | | | | | | |
Test	Hypothesis	n	$= 0$	≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	≤ 1.0
Silvermans	$H_0 : k = 1$	40	0	23	44	58	72	84	91	95	96	99	100
Silvermans	$H_0 : k = 2$	40	0	11	23	31	38	54	65	74	86	95	100
Silvermans	$H_0 : k = 3$	40	0	5	14	28	42	56	67	75	86	96	100
Adaptive	$H_0 : k = 1$	40	17	47	59	64	70	79	85	89	92	97	100
Adaptive	$H_0 : k = 2$	40	0	12	25	32	44	57	73	80	90	98	100
Adaptive	$H_0 : k = 3$	40	0	6	17	30	43	55	67	80	89	94	100
Silvermans	$H_0 : k = 1$	60	0	27	50	62	73	81	89	94	96	97	100
Silvermans	$H_0 : k = 2$	60	1	14	35	44	56	62	76	84	89	93	100
Silvermans	$H_0 : k = 3$	60	0	5	15	26	37	52	61	74	86	95	100
Adaptive	$H_0 : k = 1$	60	15	49	58	65	70	74	82	88	91	99	100
Adaptive	$H_0 : k = 2$	60	1	25	33	49	57	68	75	86	91	98	100
Adaptive	$H_0 : k = 3$	60	0	6	15	24	39	49	60	71	81	94	100
Silvermans	$H_0 : k = 1$	80	1	29	51	70	84	92	93	96	97	100	100
Silvermans	$H_0 : k = 2$	80	0	29	34	49	60	73	80	86	93	97	100
Silvermans	$H_0 : k = 3$	80	0	9	19	30	42	51	58	71	86	96	100
Adaptive	$H_0 : k = 1$	80	14	53	63	70	79	89	91	94	98	99	100
Adaptive	$H_0 : k = 2$	80	6	27	38	47	54	64	74	86	89	93	100
Adaptive	$H_0 : k = 3$	80	0	6	18	22	34	42	56	70	84	93	100
Silvermans	$H_0 : k = 1$	100	2	31	50	67	78	86	91	95	98	100	100
Silvermans	$H_0 : k = 2$	100	0	25	40	49	59	71	78	83	91	94	100
Silvermans	$H_0 : k = 3$	100	0	8	20	30	44	61	71	77	88	94	100
Adaptive	$H_0 : k = 1$	100	17	52	59	64	76	78	87	92	94	98	100
Adaptive	$H_0 : k = 2$	100	6	33	42	55	62	69	79	83	88	95	100
Adaptive	$H_0 : k = 3$	100	0	7	20	27	40	49	58	68	78	92	100

Table C.16: Cumulative frequency of the p value resulting from Silverman's (1981) test of modality and our adaptive modification to the original test (section 5.5). Results are from testing the hypothesis of unimodality, bimodality and trimodality in samples of size n from the population $\frac{1}{3}N(2.5, 1) + \frac{1}{3}N(0, 1) + \frac{1}{3}N(3, 1)$.

Test	Hypothesis	n	$= 0$	≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	≤ 1.0
							Count of p						
Silvermans	$H_0 : k = 1$	40	0	32	59	68	77	85	88	91	94	97	100
Silvermans	$H_0 : k = 2$	40	1	8	19	27	40	52	61	72	80	92	100
Silvermans	$H_0 : k = 3$	40	0	6	17	27	40	54	65	76	85	97	100
Adaptive	$H_0 : k = 1$	40	19	54	66	74	77	84	88	91	94	98	100
Adaptive	$H_0 : k = 2$	40	3	21	26	35	41	51	67	76	85	95	100
Adaptive	$H_0 : k = 3$	40	0	3	8	17	35	46	64	77	91	98	100
Silvermans	$H_0 : k = 1$	60	0	30	51	64	69	78	83	87	96	96	100
Silvermans	$H_0 : k = 2$	60	1	9	21	35	46	58	67	79	86	93	100
Silvermans	$H_0 : k = 3$	60	0	4	16	23	37	48	65	79	85	96	100
Adaptive	$H_0 : k = 1$	60	14	50	58	68	73	77	84	88	93	97	100
Adaptive	$H_0 : k = 2$	60	1	20	33	44	57	66	75	83	89	95	100
Adaptive	$H_0 : k = 3$	60	0	2	11	29	38	46	61	71	85	93	100
Silvermans	$H_0 : k = 1$	80	0	30	45	57	73	80	89	93	94	97	100
Silvermans	$H_0 : k = 2$	80	0	11	27	39	50	58	73	84	95	100	100
Silvermans	$H_0 : k = 3$	80	0	11	23	36	48	53	60	71	92	96	100
Adaptive	$H_0 : k = 1$	80	19	47	56	65	73	78	82	88	90	94	100
Adaptive	$H_0 : k = 2$	80	2	22	28	41	49	59	70	83	93	97	100
Adaptive	$H_0 : k = 3$	80	0	6	15	25	38	49	60	69	82	91	100
Silvermans	$H_0 : k = 1$	100	2	38	57	66	75	81	89	93	93	99	100
Silvermans	$H_0 : k = 2$	100	1	14	19	32	48	56	68	77	88	99	100
Silvermans	$H_0 : k = 3$	100	0	7	14	27	42	55	73	83	88	95	100
Adaptive	$H_0 : k = 1$	100	23	60	68	72	78	82	90	91	93	97	100
Adaptive	$H_0 : k = 2$	100	1	14	28	38	47	57	68	77	89	94	100
Adaptive	$H_0 : k = 3$	100	0	3	16	28	39	51	65	78	87	92	100

Table C.17: Cumulative frequency of the p value resulting from Silverman's (1981) test of modality and our adaptive modification to the original test (section 5.5). Results are from testing the hypothesis of unimodality, bimodality and trimodality in samples of size n from the population $\frac{1}{3}N(2,1) + \frac{1}{3}N(0,1) + \frac{1}{3}N(3,1)$.

163

| Test	Hypothesis	n	$= 0$	≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	≤ 1.0
Silvermans	$H_0 : k = 1$	40	0	18	34	45	54	66	75	85	94	96	100
Silvermans	$H_0 : k = 2$	40	0	9	26	42	53	61	71	81	91	99	100
Silvermans	$H_0 : k = 3$	40	0	11	18	34	49	57	71	80	88	97	100
Adaptive	$H_0 : k = 1$	40	14	33	44	50	65	72	77	84	91	98	100
Adaptive	$H_0 : k = 2$	40	0	13	23	36	52	65	73	86	93	96	100
Adaptive	$H_0 : k = 3$	40	0	12	25	38	44	62	71	87	91	98	100
Silvermans	$H_0 : k = 1$	60	1	21	39	58	68	83	89	91	94	97	100
Silvermans	$H_0 : k = 2$	60	0	14	29	41	52	61	72	83	92	95	100
Silvermans	$H_0 : k = 3$	60	0	5	15	28	38	53	70	79	91	99	100
Adaptive	$H_0 : k = 1$	60	12	42	55	65	72	73	79	86	93	97	100
Adaptive	$H_0 : k = 2$	60	2	19	30	41	50	65	72	82	91	97	100
Adaptive	$H_0 : k = 3$	60	0	7	21	32	43	54	69	76	85	95	100
Silvermans	$H_0 : k = 1$	80	3	25	44	62	73	81	86	93	97	98	100
Silvermans	$H_0 : k = 2$	80	0	9	26	39	49	63	71	74	83	94	100
Silvermans	$H_0 : k = 3$	80	1	9	18	31	50	57	68	77	87	95	100
Adaptive	$H_0 : k = 1$	80	14	44	58	64	73	80	84	90	92	96	100
Adaptive	$H_0 : k = 2$	80	2	16	26	32	44	51	65	78	86	94	100
Adaptive	$H_0 : k = 3$	80	0	8	16	28	39	54	66	77	84	92	100
Silvermans	$H_0 : k = 1$	100	2	27	45	54	66	81	85	91	96	98	100
Silvermans	$H_0 : k = 2$	100	0	29	40	51	60	73	82	89	91	97	100
Silvermans	$H_0 : k = 3$	100	0	8	22	26	34	45	57	66	81	91	100
Adaptive	$H_0 : k = 1$	100	19	48	55	64	74	82	90	93	95	98	100
Adaptive	$H_0 : k = 2$	100	2	29	39	50	60	67	77	84	87	94	100
Adaptive	$H_0 : k = 3$	100	0	7	19	25	40	49	59	73	83	93	100

Table C.18: Cumulative frequency of the p value resulting from Silverman's (1981) test of modality and our adaptive modification to the original test (section 5.5). Results are from testing the hypothesis of unimodality, bimodality and trimodality in samples of size n from the population $\frac{1}{3}N(2.5, 1) + \frac{1}{3}N(0, 1) + \frac{1}{3}N(2.5, 1)$.

Test	Hypothesis	n	$= 0$	≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	≤ 1.0
								Count of p					
Silvermans	$H_0 : k = 1$	40	0	11	32	45	59	69	79	84	90	95	100
Silvermans	$H_0 : k = 2$	40	0	5	12	22	32	42	57	77	88	99	100
Silvermans	$H_0 : k = 3$	40	0	5	22	30	45	56	66	76	86	96	100
Adaptive	$H_0 : k = 1$	40	8	32	39	48	59	71	78	88	93	97	100
Adaptive	$H_0 : k = 2$	40	0	10	20	34	40	51	60	71	78	89	100
Adaptive	$H_0 : k = 3$	40	0	6	17	27	36	45	62	79	87	95	100
Silvermans	$H_0 : k = 1$	60	0	16	42	53	60	70	79	83	91	95	100
Silvermans	$H_0 : k = 2$	60	0	9	25	35	45	56	68	75	82	92	100
Silvermans	$H_0 : k = 3$	60	0	9	18	31	43	58	75	81	94	100	100
Adaptive	$H_0 : k = 1$	60	15	39	51	58	67	73	78	85	87	94	100
Adaptive	$H_0 : k = 2$	60	1	11	23	37	45	54	67	78	89	95	100
Adaptive	$H_0 : k = 3$	60	1	10	22	36	47	58	68	77	90	97	100
Silvermans	$H_0 : k = 1$	80	0	16	29	43	58	70	81	86	93	96	100
Silvermans	$H_0 : k = 2$	80	0	18	25	38	47	57	65	77	88	92	100
Silvermans	$H_0 : k = 3$	80	0	5	16	28	35	49	68	82	88	99	100
Adaptive	$H_0 : k = 1$	80	5	37	44	55	63	70	78	85	90	99	100
Adaptive	$H_0 : k = 2$	80	0	19	30	41	52	60	72	79	87	92	100
Adaptive	$H_0 : k = 3$	80	0	7	15	26	35	44	60	76	93	99	100
Silvermans	$H_0 : k = 1$	100	0	22	39	53	63	76	86	91	94	97	100
Silvermans	$H_0 : k = 2$	100	0	15	25	36	44	57	65	77	87	95	100
Silvermans	$H_0 : k = 3$	100	0	9	22	34	46	60	67	78	88	95	100
Adaptive	$H_0 : k = 1$	100	16	43	54	63	67	75	81	88	88	94	100
Adaptive	$H_0 : k = 2$	100	1	15	27	39	48	57	69	77	88	95	100
Adaptive	$H_0 : k = 3$	100	1	9	19	33	41	52	65	74	83	93	100

Table C.19: Cumulative frequency of the p value resulting from Silverman's (1981) test of modality and our adaptive modification to the original test (section 5.5). Results are from testing the hypothesis of unimodality, bimodality and trimodality in samples of size n from the population $\frac{1}{3}N(2,1) + \frac{1}{3}N(0,1) + \frac{1}{3}N(2.5,1)$.

