Uses of Multivariate Kernel Density Estimates in Archaeology

Christian C. Beardah

Abstract

Though the technique of Kernel Density Estimation is now well established, it is still under-employed by archaeologists. For multivariate data this may be because the methodology is less well developed than in the case of univariate data. In this paper we aim to illustrate through archaeological examples some uses of multivariate Kernel Density Estimates (KDEs). In particular we shall examine the important issue of smoothing parameter selection in the bivariate case. Recent developments in the automatic data-based selection of smoothing parameters will be reviewed and particular emphasis will be placed on an illustration of the role of the so-called orientation parameter. Methods for the automatic selection of this parameter are in their infancy. We shall illustrate the effect it can have on the appearance of the KDE and therefore the archaeological conclusions drawn from it.

1 Introduction

For given bivariate data $\underline{X}_1 = (x_1,y_1)^T, ..., \underline{X}_n = (x_n,y_n)^T$ a bivariate KDE is formed by placing a "bump" at each data point (see the top left sub-figure of figure 1). The value of the KDE at any point $\underline{v} = (x,y)^T$ in the plane is found by summing the height of bumps which pass above the point $\underline{v}$. In the most general terms this is mathematically expressed as

$$\hat{f}(\underline{v}) = n^{-1}|H|^{-1/2} \sum_{i=1}^{n} K(H^{-1/2}(\underline{v} - \underline{X}_i)) \qquad (1)$$

where H is a symmetric positive definite 2 by 2 matrix with structure

$$H = \begin{bmatrix} h_1^2 & h_3 \\ h_3 & h_2^2 \end{bmatrix}.$$

Here $h_1, h_2 > 0$ and $|h_3| < h_1 h_2$. This means that we can write $h_3 = r h_1 h_2$ where $-1 < r < 1$. The values h_1, h_2 and h_3 are called the *smoothing parameters*.

The shape of the bump is defined by a mathematical function, the *kernel*, denoted by $K(v)$. This kernel function is usually a bivariate probability density function (pdf) such as the bivariate normal pdf, given by

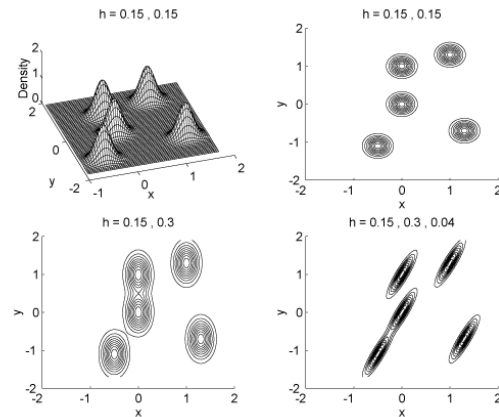$$K(\underline{v}) \equiv K(x,y) = (2\pi)^{-1} \exp\left(-\frac{1}{2}(x^2 + y^2)\right).$$

Figure 1. KDEs based upon a small dataset for illustration.

Such functions have the property that their volume is 1. The appearance of the KDE is not greatly influenced by the choice of kernel function (see Silverman (1986), or Wand and Jones (1995)). We shall therefore use the bivariate normal pdf throughout this paper. However the appearance of (and therefore the archaeological conclusions inferred from) the KDE does depend crucially upon the values of the smoothing parameters h_1, h_2 and h_3. Three smoothing strategies are possible.

1. Using $h_1 = h_2 = h$ and $h_3 = 0$ (a single smoothing parameter) causes the bumps which form the KDE to be *spherically symmetric*, i.e. they have circular contours. Smoothing is the same in both directions. With this simplification the representation of the bivariate KDE is given by

$$\hat{f}(x,y) = \frac{1}{nh^2} \sum_{i=1}^{n} K(\frac{x-x_i}{h}, \frac{y-y_i}{h}).$$

2. Using h_1, h_2 and $h_3 = 0$ (two distinct smoothing parameters) causes the bumps to have elliptical contours, where the ellipsoidal axes are parallel to the co-ordinate axes. With this simplification the representation of the bivariate KDE is given by

$$\hat{f}(x,y) = \frac{1}{nh_1h_2} \sum_{i=1}^{n} K(\frac{x-x_i}{h_1}, \frac{y-y_i}{h_2}).$$

3. Using three distinct smoothing parameters h_1, h_2 and h_3 causes the bumps to have elliptical contours, where the orientation of the ellipsoidal axes is controlled by the value of h_3. The KDE is given by equation (1).