165

Test	Hypothesis	n	Count of p										
			$= 0$	≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	≤ 1.0
Silvermans	$H_0 : k = 1$	40	0	13	28	46	53	66	71	80	90	97	100
Silvermans	$H_0 : k = 2$	40	0	10	22	33	47	56	74	86	92	98	100
Silvermans	$H_0 : k = 3$	40	0	13	27	43	56	65	76	88	95	99	100
Adaptive	$H_0 : k = 1$	40	7	34	47	58	65	72	77	86	90	94	100
Adaptive	$H_0 : k = 2$	40	0	14	26	39	53	63	73	85	92	97	100
Adaptive	$H_0 : k = 3$	40	0	9	20	31	48	59	70	84	92	98	100
Silvermans	$H_0 : k = 1$	60	0	15	22	35	47	60	74	83	91	98	100
Silvermans	$H_0 : k = 2$	60	0	9	19	38	46	57	71	80	90	95	100
Silvermans	$H_0 : k = 3$	60	1	11	22	29	44	63	72	82	95	100	100
Adaptive	$H_0 : k = 1$	60	11	30	38	47	51	66	71	78	93	96	100
Adaptive	$H_0 : k = 2$	60	1	15	24	37	50	63	73	77	85	97	100
Adaptive	$H_0 : k = 3$	60	1	11	21	29	42	55	68	77	86	94	100
Silvermans	$H_0 : k = 1$	80	2	15	32	50	56	66	74	81	89	97	100
Silvermans	$H_0 : k = 2$	80	0	5	17	34	48	57	71	81	92	96	100
Silvermans	$H_0 : k = 3$	80	0	11	19	28	39	51	66	83	88	96	100
Adaptive	$H_0 : k = 1$	80	13	35	51	62	66	71	79	87	92	97	100
Adaptive	$H_0 : k = 2$	80	1	13	21	31	42	53	59	69	82	93	100
Adaptive	$H_0 : k = 3$	80	0	9	15	29	41	54	68	79	90	95	100
Silvermans	$H_0 : k = 1$	100	0	20	31	49	57	68	77	80	85	95	100
Silvermans	$H_0 : k = 2$	100	1	6	15	25	39	52	65	76	87	93	100
Silvermans	$H_0 : k = 3$	100	0	7	22	28	39	53	65	76	88	95	100
Adaptive	$H_0 : k = 1$	100	16	38	45	55	63	74	81	86	91	94	100
Adaptive	$H_0 : k = 2$	100	1	5	15	32	42	57	68	75	86	95	100
Adaptive	$H_0 : k = 3$	100	0	15	25	42	55	67	71	80	83	94	100

Table C.20: Cumulative frequency of the p value resulting from Silverman's (1981) test of modality and our adaptive modification to the original test (section 5.5). Results are from testing the hypothesis of unimodality, bimodality and trimodality in samples of size n from the population $\frac{1}{3}N(2,1) + \frac{1}{3}N(0,1) + \frac{1}{3}N(2,1)$.

Appendix D

Publications

The following publications have been based, in part, on the research contained within this thesis and are included in the thesis.

1. Westwood S., Baxter M.J. and Beardah C. (1998). Sample size problems with archaeometric data. Computer Applications and Quantitative Methods in Archaeology 1998 (eds. J. Barcelo, I. Briz and A. Vila), 155-7 (figures on CD-ROM). BAR International Series 757, Oxford: Archaeopress (1999).

2. Westwood S. and Baxter, M.J. (2000). Exploring archaeometric data using projection pursuit methodology, UK Chapter of Computer Applications and Quantitative Methods in Archaeology 1999 (eds. C. Buck et al.), 81-90. BAR International Series 844, Oxford: Archaeopress (2000).

3. Baxter M.J., Beardah C.C. and Westwood S. (2000). Sample size and related issues in the analysis of lead isotope data, *Journal of Archaeological Science*, **27**, 973-980.
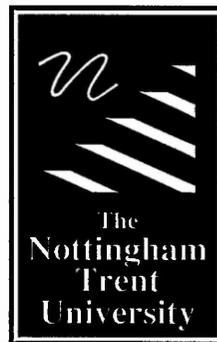
Libraries &
Learning
Resources

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

156

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)

Exploring archaeometric data using projection pursuit methodology

S. Westwood and M. J. Baxter

Department of Mathematics, Statistics and Operational Research

Clifton Campus, The Nottingham Trent University

Nottingham NG11 8NS, UK.

email: simon.westwood@ntu.ac.uk

1 Introduction

Principal component analysis (PCA) is widely used in quantitative archaeological studies for investigating structure in multivariate data. It can be viewed as a particular application of projection pursuit (PP) methodology. The focus of this paper is on whether more general PP methods have much to offer in comparison to PCA for analyzing artefact compositional data in archaeometric studies.

Applications of multivariate methodology to chemical compositional data for artefacts, in the form of an $n \times p$ data matrix, $\mathbf{X}$, are among the most common uses of multivariate statistics in archaeology (Baxter 1994). Cluster analysis and PCA are the most widely used methods. The principal components, of which there are p, are uncorrelated linear combinations of the original variables. The first component has maximum variance, subject to a normalizing constraint on the coefficients; the second component has the second highest variance, and so on.

Results are often presented as plots based on the first few components, in the hope of revealing interesting structure in the data. Projection pursuit, as used here, might be similarly described, except that linear combinations of variables are sought which optimise an index other than variance and attempt, more directly, to measure 'interesting' structure in the data. It has been claimed that PCA is 'something of a blunt instrument' for detecting interesting structure because large variation need not be interestingly structured variation (Jones and Sibson 1987, 2).

Accessible methodological discussions of PP have been available in the British statistical literature since Jones and Sibson (1987), and in the American literature for somewhat longer (Friedman and Tukey 1974). Despite claims that applications of PP have 'flourished' (Posse 1995a, 84) and been 'promoted extensively in the literature and in implementation' (Nason 1995, 413), published practical applications—as opposed to theoretical papers—are quite hard to find. Flenley and Olbricht (1993) and Wilhelm et al. (1999) are the only applications to archaeological data that we know of, other than our own noted in the first example. Applications to data from other subject areas can be found in Friedman (1987), Jones and Sibson (1987), Nason (1995), Ripley (1996), Clements and Jones (1991), Glover and Hopke (1992, 1994), Lendzionowski et al. (1990) and Walden (1994). Several of these papers were written to explore the potential of PP in particular application areas, and this is the spirit in which this paper has been written.

Some of the theory of PP is discussed in the next section, along with practicalities of application. The heart of the paper is the third section, where a variety of applications are discussed. Our current thoughts on the usefulness of PP for archaeometric data analysis are presented in the final section.

2 Projection pursuit

2.1 Theory

Mathematical discussions are given in Huber (1985), Jones and Sibson (1987), Friedman (1987), Hall (1989), Cook et al. (1993), Sun (1991; 1993), Li and Cheng (1993), Eslava and Marriott (1994), Posse (1995a; 1995b) and Nason (1995), with a useful overview being provided by Ripley (1996, 296–303).

The central idea is to find k linear functions (projections) of the original variables that, when plotted, show interesting structure in the data. Usually $k = 1$ or $k = 2$ functions are sought, though $k = 3$ is possible (Nason 1995; Glover and Hopke 1992; 1994). Many approaches begin by equating *uninteresting* structure with the normal distribution and seek projections that are as non-normal as possible. For ease of exposition $k = 1$ is assumed in what follows.

Let $f(x)$ be the probability density of a random variable, X, with $\phi(x)$ the 'null' density when $f(x)$ has a normal distribution. A measure of weighted distance between $f(x)$ and the normal distribution is

$$I = \int (f(x) - \phi(x))^2 w(x)\, dx \qquad (1)$$

where $w(x)$ is a weight function. Equation 1 can be used as the basis for a variety of indices of interestingness; Cook *et al.* (1993) discuss indices for which $w(x) = \phi(x)^a$ for $a = -1$ (Friedman 1987), 0 (Hall 1989) and 1. Other forms of index are discussed in the references given above. Some of the indices that have been proposed for PP can be used as omnibus tests of normality (for example Mardia 1987) and Simonoff (1996, 117) has noted that any reasonable test statistic for normality is a candidate index.

For practical purposes $f(x)$ must be estimated. Using orthogonal series estimates of $f(x)$ and expansions of $\phi(x)$ Ripley (1996) notes that the index in equation 1 can be estimated as

$$\hat{I} = \sum_{i=0}^{\infty} w_i (a_i - b_i)^2 \qquad (2)$$

where a_i are the coefficients in the orthogonal series estimator; b_i are constants arising from the expansion for a normal distribution; and w_i depends on the weight function being used. For practical use the series in equation 2 must be truncated. Cook *et al.* (1993) call indices of the form given in equation 2 the *Legendre* (a = -1), *Hermite* (a = 0) and *Natural Hermite* (a = 1) indices after the orthogonal series expansions used to obtain the coefficients. Generalization to two-dimensional PP, of indices of the above type, is discussed in Cook *et al.* (1993) and Posse (1995a).

In the examples to follow we have used the two-dimensional Legendre, Natural Hermite and Friedman-Tukey indices, the last of which can be viewed as an estimate of the index $\int f(x)^2\, dx$ based on kernel density estimates (Jones and Sibson 1987, 5).

2.2 Practicalities

To implement two-dimensional projection pursuit the XGobi program (Swayne *et al.* 1991), which is freely available and can be run under X-Windows or from within the S-Plus package (Venables and Ripley 1997), has been used. The source of this, and other, software is discussed in the Appendix.

Many practitioners (for example Jones and Sibson 1987; Cook *et al.* 1993; Nason 1995) recommend 'sphering' the data before applying PP. This involves transforming the data to new variables that are uncorrelated and have the same variance. For most of our applications the data have been standardized to have zero mean and unit variance; principal components have been extracted; and the resultant principal component scores have been renormalized to have unit variance for each component. This results in a set of uncorrelated variables (essentially the principal components of standardized data) with equal variance. Some approaches to data analysis in archaeometry work routinely with logarithmically transformed but unstandardized data (for example Glascock 1992). It would be equally possible to sphere by finding the principal components of such data, and then renormalizing, though we have not used this approach below.

Typically p is in the range 8–30, but multicollinearity among the variables means that the effective dimensionality of the data is often much less than p. Typically more than 90% of the variation in the data will be accounted for by fewer than 10 principal components. The leading components, rather than the original variables, may be used in the PP.

A limitation of some approaches to PP is the sensitivity to outliers in the data, or to the tails rather than centre of the data. The moment index of Jones and Sibson (1987) has been criticised for the former reason (Ripley 1996, 300), and the Legendre index of Friedman (1987) for the latter reason (Cook *et al.* 1993, 228). It is possible to use PP as an informal method of multivariate outlier detection; if this is not of interest it is sensible to remove obvious outliers before analysis by PP, as they will often be the dominant feature of 'interesting' projections.

Sample size may also be a practical problem. Cook *et al.* (1993, 244–5) provide examples to show that for sparse, high-dimensional data PP can suggest spurious structure in random data. One of their cautionary examples shows apparent structure in a 100×5 data set, generated

randomly from a normal distribution. Many archaeometric data sets have smaller n and larger p, so that particular heed should be paid to their warning that 'exploratory projection pursuit will always find structure, albeit weak, but care must be taken when emphasizing the significance of that structure'. In similar vein Ripley (1996, 301) suggests that, with large p PP 'may be used for hypothesis formation, but we will need independent evidence of the validity of the structure suggested by the plots'.

In the indices that have been discussed z, in the one-dimensional case, is a linear combination of p variables. The coefficients of this linear combination must be estimated and this leads to a non-linear optimisation problem in p dimensions which must be solved numerically. For higher dimensions the problem is obviously compounded. In XGobi the optimisation can be tracked, and numerous local optima will be found that may correspond to interesting views of the data. This will be illustrated in our second example. The ability of PP to produce multiple views of the data is widely seen as an attraction of the method as will be discussed in the final section, but is also time consuming.

Ripley (1996) observes that there is no unanimity in practice about which indices to use and advises that several should be tried. This has been done here, and the examples that follow represent only a small selection of the analyses that have been undertaken.

3 Examples

3.1 Example 1 - Lead isotope data

The following example, based on Baxter (1999), will be discussed in summary form only. We have used the method to be presented routinely in past work, but have not previously noted its interpretation as a PP method. Lead isotope ratio data are three-dimensional and in their analysis it has sometimes been assumed that data from an ore source can be treated as a sample from a multivariate normal distribution (Sayre et al. 1992). Recent work by Baxter and Gale (1998) and Baxter (1999) has called into question the general validity of this assumption. In particular, Baxter (1999) used a variety of tests of multivariate normality to demonstrate that many of the data sets in Stos-Gale et al. (1996) could not reasonably be regarded as samples from normally distributed data.

One test used was the multivariate extension of the univariate Shapiro-Wilk test statistic for normality (Malkovich and Afifi 1973). In this test the linear combination of the three lead isotope ratios is sought that minimizes the univariate statistic. This can be viewed as a PP method that results in a linear function, $k = 1$, that best displays the non-normality of the data. Figure 1 shows a kernel density estimate of the most non-normal linear combination for the Kea field, with $n = 62$. Formal tests of normality suggest that the data are non-normal and this particular application of PP methodology suggests that the data are strongly multi-modal. Further illustrations of this kind of use can be found in Baxter and Gale (1998) and Baxter (1999).

3.2 Example 2 - Blue soda glass from York

Cox and Gillies (1986) published analyses of blue soda glass from the windows of York Minster and archaeological excavations that has been used elsewhere to illustrate a variety of methodologies (Baxter 1989; Baxter and Buck 2000; Bell and Croson 1998). There are 27 specimens, measured with respect to the concentration of 12 oxides and elements. Most analyses clearly show three main groups in the data, with some analyses suggesting possible sub-groups or outliers.

Figure 2 shows four analyses of the data. The PCA analysis (of standardized data) in Figure 2(a) shows the three groups, one of which is dispersed relative to the other two. This structure is readily found using PP, and an example is given in Figure 2(b), where the structure is even more apparent. The view illustrated in Figure 2(c), in which the structure is 'circular', occurs quite commonly in our experience with similar data sets, and has no useful practical interpretation. Similar examples can be found in Cook et al. (1993, 248) and Ripley (1996, 302). Figure 2(d) shows a view in which outliers are the predominant feature.

We may remark that for this data set the structure is fairly obvious and found almost 'instantaneously' by PCA. Other useful views were not found in the course of exploration using PP. The PP view in Figure 2(b) is 'sharper' than the PCA view, but tells essentially the same story.
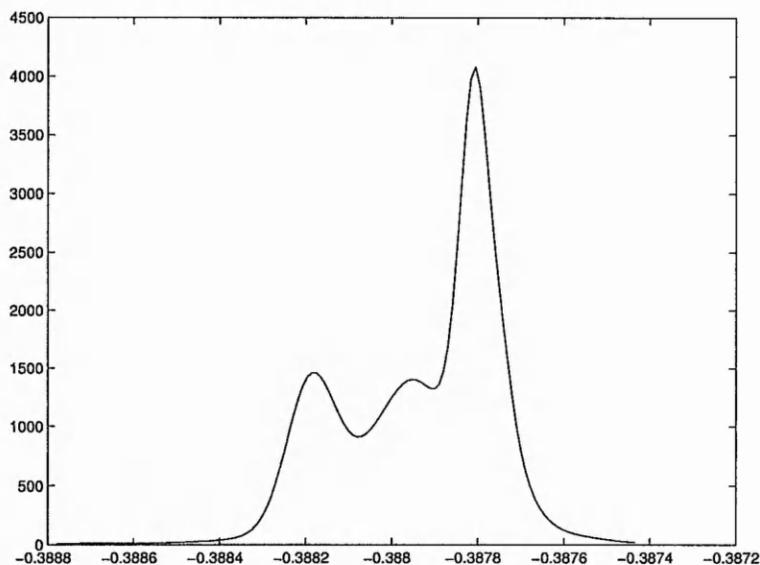
Figure 1: A kernel density estimate of the most non-normal linear combination of three-dimensional lead isotope ratio data for the Kea field.

3.3 Example 3 - Waste glass from Leicester and Mancetter

The data used in this example consist of 105 specimens of waste glass found on furnace sites at Leicester and Mancetter and measured with respect to the concentration of 11 major and minor oxides. It is of interest to see if there are distinct chemical groups in the data, and if these correspond to the furnace sites. The data were collected and published by Jackson (1992) and are reproduced in Baxter (1994) where extensive analysis was undertaken using a variety of multivariate methods. These analyses suggest three concentrations in the data with some correspondence—by no means exact—to the furnace groups.

This is shown in the PCA plot in Figure 3(a), where labelling is by site. Without a knowledge of the sites it is possible, visually or with the aid of techniques such as kernel density estimation (Baxter *et al.* 1997) to detect three main concentrations in the data. There are no obviously distinct clusters. The densest concentration to the right consists mainly of glass from Leicester; the other two concentrations contain most of the Mancetter specimens, with 11 to 14 Leicester specimens mixed in (depending on how boundaries of concentrations are visualised).

The PP view in Figure 3(b) quite clearly isolates a cluster of cases in the bottom half of the plot consisting, with one exception, of Leicester specimens. The remaining dispersed group, possi-

bly sub-dividing into two, contains the Mancetter specimens with the same number of Leicester specimens mixed in as in the PCA.

Arguably the PCA and PP analyses lead to similar conclusions, but the separation between material from the two sites, and the fact that it is less than perfect, is clearer in the latter analysis because of the clearer clustering revealed. We remark that we can be confident that PP is not revealing spurious structure in this case because information not used in the PP, concerning site of origin, allows us to interpret the revealed structure in a useful archaeological manner.

3.4 Example 4 - Oriental Greenwares

This example is based on a 133×9 data set published by Pollard and Hatcher (1986) showing the chemical composition of 133 oriental greenwares which are suspected to have originated from several areas of manufacture. We follow them in omitting three clear outliers and one variable, SiO_2, in our analysis.

There are two very obvious chemical groups in the data, as the PCA in Figure 4(a) shows. The group to the left is associated with Northern Zhejiang Yue wares and that to the right with Longquan celadons. It is easy to get the same separation using PP and one such view is shown in the Figure 4(b). This additionally suggests a small group at the bottom of the plot that is a
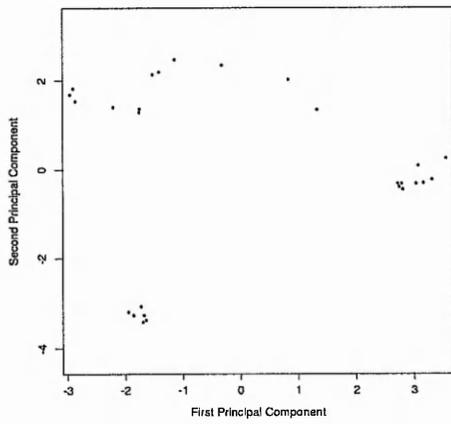
(a) PCA of standardised data
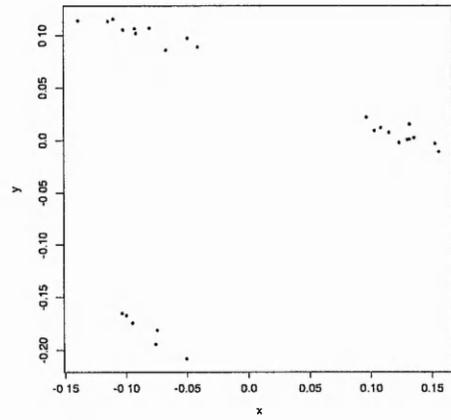
(b) PP using Friedman-Tukey index
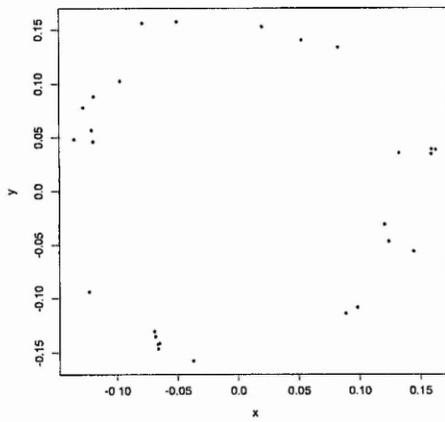
(c) PP using Natural Hermite index
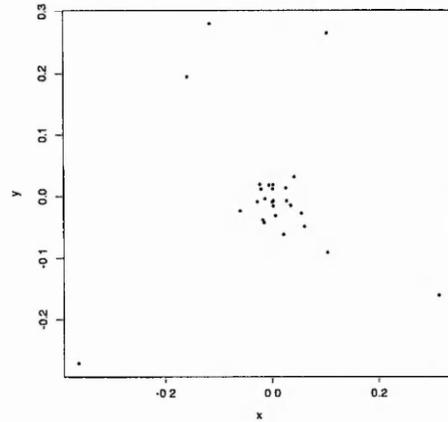
(d) PP using Natural Hermite index

Figure 2: Plots of York Minster data.
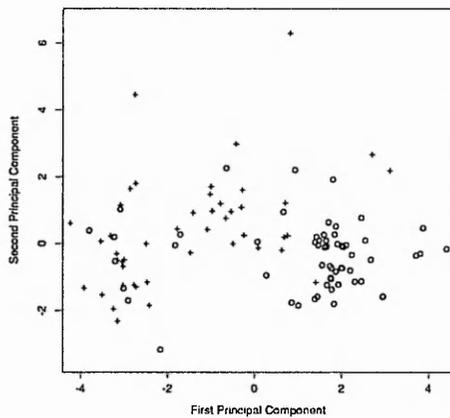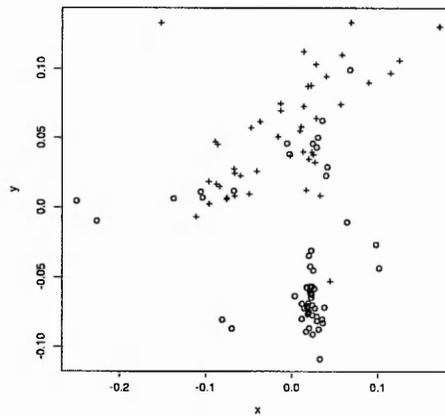
| (a) PCA of standardised data | (b) PP using Legendre index |

Figure 3: Plots based on analyses of the waste glass compositions from Mancetter and Leicester. Labelling is by site with '+' cases from Mancetter and 'o' cases from Leicester.

subset of the earlier wares, but we have been unable to interpret this as archaeologically distinct in any way.