Figure 1 shows KDEs based upon a small bivariate dataset consisting of just five points. (A small dataset has been used for illustration only. In normal circumstances it would be unwise to apply KDE methods to such a small dataset.) The top left and top right sub-figures show surface and contour representations of the KDE based upon a single h value. The bottom left and bottom right sub-figures show contour plots of the KDE with two and three h values respectively. It is clear from the figure that h_1 and h_2 control the dimensions of each bump in the two co-ordinate directions while $h_3 = r\,h_1\,h_2$ controls the orientation (or more exactly, the correlation) of the bump. In the bottom right sub-figure the value of r is $h_3 / (h_1\,h_2) = 0.04 / (0.15 \times 0.3) \approx 0.89$, giving rise to a bump with high positive correlation. Using $r = -0.89$ leads to bumps with a high negative correlation. Finally, using $r = 0$ and hence $h_3 = 0$, leads to the bumps shown in the bottom left sub-figure.

Wand and Jones (1995) state that in general it is bad practise to use a single smoothing parameter for both co-ordinate directions. Furthermore, while in many cases two smoothing parameters will lead to an adequate representation of the underlying density, there is a strong case for using all three smoothing parameters in certain circumstances.

For example, figure 2 shows contour plots of KDEs based upon data representing two ratios from the Kea lead isotope field (see Stos-Gale et al. (1996)). For this dataset $n = 62$. Contour lines join points on the KDE surface with equal height, and the heights at which contours are drawn are equally spaced. Other methods of contouring exist, for example see

Bowman and Foster (1993) for an alternative methodology and Baxter and Beardah (1995), Baxter et al. (1997) for some archaeological examples.
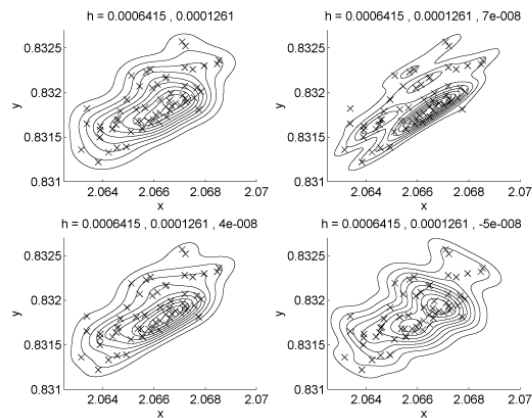
Figure 2. KDEs based upon the Kea lead isotope field data with various choices of h_3.

In figure 2 the values of h_1 and h_2 shown above each plot were chosen using an automatic data-based h selection algorithm. (Specifically a Direct Plug-in (DPI) rule, see section 2 for more details.) Keeping h_1 and h_2 fixed, figure 2 shows the effect of varying the value of h_3. Proceeding clockwise from the top left sub-figure, the values of r used are respectively $r = 0$, 0.87, -0.62, 0.49.

It is clear from figure 2 that the value of h_3 can make a great deal of difference to the conclusions we can draw from these KDEs. Two of the KDEs suggest a unimodal structure to the underlying density, while the other two KDEs suggest very different bimodal structures. The question is, which of the KDEs in figure 2 is the most appropriate? We shall attempt to answer this question later. For now we simply note that the dataset has a natural orientation which is not parallel to the co-ordinate axes. When forming a KDE based upon such data it seems sensible to use kernel "bumps" which reflect this orientation.

In section 2 we briefly review some methods for the automatic selection of smoothing parameters in the bivariate case. These methods are illustrated using an example. Finally, in section 3, we present the results of simulations which demonstrate the importance of the smoothing parameter h_3.

2 Automatic smoothing for bivariate data

Methods for the automatic data-based choice of h_3 are in their infancy and we shall not discuss such

methods here. Instead we offer a brief overview of available methods for the automatic choice of h_1 and h_2.

The most basic method of choosing smoothing parameters h_1 and h_2 is to use a univariate method of h selection for each variable separately. By doing so we can draw upon several well known methods (see Jones et al. (1996) for technical details and Baxter and Beardah (1996) for some archaeological examples). This technique at least gives initial values for h_1 and h_2 which can be further adjusted, subjectively if necessary. However, it is often the case that the appearance of a KDE is fairly robust to the choice of smoothing parameters. Because of this, in many applications univariate methods give reasonable results when compared to the bivariate methods detailed below.

Currently one of the best univariate h selection methods is the "solve the equation" method of Sheather and Jones (1991). Unfortunately this method does not readily extend to the bivariate case. However the closely related Direct Plug-in methods do have bivariate (indeed multivariate) extensions which we briefly discuss here.