Pollard and Hatcher (1986) applied cluster analysis to the 53 specimens in the earlier group and concluded there were three subgroups. After applying stepwise discriminant analysis to these, five outliers were removed and a discriminant analysis plot for the remaining 48 cases was shown on page 268 of their paper. A similar analysis is shown in Figure 4(c), the only difference being our use of all eight variables rather than the five selected in the original publication. Interpretation of the groups is not absolutely clear-cut, but they can be associated with regional differences in composition. Given a knowledge of this classification we have been unable to obtain a PP view that separates out the groups as well as the discriminant analysis. In Figure 4(d) one PP view for the 48 cases is shown which separates out the smaller group but not the two larger ones. It may be noted that a PCA analysis of this subset (not shown) did as well as the PP in separating the groups.

Our PP analysis of this data set cannot be regarded as especially successful. Although the PP for the full data set did suggest structure additional to that revealed by PCA we were unable to interpret the results in an archaeologically useful fashion, so have no real way of determining whether the structure is spurious or not. Similar remarks apply to other PP analyses of the subset

used in Figure 4 that are not shown here.

4   Discussion

For the specialized problem of Example 1 there is no doubt that one-dimensional PP, as illustrated there, has a useful role to play. Our current practice is to use PP in conjunction with tests of normality to explore the nature of the non-normality when it occurs. Unpublished work in progress suggests that the use of PP in isolation can mislead if the sample sizes are small.

Examples 2 and 3, particularly the latter, show that two-dimensional PP can produce a sharper view of structure in the data than that provided by PCA, but it was also the case that PP did not lead to an interpretation different from that achieved with PCA. In example 4 the PP analysis did not lead to any new insights into the data. Although some additional structure was suggested there is no obvious way of determining whether it is spurious or not. Our experience with these data sets is representative of others we have worked with.

In the wider literature there undoubtedly exist examples where PP does produce informative views of the data that PCA does not reveal. This sometimes occurs when the structure in the data is 'unusual' (see, for example, the structures used in Posse's (1995a, 91) simulation study), and of a kind that we suspect would often be regarded as
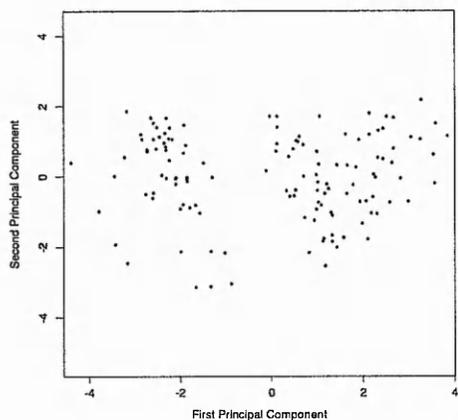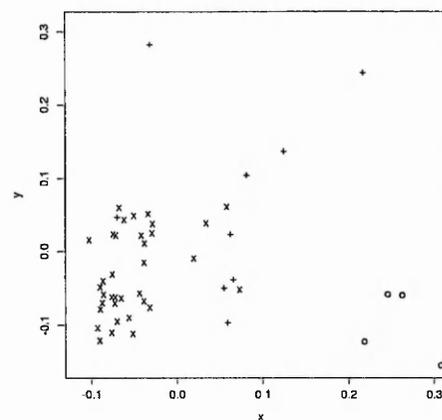
(a) PCA of standardised data
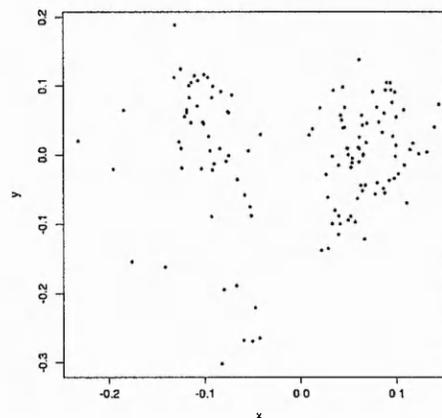
(b) PP using Friedman-Tukey index
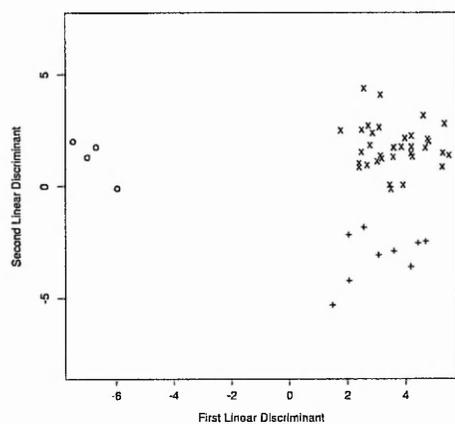
(c) Discriminant analysis plot for the smaller group from Figure 4(a)

(d) PP using Legendre index for the smaller group from Figure 4(a)

Figure 4: Plots based on analyses of the oriental Greenware compositions. Figures 4(c) and 4(d) use data from the left-hand group showing in Figure 4(a), after omitting five outliers, labelled according to the groups determined by Pollard and Hatcher (1986).

Label in Appendix	URL for software	Accessed
1	http://www.research.att.com/~andreas/xgobi/	21/02/00
2	http://www.stats.ox.ac.uk/pub/SWin/xgobi.zip	21/02/00
3	http://lib.stat.cmu.edu/general/projpurs	21/02/00
4	http://www.stats.bris.ac.uk/~guy/Research/PP/PP.html	21/02/00
5	http://www.stats.bris.ac.uk/pub/software/pp2/mcj_pp.shar.gz	21/02/00

Table 1: Freely available projection pursuit software

uninterpretable in the context of the type of data used here. The model, often implicit, in studies that produce data similar to those used in examples 2, 3 and 4 is that the data may be viewed as a sample from a mixture of distributions which, in those studies that make statistical assumptions, are multivariate normal. In p-dimensional space the expectation is either that there will be distinct point clouds, or that there will be overlapping point clouds with distinct high-density regions. We suspect that methods such as PCA or cluster analysis will often be adequate to detect this, and that a PP view showing a marked departure from the underlying model might be difficult to interpret (we also recognise that this is not a good argument for not using PP).

Where PP has been contrasted with PCA and judged to be superior (for example Glover and Hopke 1992) the judgement is sometimes a fine one. It is also the case that in order to select a PP view and judge that it is superior to PCA it may be necessary to use additional information (for example a prior classification of the data) to confirm that the PP view is a useful one. Given the ease with which PP can suggest spurious structure with 'small' data sets we have found it very difficult to interpret results where such prior knowledge has not been available. The superiority of PP compared with PCA has sometimes been exaggerated. Posse (1995a, 83–84) analyses data on five measurements for 200 Australian crabs, most belonging to four groups. He claims that PP is able to reveal a 'clustered projection' that was 'not found by principal component analysis'. In fact the first component has an obvious size interpretation, and any of several standard approaches to PCA that aim to remove size effects (including a plot of the second and third components) will reveal a clustered projection similar to that found by PP.

Thus, while not disputing the theoretical interest of PP or its potential for revealing unusual and unexpected structure in large data sets, we remain agnostic about its value as a tool for the *routine* analysis of data of the kind discussed in

examples 2–4. In addition to the reasons discussed so far, there are two pragmatic reasons that give rise to this agnosticism. The first concerns the size of the data sets typically available. Most examples of two-dimensional PP that we have seen use $p < 10$; our examples used $p = 8$, 11 and 12; it is now quite common to see analyses based on data sets for which $p > 20$. There has not been a commensurate increase in the size of samples typically collected, so that $n < 100$ is quite usual. In the context of a technique that can easily suggest spurious structure in small data sets, and where 100×5 is considered to be small (Cook *et al.* 1993), many archaeometric data sets are small and subject to the problems that this entails.

Our second reason concerns the time required to carry out PP. A large number of local optima arise in analysis, and the views they are associated with need to be inspected to see if they are 'interesting' and have a useful archaeological interpretation. In XGobi the plots produced in the course of pursuit can be viewed in real time and visually 'interesting' projections, including some used here, do not necessarily even correspond to local optima. These also need to be assessed and this is very demanding of time and has not, in the many analyses that we have undertaken, led to much extra insight into the data being gained, beyond that provided by PCA and cluster analysis—the tools most usually deployed in the literature. For these and other reasons we do not view PP, in its current state of development, as a tool we would recommend for routine archaeometric data analysis.

Appendix - Software availability

Freely available projection pursuit software is listed in Table 1. The software used in this paper was obtained from the first of the listed sites. Other sources of PP software are also listed, but we should stress that we have no experience in

using these.

The XGobi software we used, running under X-Windows, is freely available and may be obtained from source 1 in Table 1. A Microsoft Windows version can be found at 2. FORTRAN software for two-dimensional PP (Friedman, 1987) can be found at 3 and 4 where software for Nason's (1995) three-dimensional PP is also available. FORTRAN code for two-dimensional PP from Jones and Sibson (1987) can be found at 5.

Acknowledgements

References

[1] Baxter, M. J. 1989. Multivariate analysis of data on glass compositions: a methodological note. *Archaeometry*, 31:45–53.

[2] Baxter, M. J. 1994. *Exploratory Multivariate Analysis in Archaeology*. Edinburgh, Edinburgh University Press.

[3] Baxter, M. J. 1999. On the multivariate normality of lead isotope fields. *Journal of Archaeological Science*, 26:117–124.

[4] Baxter, M. J., Beardah, C. C. and Wright, R. V. S. 1997. Kernel density estimation for archaeology. *Journal of Archaeological Science*, 24:347–354.

[5] Baxter, M. J. and Buck, C. E. 2000. Data handling and statistical analysis. In E. Ciliberto and G. Spoto, (eds.), *Modern Analytical Methods in Art and Archaeology*. New York (in press), Wiley.

[6] Baxter, M. J. and Gale, N. H. 1998. Testing for multivariate normality via univariate tests: a case study using lead isotope ratio data. *Journal of Applied Statistics*, 25:679–691.

[7] Bell, S. and Croson, C. 1998. Artificial neural networks as a tool for archaeological data analysis. *Archaeometry*, 40:139–151.

[8] Clements, A. M. and Jones, M. C. 1991. An ecological example of the application of projection pursuit to compositional data. *Vegetatio*, 95:101–117.

[9] Cook, D., Buja, A. and Cabrera, J. 1993. Projection pursuit indexes based on orthonormal function expansions. *Journal of Computational and Graphical Statistics*, 2:225–250.

[10] Cox, G. A. and Gillies, K. J. S. 1986. The X-ray fluorescence analysis of medieval durable blue soda glass from York Minster. *Archaeometry*, 28:57–68.

[11] Eslava, G. and Marriott, F. H. C. 1994. Some criteria for projection pursuit. *Statistics and Computing*, 4:13–20.

[12] Flenley, E. C. and Olbricht, W. 1993. Classification of archaeological sands by particle size analysis. In O. Opitz, B. Lausen and R. Klar, (eds.), *Information and Classification. Concepts, Methods and Applications. Proceedings of the 16th Annual Conference of the Gesellschaft fr Klassifikation e. V.*, pp. 478–489. Berlin, Heidelberg, Springer.

[13] Friedman, J. H. 1987. Exploratory projection pursuit. *Journal of the American Statistical Association*, 82:249–266.

[14] Friedman, J. H. and Tukey, J. W. 1974. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, 23:881–889.

[15] Glascock, M. D. 1992. Characterization of archaeological ceramics at MURR by neutron activation analysis and multivariate statistics. In H. Neff, (ed.), *Chemical Characterization of Ceramic Pastes in Archaeology*, pp. 11–26. Madison, Wisconsin, Prehistory Press.

[16] Glover, D. M. and Hopke, P. K. 1992. Exploration of multivariate chemical data by projection pursuit. *Chemometrics and Intelligent Laboratory Systems*, 16:45–59.

[17] Glover, D. M. and Hopke, P. K. 1994. Exploration of multivariate atmospheric particulate compositional data by projection pursuit. *Atmospheric Environment*, 28:1411–1424.

[18] Hall, P. 1989. Polynomial projection pursuit. *The Annals of Statistics*, 17:589–605.

[19] Huber, P. J. 1985. Projection pursuit (with discussion). *Annals of Statistics*, 13:435–525.

[20] Jackson, C. M. 1992. *A Compositional Analysis of Roman and Early Post-Roman Glass and Glass-working Waste from Selected British Sites.* PhD thesis, University of Bradford, UK.

[21] Jones, M. C. and Sibson, R. 1987. What is projection pursuit? (with discussion). *Journal of the Royal Statistical Society A*, 150:1–36.

[22] Lendzionowski, V., Walden, A. T. and White, R. E. 1990. Seismic character mapping over resevoir intervals. *Geographical Prospecting*, 38:951–969.

[23] Li, G. Y. and Cheng, P. 1993. Some recent developments in projection pursuit in China. *Statistica Sinica*, 3:35–51.

[24] Malkovich, J. F. and Afifi, A. A. 1973. On tests for multivariate normality. *Journal of the American Statistical Association*, 68:176–179.

[25] Mardia, K. V. 1987. Discussion of 'What is projection pursuit?' by Jones and Sibson (1987). *Journal of the Royal Statistical Society A*, 150:22–23.

[26] Nason, G. P. 1995. Three-dimensional projection pursuit. *Applied Statistics*, 44:411–430.

[27] Pollard, A. M. and Hatcher, H. 1986. The chemical analysis of oriental ceramic body compositions: part 2 - greenwares. *Journal of Archaeological Science*, 13:261–287.

[28] Posse, C. 1995a. Tools for two-dimensional exploratory projection pursuit. *Journal of Computational and Graphical Statistics*, 4:83–100.

[29] Posse, C. 1995b. Projection pursuit exploratory data analysis. *Computational Statistics and Data Analysis*, 20:669–687.

[30] Ripley, B. D. 1996. *Pattern Recognition and Neural Networks.* Cambridge, UK, Cambridge University Press.

[31] Sayre, E. V., Yener, K. A., Joel, E. C. and Barnes, I. L. 1992. Statistical evaluation of the presently accumulated lead isotope data from anatolia and surrounding regions. *Archaeometry*, 34:73–105.

[32] Simonoff, J. S. 1996. *Smoothing Methods in Statistics.* New York, Springer.

[33] Stos-Gale, Z. A., Gale, N. H. and Annetts, N. 1996. Lead isotope data from the Isotrace laboratory, Oxford *Archaeometry* data base 3, ores from the Aegean, part 1. *Archaeometry*, 38:381–390.

[34] Sun, J. 1991. Significance levels in exploratory projection pursuit. *Biometrika*, 78:759–769.

[35] Sun, J. 1993. Some practical aspects of exploratory projection pursuit. *SIAM Journal on Scientific Computing*, 14:68–80.

[36] Swayne, D. F., Cook, D. and Buja, A. 1991. *User's manual for XGobi, a dynamic graphics program for data analysis implemented in the X Windows system (Release 2).* Bellcore Technical Memorandum.

[37] Venables, W. N. and Ripley, B. D. 1997. *Modern Applied Statistics with S-Plus: Second Edition.* New York, Springer.

[38] Walden, A. T. 1994. Spatial clustering: using simple summaries of seismic data to find the edge of an oil-field. *Applied Statistics*, 43:385–398.

[39] Wilhelm, A. F. X., Wegman, E. J. and Symanzik, J. 1999. Visual clustering and classification: The Oronsay particle size data set revisited. *Computational Statistics*, 14:109–146.

Sample Size and Related Issues in the Analysis of Lead Isotope Data

M. J. Baxter, C. C. Beardah and S. Westwood

Department of Mathematics, Statistics and Operational Research, Clifton Campus, The Nottingham Trent University, Nottingham NG11 8NS, U.K.

The statistical analysis of lead isotope ratio data in archaeology has attracted considerable controversy, but one area of consensus seems to be that a minimum sample size of 20 is adequate for the satisfactory characterisation of a lead isotope field. The argument in the present paper is that this is too small. Twenty would be satisfactory if the assumption of normality sometimes used in analysing lead isotope was correct, but it is inadequate for checking this assumption or detecting non-normal structures within a field. Evidence based on both real and simulated data suggests that 40 may be a more realistic minimum, and even this is not always adequate. The consequences of incorrectly assuming normality, and alternative methods of analysis that do not involve this assumption, are investigated.   © 2000 Academic Press

Keywords: LEAD ISOTOPE DATA, KERNEL DENSITY ESTIMATES, MULTIMODALITY, NORMALITY, SAMPLE SIZE.

Introduction

The subject of the statistical treatment and interpretation of lead isotope data in archaeology has occasioned considerable debate. The positions of many of the protagonists are summarised in the discussion following Budd *et al.* (1995), and later contributions include Scaife *et al.* (1996); Tite (1996) and Stos-Gale *et al.* (1997). A recent review of many aspects of this debate is provided by Scaife (1998).

In a subject where there has been little consensus, one area of agreement seems to be the belief that 20 is the minimum sample size necessary for the satisfactory statistical treatment of lead isotope data (Pollard & Heron, 1996: 328). The main purposes of this paper are to question this belief; to suggest alternative guidelines; and to examine some of the consequences for data analysis. This last issue is related to the fact that 20 *may* be adequate *if* data are normally distributed, but not otherwise. The issue of normality and the consequences of non-normality for sample size requirements are discussed in the next two sections.

If data are non-normal, then alternatives to the normal-based methods favoured by some (e.g. Sayre *et al.*, 1992, 1995) are needed. Alternatives, based on the use of kernel density estimates (KDEs), are reviewed but are demanding of data in their own right. A discussion of the issues raised concludes the paper.

The Normal Distribution of Lead Isotope Data

Lead isotope data

A specimen from an ore-body may be characterised by measurements on three lead isotope ratios, $^{208}Pb/^{206}Pb$, $^{207}Pb/^{206}Pb$ and $^{206}Pb/^{204}Pb$, that define a point in three-dimensional space. A sample of N specimens defines a three-dimensional (trivariate) cloud of points that is a sample from the lead isotope field of the ore-body. The sample may be used to estimate the lead isotope field of the ore-body, often presented in the form of bivariate plots of ratios with confidence ellipsoids delimiting the estimated extent of the field (e.g. Gale & Stos-Gale, 1992; Sayre *et al.*, 1992). The construction of these ellipsoids requires the assumption that the data are sampled from a trivariate normal distribution.

If fields for ore-bodies are distinct, the possibility exists that the lead isotope signatures of metal artefacts can be matched with fields, and identify possible sources. Opinion is divided on whether simple graphical methods suffice for such matching, or whether there are benefits to be gained in using methods of multivariate statistical analysis. Proponents of the latter view (Sayre *et al.*, 1992, 1995) determine the (Mahalanobis) distance of an artefact from an ore-body, and convert this to a probability that is used to assess whether the ore body could be the source of the artefact. The probability calculations also require the assumption of trivariate normality of the field.

The normality assumption

In those publications that assume normality, the assumption has not usually been rigorously tested. Sayre *et al.* (1992) note that they "often create histograms of the distributions . . . for source group specimens along the characteristic vectors for the groups. If the specimens of a group are normally distributed along each of its characteristic vectors one can conclude that the group is normally distributed . . . These histograms for our groups do approximate Gaussian curves." Limitations of this approach are the small sample sizes often used and the inefficiency of the histogram as a method of assessing normality. The assertion that normality of the characteristic vectors implies multivariate normality is also wrong.

The statement of Gale & Stos-Gale (1993: 256) that examination of many ore data sets "shows that bivariate normality is in most cases well satisfied" is based on testing the normality of univariate ratios. Unfortunately, as is clear from examples in Baxter & Gale (1998), univariate normality of the ratios does not establish multivariate normality.

In contrast to this confidence that normality is the rule, Scaife *et al.* (1996) assert that "some of the larger fields are clearly non-normal in their distributions", and provide an example based on univariate tests of ratios for the Cyprus field, as then defined, to support their claim.

A problem with some analyses on which these statements are based, explicitly acknowledged by Sayre *et al.* (1992: 97), is the need to work with "less than optimum amounts of data". Analysis of some of the larger data sets published in Stos-Gale *et al.* (1996), using univariate tests of normality adapted to multivariate data, suggested that they were, in fact, nonnormal (Baxter & Gale, 1998). This was confirmed using a battery of truly multivariate tests in Baxter (1998), which additionally produced evidence that several smaller data sets ($N \geq 15$) also exhibited signs of non-normality. It was concluded in that paper that non-normality was possibly the rule rather than the exception, and independent work by Scaife (1998) on other data sets is consistent with this conclusion.