Under certain non-restrictive assumptions on the unknown density f it is possible to develop a simple asymptotic (large sample) approximation to the Mean Integrated Squared Error (MISE) of a bivariate KDE. This is simply a measure of the amount by which the true density f and the KDE $\hat{f}$ differ. Since the KDE, $\hat{f}$, depends upon the values of h_1 and h_2, then so does the asymptotic MISE, denoted by A(h_1,h_2). Explicit expressions for the values of h_1 and h_2 which minimise the function A can be found. These expressions depend upon quantities φ_{y1} which in turn depend upon the unknown true density f. For particular values of r_1, the value of φ_{y1} can be estimated using KDEs based upon a single smoothing parameter g. Unfortunately the formula for g depends upon a different φ value, say φ_{y2}. To estimate φ_{y2} we again use a KDE where the value of the single smoothing parameter depends upon a further φ value, say φ_{y3}. This circular calculation process can be resolved at some stage by "plugging in" a simple estimate for the latest φ_r value, where the estimate is usually formed by assuming that the true density is normal. The stage at which this simple estimate is plugged-in determines the stage of the so-called *Direct Plug-in* method which results. Theoretical considerations (see Wand and Jones (1995)) lead to a recommendation that at least two stages are used. If the simple estimate is made immediately (at stage

zero) then the resulting method is called the *normal scale* (NS) rule. This is the most basic method of h selection and can also be used in the univariate case (see Silverman (1986)).

Upon application to replicated samples from known densities, bivariate DPI methods result in values of h_1 and h_2 which have low variability. However there does seem to be a tendency towards over-smoothing, i.e. h_1 and h_2 values which are too large, especially in cases where the density consists of widely separated modes (see section 2.1 below for an illustration).

In fact it is interesting to speculate whether use of the univariate DPI rule to separately determine h_1 and h_2 may in some way compensate for the inherent tendency of the bivariate DPI rule to oversmooth. The formulae for h_1 and h_2 used by the bivariate DPI rule essentially reduce to

$$h_1 = K_2 n^{-1/6},$$

$$h_2 = L h_1,$$

where K_2 and L are constants which depend on the data. On the other hand, the formula for the univariate DPI rule can be reduced to $h = K_1 n^{-1/5}$ where K_1 is a data dependent constant. The important thing to note here is the difference in the power to which n is raised in the two formulae. Since $n^{-1/5} < n^{-1/6}$ the univariate DPI rule tends to result in smaller smoothing parameters than the bivariate DPI rule (see section 2.1 below for an illustration).

Another univariate method of h selection which generalises to the multivariate case is that of biased cross-validation (CV), see Sain et al. (1994). This method works by finding the values of h_1 and h_2 which minimise a *criterion function*, B(h_1,h_2). The criterion function has the same form as the asymptotic MISE function, A(h_1,h_2), except that terms involving the unknown true density f are replaced by a cross-validation approximation. For a dataset with n data points, cross-validation involves using information about n reduced datasets (each with n-1 data points, formed by leaving one of the original data points out) to give information about the dataset as a whole.

2.1 An archaeological example

Figure 3 shows KDEs based upon four different h selection routines. These data represent the co-

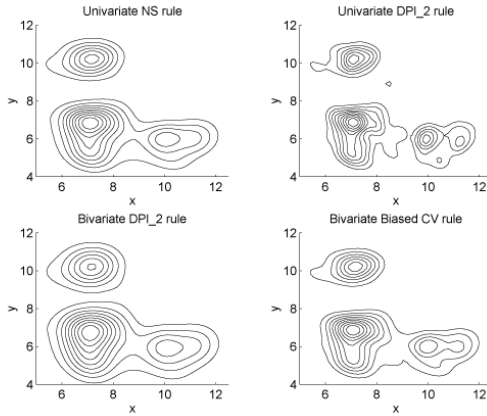ordinates of 276 bone splinters and are a subset of Binford's (1978) Mask Site data.

Figure 3. KDEs generated by four different h selection routines.

While the four KDEs share features such as the clustering of the dataset into three main groups, only the univariate 2-stage DPI rule (and possibly the bivariate biased CV rule) split the lower right cluster into two further groups. Since these two sub-groups correspond to two separate hearth locations we can be reasonably sure that this is the true nature of the density and that h selection methods which do not reveal this structure are over-smoothing the KDE. In particular, note that the bivariate 2-stage DPI rule seems to have over-smoothed the KDE (as might have been expected with such widely separated modes in the underlying density). The bivariate 2-stage DPI rule gives $h_1 = 0.53$, $h_2 = 0.57$ while separately applying the univariate 2-stage DPI rule to each variable gives $h_1 = 0.24$, and $h_2 = 0.30$.