Does normality matter?

The assumption of normality is a common one in applications of statistical methodology. It is rarely, if ever, exactly true but is often a sufficiently good approximation that the methodology is not compromised. There are also circumstances when the normality assumption is manifestly untrue, but the output of a statistical analysis is insensitive to this. In such circumstances the methodology is said to be "robust" to the assumption. Robustness, like normality, should not be assumed in applications.

This may be seen by examining the way in which lead isotope fields for Cyprus have been defined and used. At one stage a single field for Cyprus (the "old" Cyprus field), based on 43 specimens from several sources on the island, was defined and delineated using confidence ellipsoids (Stos-Gale *et al.*, 1997). The definition of this field, whether or not it could be subdivided, and its use for provenancing artefacts such as oxhide ingots, has led to extensive published debate (e.g. Budd *et al.*, 1995). Gale & Stos-Gale (1992) suggested that fields for different ore-bodies within Cyprus might be distinct, but the data available at that time did not allow this to be checked in a rigorous way. Subsequent publication and analysis of extensive new data (Gale *et al.*, 1997; Stos-Gale *et al.*, 1997) has confirmed this thesis and rendered much previous publication and argument redundant. The important point here is that, because of an inadequate sample size, it could not be recognised that Cyprus field was non-normal and extremely multimodal. Had larger sample sizes been available at an earlier data much argument could have been avoided.

The foregoing discussion is concerned with the way confidence ellipsoids, calculated using the normality assumption, can mislead if the sampled field is nonnormal. Arguments may also be adduced to show that probability calculations based on the assumption of normality may mislead, and these will be raised later.

Sample Size Considerations

On the belief that 20 is an adequate sample size

The idea that 20 is an acceptable sample size for delineating fields and using them in statistical analysis seems to be generally accepted. Sayre *et al.* (1992: 97) state that "the spread of uncertainty about a source field steadily contracts as the number of specimens describing the source becomes larger, tending to level off when one has something of the order of 20 such data points". Gale & Stos-Gale (1992: 312), Leese (1992: 318) and Reedy & Reedy (1992: 327) concur with the suggestion that a sample size of 20 is desirable, and Pollard & Heron (1996: 328) summarise the consensus by noting that, in characterising source fields, 20 geologically well selected ore samples define "an agreeable minimum level".

The case of the Cyprus field suggests that 20 (and even 40) can be seriously inadequate. It is thus of interest to ask how a value of 20 has come to be accepted as a reasonable minimum. It would appear to be based on practical experience; in the case of Sayre *et al.* (1992), explicitly in the context of statistical analyses that assume normality.

A more "theoretical" justification might also be advanced. Construction of confidence intervals and probability calculations require estimation of the covariance matrix of the data. For stable estimation, with k variables, Harbottle (1976) suggests that $N > 5k$ is desirable, so for $k = 3$ a value of N close to 20 emerges as a desirable minimum. This rule also envisages that the data are normally distributed.
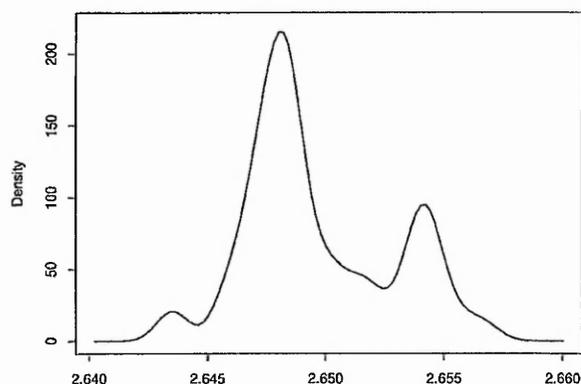
Figure 1. A univariate KDE for the Lavrion field based on a linear combination of the $x_1 = {}^{208}\mathrm{Pb}/{}^{206}\mathrm{Pb}$ and $x_2 = {}^{206}\mathrm{Pb}/{}^{204}\mathrm{Pb}$ ratios, of the form $0.9995x_1 + 0.0314x_2$. The Sheather–Jones (1991) estimate of $h = 0.0006651$ has been used.

To summarise, if the data are sampled from a normal distribution then 20 may be an acceptable minimum sample size for analysis. If the data are sampled from a non-normal distribution 20 may be seriously inadequate, as is now argued.

Sample size requirements

The appropriate sample size depends both on the purpose of sampling and on the structure of the sampled population. The latter is unknown in advance of sampling so a simple answer to questions of sample size is impossible if the possibility of non-normality is admitted. It seems reasonable to suggest that sample sizes larger than 20 will be needed where there is a non-trivial departure from normality.

A possible alternative model to that of normality is that the population is multimodal, suggested by the results of Baxter (1998) using several of the data sets in Stos-Gale *et al.* (1996). For example, Figure 1 shows a kernel density estimate (KDE—see the Appendix for technical details) for the most non-normal linear combination (or projection) of two of the ratios for the Lavrion field, as identified by Malkovich & Afifi's (1973) multivariate extension of the Shapiro–Wilk statistic. This suggests that the field is bimodal, as does a more direct three-dimensional representation explored in the next section (Figure 4).

Distributions such as that in Figure 1 can be modelled as a two-component mixture of univariate normals:

$$f(x) = pN(\mu_1, \sigma_1^2) + (1-p)N(\mu_2, \sigma_2^2)$$

where p is the mixing proportion ($0 \leq p \leq 1$) and $N(\mu_i, \sigma_i^2)$ is the normal distribution with mean μ_i and variance σ_i^2. One way of getting a feel for sample size requirements needed to detect non-normality is to simulate data from a normal mixture; repeatedly draw

Table 1. *The table showing the power, in %, of the ISE test for two-component mixtures of normals with p=0.5*

n	$\|\mu_1 - \mu_2\|$		
	3	4	5
20	11	38	75
30	27	70	97
40	39	89	99
50	47	95	100
100	86	100	100

random samples of a fixed size from it; and estimate the power of a formal test of normality (the proportion of times that the hypothesis of normality is rejected).

The results of such an exercise depend on the values of p, μ_i, σ_i^2 and the test used, and a detailed account of our work on this is not provided here. It is, however, clear that a sample size of 20 will often be inadequate for detecting obvious bimodality in the population. For example, for $p = 0.5$ and assuming equal variances (which without loss of generality can be modelled as $\sigma_1 = \sigma_2 = 1$; Titterington *et al.*, 1985: 161), the population is bimodal if $|\mu_1 - \mu_2| > 2$. Table 1 shows the estimated power, based on 1000 simulated samples, for $p = 0.5$ and different values of N and $|\mu_1 - \mu_2|$ using the integrated squared error (ISE) test developed by Bowman (1992).

The results show that even for quite clear separation of the components (e.g. $|\mu_1 - \mu_2| = 4$ a sample size of $N = 20$ is generally inadequate to detect the departure from normality. For this degree of separation sample sizes of the order of 40–50 are needed to achieve reasonable power. Essentially similar results are obtained on varying the mixing proportion, p, provided the variances do not differ by too much.

A more direct way of assessing sample size requirements is to take data sets that exhibit non-normality; subsample from these; and estimate the proportion of times that the non-normality is detected. Using the ISE test of multivariate normality of Bowman & Foster (1993) that generalises Bowman's (1992) univariate test, and other tests, the results of Baxter (1998) provide strong evidence for the non-normality of the Lavrion ($N = 59$) and Kea ($N = 62$) fields (data in Stos-Gale *et al.*, 1996). Using KDEs for selected projections, as in Figure 1, provides good evidence that the non-normality takes the form of multimodality.

If 1000 subsamples of size 20 are selected from the data sets for each field, and tested for trivariate normality using the ISE test, the hypothesis of normality is rejected 37% of the time for Lavrion, and 20% of the time for Kea. In other words a sample size of 20 is wholly inadequate to detect the non-normality manifest in the full data sets. For Lavrion and subsamples of size 40 the normality hypothesis was rejected 86% of the time; for Kea and subsamples of size 50 the normality hypothesis was rejected 68% of

the time. It is possible for projections of the data based on subsets of the variables to be more revealing of non-normality than use of the full data set (Baxter & Gale, 1998). This was investigated, but essentially the same results were obtained.

The results reported above are specific to two particular data sets but show clearly that much larger sample sizes than those commonly recommended are necessary to detect quite evident departures from normality. An implication of this is that the use of confidence ellipsoids to represent the extent and structure of fields can be seriously misleading. The history of the Cyprus field is additional evidence for this, and Scaife (1998) has independently reached similar conclusions. What is needed is a way of representing fields that respect their structure, without imposing assumptions such as that of normality. This is investigated in the next section, followed by a brief exploration of the effects of non-normality on Mahalanobis distance and probability calculations.

Data Presentation using KDEs

The construction of confidence ellipsoids for fields requires the assumption of normality. This is a parametric statistical procedure; if the assumption is correct an efficient representation of the data is achieved, but if it is wrong the outcome can be misleading. Nonparametric methods of data display avoid assumptions such as normality, and allows the data to "speak for itself". Kernel density estimates (KDEs) are a nonparametric alternative to confidence ellipsoids, and their application to two- and three-dimensional data is discussed in turn.

Bivariate KDEs

The use of two-dimensional KDEs for archaeological data presentation is discussed in Beardah & Baxter (1996) and Baxter *et al.* (1997), and has been applied to lead isotope data independently by Beardah (1999) and Scaife (1998). Technical details are given in the Appendix. The appearance of a univariate KDE, such as in Figure 1, is controlled by a smoothing parameter, h_1. For the bivariate case two smoothing parameters, h_1 and h_2, are needed that control the smoothing in two orthogonal directions. A third parameter, h_3, determines the orientation of these directions.

There is less theory to guide the choice of these parameters than for the univariate case, though Wand & Jones (1993) suggest that the simplification $h_3=0$ may often be acceptable, but not $h_1=h_2$. For lead isotope data the $h_3=0$ simplification is generally unsatisfactory. Figure 2 shows a plot of the Kea data for the $^{208}Pb/^{206}Pb$ and $^{207}Pb/^{206}Pb$ ratios with a 90% confidence ellipsoid superimposed. The left-hand side of Figure 3 shows a contour plot of a KDE, calculated assuming $h_3=0$. The contours are approxi-
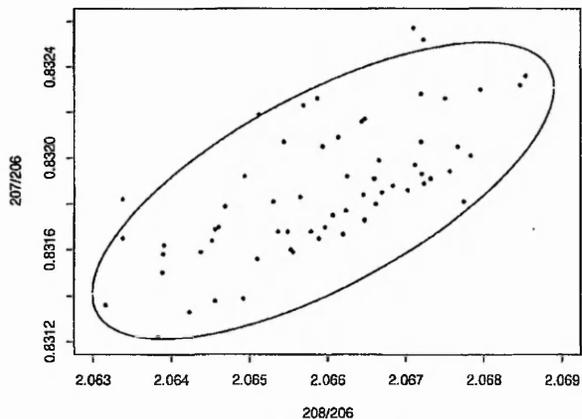

Figure 2. A plot of the $^{208}Pb/^{206}Pb$ and $^{207}Pb/^{206}Pb$ lead isotope ratios for the Kea field with a 90% confidence ellipsoid.

mately elliptical, and suggest a unimodal distribution consistent with the presentation in Figure 2.

However, inspection of Figure 2 reveals that the data have a "natural" orientation (at about 45° to the axes) and that the pattern of points is possibly striated. A bivariate KDE that ignores this may produce misleading results. Recognising this, Scaife (1998) adopted the *ad hoc* solution of transforming to principal components; fitting a bivariate KDE with $h_3=0$; and transforming back. Beardah (1999) uses a non-zero h_3 which is, in principle, more satisfactory but in practice, as he notes, also *ad hoc* in that little guidance exists on the choice of h_3 and some experimentation is necessary. The right-hand diagram shows a contoured KDE for the Kea data using non-zero h_3 and suggests that the data are both non-normal and multimodal.

This example begs questions, both about which is the best representation of the data in Figure 3, and what sample sizes are necessary for the use of bivariate KDEs to be effective. Beardah (1999) addresses the former issue via a series of simulation experiments, using mixtures of bivariate normal distributions. He concludes that the KDE should reflect the natural orientation of the data, and that the right-hand, non-normal representation presents the truer picture. The sample size question is more difficult to address; once again, the problem is that what constitutes an adequate sample depends on the underlying structure of the data. For the data used in Beardah (1999) the sample size, 62, is probably adequate but may be close to the limit of what is satisfactory. Scaife (1998) constructs a more complex, very multimodal, example where even $N=100$ appears inadequate to reflect all the structures in the data.

Trivariate KDEs

Several authors (e.g. Reedy & Reedy, 1992; Scaife *et al.*, 1996) have noted that lead isotope data, being naturally trivariate, are well suited to exploration using
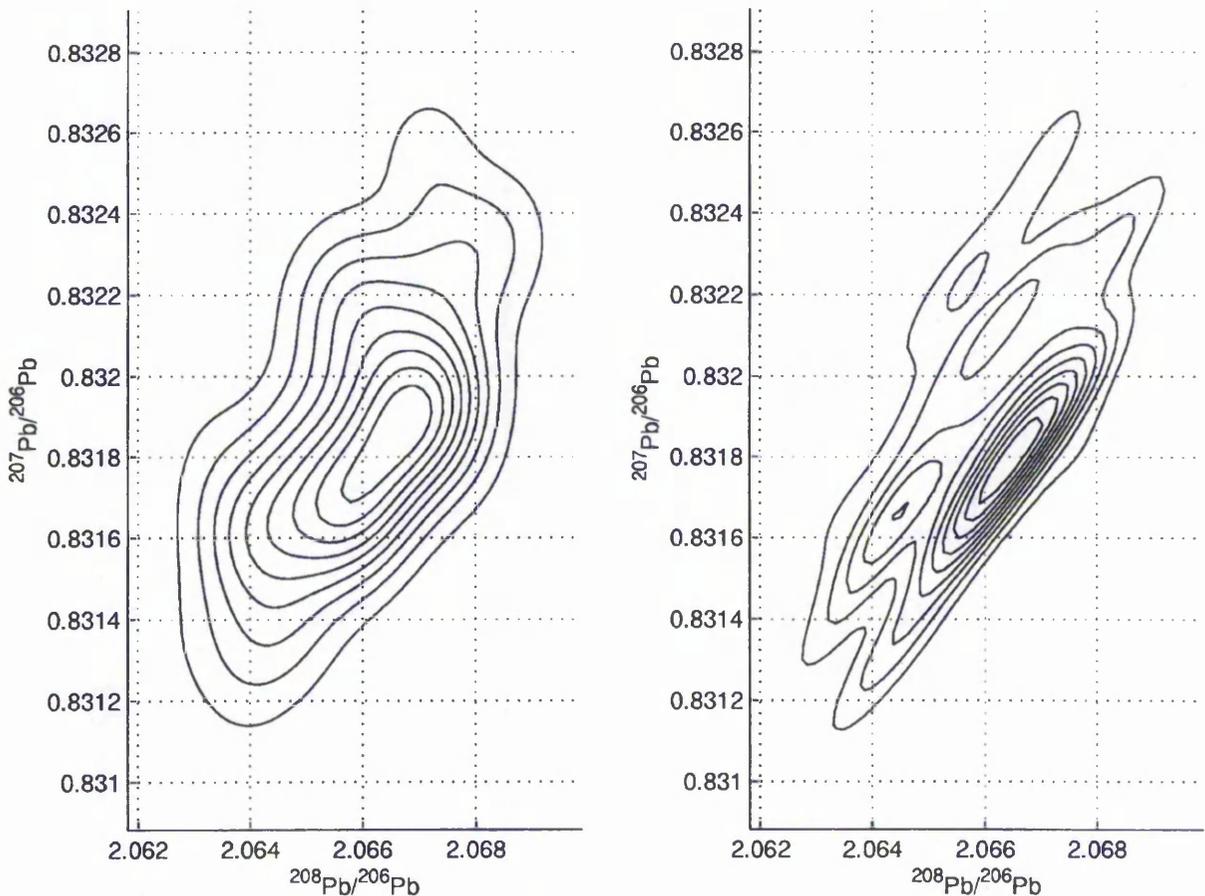
Figure 3. Two bivariate KDEs for the $^{208}Pb/^{206}Pb$ and $^{207}Pb/^{206}Pb$ ratios for the Kea data. The graph on the left has the orientation parameter $h_3=0$; and on the right has $h_3=7 \times 10^{-8}$; $h_1=0.0006415$ and $h_2=0.0001261$ in both cases.

statistical graphics that allow the continuous rotation and labelling of three-dimensional point clouds. Given the use of bivariate KDEs reported in the previous subsection, it is natural to ask if trivariate KDEs might also be used for data visualisation. A k-variate KDE is a $(k+1)$ dimensional construct (see the Appendix); for example, in Figure 1 the univariate KDE is displayed in two dimensions. The bivariate KDEs used in the previous section were reduced to two-dimensional displays through the use of contouring.

For trivariate KDEs a direct four-dimensional representation is not possible, but three-dimensional contour shells that enclose some percentage of the most dense points of an estimated KDE can be used. Figure 4 illustrates the idea, using the Lavrion data ($N=59$). The 50% shell isolates two separate clouds of points suggesting non-normality in the form of bimodality. This is consistent with uni- and bivariate analyses, and the 70% inclusion shell in Figure 4 confirms the impression. Since tests of normality provide reassurance that the data are non-normal, displays

such as those in Figure 4 provide a convenient way of exploring the form of non-normality.

The use of trivariate KDEs in isolation for investigating field structure is possible but may be limited by sample size considerations. Some limited simulation results, reported in Beardah & Baxter (1999), based on simulating samples of size 60 from a mixture of trivariate normals (to mimic the structure of the Lavrion data), suggest that non-normality is clear about 85% of the time. This suggests that larger samples than are commonly available may be needed to detect even this relatively simple kind of departure from normality with a high degree of confidence, using purely graphical methods.

Effects of non-normality on probability calculations

As well as giving rise to misleading graphical presentations, an incorrect assumption of normality can give rise to misleading probability calculations. Scaife *et al.* (1996) argue that lead isotope fields have

50% contour shell   70% contour shell
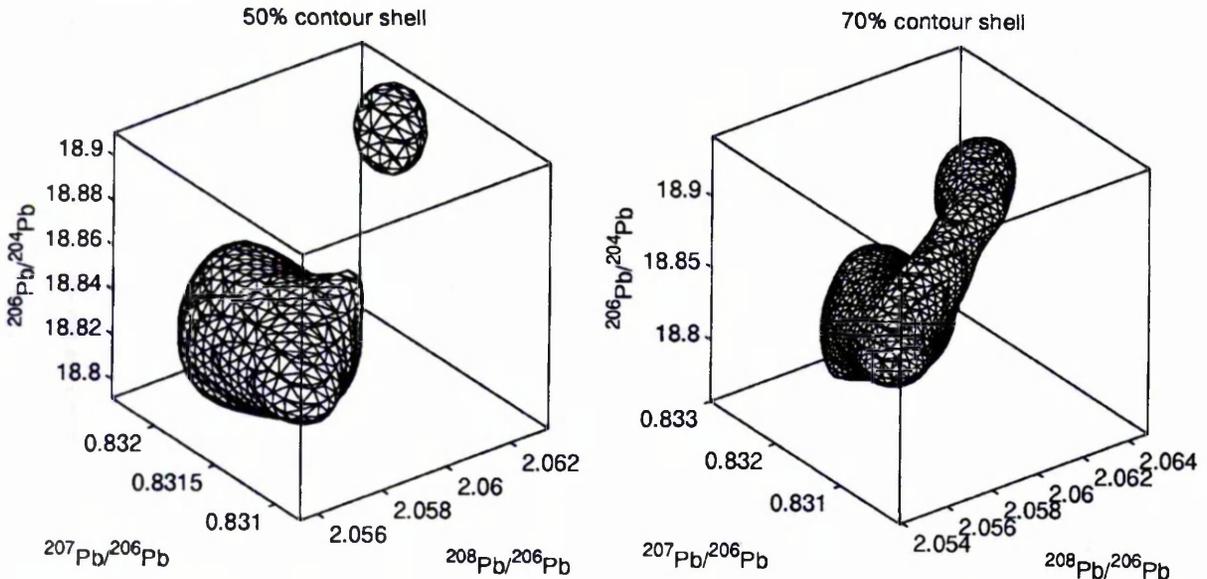

Figure 4. Two contour shells based upon a trivariate KDE for the full Lavrion data set showing 50% (left) and 70% (right) contour shells. $h_1=0.0011$, $h_2=0.0002$, $h_3=0.0173$.

sharp boundaries, so that any case lying outside the boundaries (once measurement error is allowed for) cannot belong to that field (i.e. has zero probability of membership), however close it appears to be visually, and whatever normal-based calculations suggest. In other words, any probability calculation is dependent on the model on which it is based.

Even if the specific model proposed by Scaife *et al.* (1996) is rejected, the more general point about the dependency of probabilities on specific models remains valid. Figure 1 suggested that the Lavrion was bimodal. Figure 5 shows a bivariate plot of the 208/206 and 207/206 ratios, labelled according to which of two

Figure 5. A plot of the ^{208}Pb/^{206}Pb and ^{207}Pb/^{206}Pb ratios for the Lavrion field, labelled (1 or 2) according to which of two concentrations cases seem to belong to (in three dimensions). An artificial data point has been created and is labelled 0.

main concentrations a case appears to belong to (suggested by Figure 4). A new data point (2.0630, 0.8313, 18.90) has been constructed, and is shown on the plot close to the smaller of the two groups.

Assuming normality of all the data the estimated probability that the new case could come from the field is 0.137, leading to the conclusion that the new case could comfortably belong to the field. If, however, we treat the field as a two-component mixture of normals, and argue that calculations should be based on the component nearest to the new case, a probability of 0.048 is obtained, leaving one in doubt about the field as a possible provenance. If the larger, more distant, component is used as the basis for calculation the probability of membership is estimated as 0.001. It must be emphasised that the force of this argument does not depend on the precise truth of the assumptions made; it simply demonstrates that results can depend in a non-trivial way on assumptions, whatever these might be. To some this conclusion may seem obvious; it does show that normal-based procedures cannot be assumed to be robust to the normality assumption.

Discussion

The results of Baxter & Gale (1998) and Baxter (1998) suggested that non-normality of lead isotope fields was the rule rather than the exception. That this had not always been previously accepted or demonstrated may be attributed to the small sample sizes with which researchers have had to operate. That 20 is, ideally, a minimum sample size requirement seems to have been widely accepted.
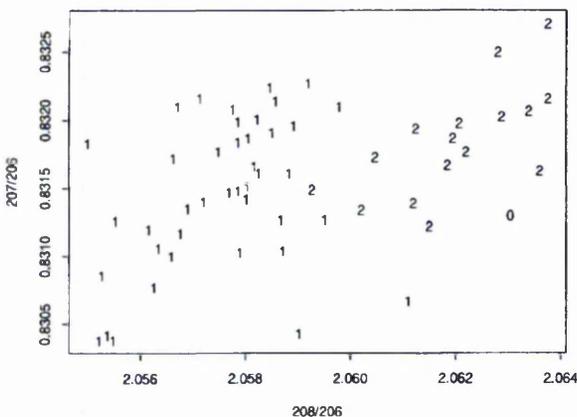
The argument in this paper has been that this may be acceptable if lead isotope fields have a normal distribution, but may be seriously inadequate otherwise. This latter contention is supported by a variety of examples. If normality is assumed and the sample size is not large enough to investigate this assumption (and 20 will usually be inadequate) the use of confidence ellipsoids to estimate the extent of fields, and probability calculations to investigate likely provenances, may be misleading.

It is impossible to make any precise statement about what is an adequate size in any specific case, as this depends on the (unknown) model appropriate for describing a field. For two of the cases studied here, Lavrion and Kea, sample sizes in excess of 40 and 50 would appear to be necessary to identify non-normality with reasonable confidence, and simulation results confirm that samples of this size will be needed to detect non-normality even when this is quite obvious in the population. It is important to emphasise that the departures from normality being discussed are not trivial ones that have little effect on normal-based procedures. The case of the old Cypriot field provides a salutary instance, where even a sample in excess of 40 was inadequate for revealing some quite clear differences between the ore sources that comprised it.

Bivariate KDEs have been explored as a non-parametric alternative to confidence ellipsoids that do not impose inappropriate structure on the data. Our work, and that of Scaife (1998), shows that the effective use of KDEs requires quite demanding sample sizes (in the context of lead isotope analysis), quite apart from technical problems concerning the choice of smoothing and orientation parameters. Once again, the sample sizes needed depend on the data structure; our examples appear to need about $N=60$, and Scaife (1998) presents an example where $N=100$ is not really adequate. The possibility of using trivariate KDEs has been noted, but these are even more demanding of data.

In summary, given the possibility of non-normality, it has to be concluded that the widely recommended minimum sample size of 20 may often be seriously inadequate, if the normality assumption is to be checked and/or if methods not based on the normality assumption are to be used.

For small sample sizes (say $N \leq 20$) normality cannot easily be checked and it may be best to avoid multivariate methods that assume them, as they possibly impart a spurious sense of authority to the results obtained. The use of simple graphical inspection has been advocated, both for assessing the distinctiveness of fields and for assessing possible provenance (Scaife et al., 1996). Although decision-making in such cases may, on occasion, be subjective and uncertain, the uncertainty is unlikely to be any greater than that associated with more "objective" inferential statistical procedures where the underlying assumptions are incorrect.

For larger sample sizes (say $N>40$) formal testing of normality is possible. If such tests do not reject the hypothesis of normality, normal-based procedures can be considered as an option. (Note that for the old Cyprus field, tests were indicative of non-normality: Scaife et al., 1996.) Where non-normality is evident KDEs provide an alternative method of data presentation, but should not be used in isolation and may need sample sizes somewhat greater than 40 to be fully effective.

For intermediate sample sizes, $20<N<40$, precise and positive guidelines are difficult to lay down. Testing normality is certainly possible but the tests may have low power against even quite non-normal alternative models. Normal-based procedures can be misleading in the presence of undetected non-normality, but non-parametric KDEs may also be difficult to interpret.

References

Baxter, M. J. (1998). On the multivariate normality of data arising from lead isotope fields. *Journal of Archaeological Science* 26, 117–124.

Baxter, M. J. & Gale, N. H. (1998). Testing for multivariate normality via univariate tests: a case study using lead-isotope ratio data. *Journal of Applied Statistics* 25, 679–691.

Baxter, M. J., Beardah, C. C. & Wright, R. V. S. (1997). Some archaeological applications of kernel density estimates. *Journal of Archaeological Science* 24, 347–354.

Beardah, C. C. (1999). Uses of multivariate kernel density estimates in archaeology. In (L. Dingwall, S. Exon, V. Gaffney, S. Laflin & M. van Leusen, Eds) *Archaeology in the Age of the Internet: CAA 97*. Oxford: BAR International Series 750, p. 107.

Beardah, C. C. & Baxter, M. J. (1996). The archaeological use of kernel density estimates. *Internet Archaeology* 1 (http://intarch.ac.uk/journal/issue1/beardah_index.html).

Beardah, C. C. & Baxter, M. J. (1999). Three-dimensional data display using kernel density estimates. In (J. A. Barcelo, I. Briz & A. Vila, Eds) *New Techniques for Old Times: CAA 98*. Oxford: BAR International Series 757, pp. 163–169.

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

Bowman, A. W. & Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis*. Oxford: Oxford Science Publications.

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

Budd, P., Pollard, A. M., Scaife, B. & Thomas, R. G. (1995). Oxhide ingots, recycling and the Mediterranean metals trade. *Journal of Mediteranean Archaeology* 8, 1–32.

Gale, N. H. & Stos-Gale, Z. A. (1992). Evaluating lead isotope data: comments I. *Archaeometry* 34, 311–317.

Gale, N. H. & Stos-Gale, Z. A. (1993). Evaluating lead isotope data: further observations: comments II. *Archaeometry* 35, 252–259.

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, 237–246.

Harbottle, G. (1976). Activation analysis in archaeology. *Radiochemistry* 3, 33–72.

Leese, M. N. (1992). Evaluating lead isotope data: comments II. *Archaeometry* 34, 318–322.

Malkovich, J. F. & Afifi, A. A. (1973). On tests for multivariate normality. *Journal of the American Statistical Association* 68, 176–179.

Pollard, A. M. & Heron, C. (1996). *Archaeological Chemistry.* Cambridge: The Royal Society of Chemistry.

Reedy, T. J. & Reedy, C. L. (1992). Evaluating lead isotope data: comments IV. *Archaeometry* **34**, 327–329.

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**, 73–105.

Sayre, E. V., Yener, K. A. & Joel, E. C. (1995). Comments on "Oxhide ingots, recycling and the Mediterranean metal trade". *Journal of Mediterranean Archaeology* **8**, 45–53.

Scaife, B. (1998). *Lead isotope analysis in archaeology.* Ph.D. Thesis. University of Bradford.

Scaife, B., Budd, P., McDonnell, J. G., Pollard, A. M. & Thomas, R. G. (1996). A reappraisal of statistical techniques used in lead isotope analysis. In (S. Demirci, A. M. Ozer & G. D. Summers, Eds) *Archaeometry 94: Proceedings of the 29th International Symposium on Archaeometry. Ankara 9–14 May 1994.* Ankara: Tubitak, pp. 301–307.

Sheather, S. J. & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society B* **53**, 683–690.

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**, 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**, 83–123.

Tite, M. S. (1996). In defence of lead isotope analysis. *Antiquity* **70**, 959–962.

Titterington, D. M., Smith, A. F. M. & Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* New York: Wiley.

Wand, M. P. & Jones, M. C. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association* **88**, 520–528.

Appendix

Let **v** be a k-dimensional variable whose ith realisation will be denoted as X_i, where $X_i = x_i$, $(x_i, y_i)^T$ or $(x_i, y_i, z_i)^T$ for $k=1,2,3$, respectively, and ^{T} is the vector or matrix transpose. A k-variate kernel density estimate (KDE) may be written in the general form

$$\hat{f}(\mathbf{v}) = N^{-1}|\mathbf{H}|^{-1/2}\sum_{i=1}^{N} K(\mathbf{H}^{-1/2}(\mathbf{v}-X_i)) \quad (1)$$

where **H** is a symmetric positive definite $k \times k$ matrix with ith diagonal h_i^2 and $h_i>0$; $K(\cdot)$, the kernel function, is the standard k-variate normal probability density function; and N is the sample size.

For the univariate case, $k=1$, there is a single smoothing parameter h_1 that determines the precise appearance of the KDE; Figure 1 provides an example. Other kernels than the normal are possible, but the appearance of the KDE is relatively insensitive to this choice. The choice of the smoothing parameter is more important and the selection method proposed by Sheather & Jones (1991), which works well in practice, has been used.

For the general bivariate case, $k=2$, **H** has off-diagonal elements h_3 that control the orientation of the KDE. The right-hand side of Figure 3 provides an example and may be contrasted with the left-hand side where $h_3=0$ is assumed. In this latter case the KDE takes a simpler form that generalises, for the trivariate case, to

$$\hat{f}(x,y,z) = \frac{1}{Nh_1h_2h_3}\sum_{i=1}^{N} K\left(\frac{x-x_i}{h_1}, \frac{y-y_i}{h_2}, \frac{z-z_i}{h_3}\right) \quad (2)$$

and it is this form that is used to define the contour shells shown in Figure 4.

In the univariate case the plot is based on pairs of points of the form $(x, \hat{f}(x))$, so that a two-dimensional plot is obtained. The estimated density, $\hat{f}(x)$, is the "height" at x on the plot. For the bivariate case, triplets of points, $(x,y,\hat{f}(x,y))$, can be represented in two-dimensions by contouring the resultant three-dimensional construct at equally spaced heights or densities. The trivariate case gives rise to a four-dimensional construct based on $(x,y,z,\hat{f}(x,y,z))$ that can be represented in three-dimensions by selecting some specified percentage of the data, $q\%$ say; determining the density estimate that defines the most dense $q\%$ of the data; and representing this as a three-dimensional contour shell or envelope. It is advisable to look at this for several choices of q, as in Figure 4.

The KDEs presented in this paper were obtained using routines written by the second author for the MATLAB package. Earlier versions of some of these can be downloaded from the online paper Beardah & Baxter (1996); updated versions can be obtained by e-mailing christian.beardah@ntu.ac.uk. A library of S-Plus functions for obtaining KDEs, written to accompany the book by Bowman & Azzalini (1997), can be obtained from World Wide Web sites listed in that book.