3 Choice of orientation parameter h_3

In order to investigate the effect of the smoothing parameter h_3 on the resulting KDE, we now present the results of some simulations using known densities. The basic idea is as follows. Given a sample from a known density we generate several KDEs based upon different values of h_3 (keeping h_1 and h_2 fixed). For each sample we can then find which value of h_3 (or equivalently, r) minimises the error between the KDE and the true density. This process is repeated for many samples (in practise 100 repetitions were made). The measure of error that we use is the *integrated squared error* (ISE), i.e. the volume under the surface defined by $(f - \hat{f})^2$, where f is the true density and $\hat{f}$ is the approximate density given by the KDE.
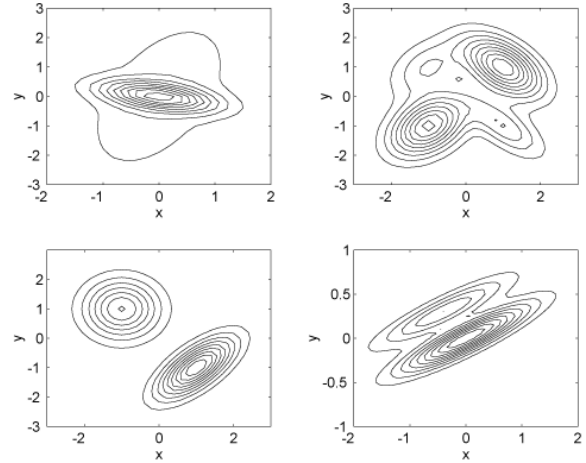

Figure 4. Four example bivariate NMDs.

3.1 Simulations using bivariate normal mixture densities

The known densities which we sample from are the family of bivariate normal mixture densities (NMDs). These are given by

$$f(x,y) = \sum_{i=1}^{k} w_i N_{\underline{\mu}_i, \Sigma_i}(x,y),$$

where k is the number of normal densities used in the mixture, and $N_{\underline{\mu}_i, \Sigma_i}$ denotes the bivariate normal pdf with mean $\underline{\mu}_i$ and covariance matrix Σ_i, and

$$\sum_{i=1}^{k} w_i = 1.$$

The family of NMDs can be used to mimic all other type of density. For example figure 4 shows contour plots of four example bivariate NMDs (three of which are taken from Wand and Jones (1993)).

The density in the lower right sub-figure is given by a mixture of $k = 2$ densities where $w_1 = 2/3$, $w_2 = 1/3$ and

$$\underline{\mu}_1 = (0,0)^T,$$
$$\underline{\mu}_2 = (-0.4, 0.3)^T,$$
$$\Sigma_1 = \Sigma_2 = \begin{bmatrix} 0.75^2 & 0.9 \times 0.75 \times 0.25 \\ 0.9 \times 0.75 \times 0.25 & 0.25^2 \end{bmatrix}.$$

This density will be denoted NMD_1. Note that both of the densities defining NMD_1 have the same covariance matrix with a correlation coefficient of

$r_1 = 0.9$. Samples from NMD_1 show similarity of structure to that of the Kea lead isotope data illustrated in figure 2, for which $n = 62$.

As described earlier, 100 samples (of size $n = 100$) were taken from NMD_1. For each sample KDEs were formed using h_1 and h_2 as automatically selected using the 2-stage DPI rule described in section 2 and various values of h_3 (or equivalently, r). The value of r, $\hat{\rho}$, which minimised the ISE between the true density and the KDE was recorded for each sample. The lower of the two histograms in figure 5 shows how the 100 values of $\hat{\rho}$ were distributed. The upper histogram shows the distribution of $\hat{\rho}$ when the univariate normal scale rule is used to select h_1 and h_2 for each sample.
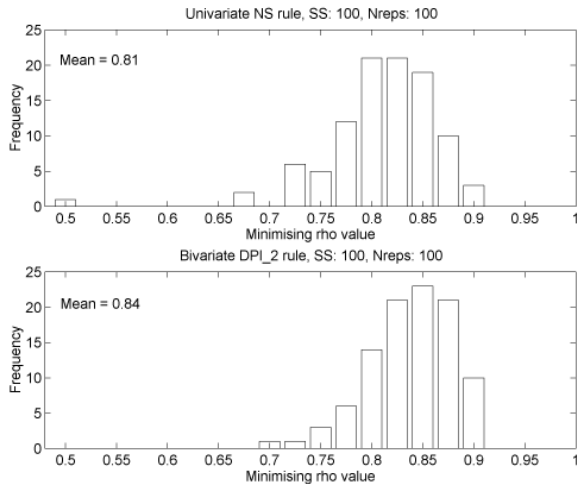
**Figure 5. Histograms of $\hat{\rho}$ values for 100 samples from NMD_1.**

Figure 6a shows a typical graph of ISE against r for a sample from NMD_1. The lower (dashed) curve illustrates how the ISE varies when the 2-stage DPI rule is used to select h_1 and h_2. For this sample the error between the true density and the KDE is minimised when we use a r value of approximately 0.88 (i.e. $h_3 = 0.88 h_1 h_2$). So for this sample, $\hat{\rho} = 0.88$. It is clear that the error is minimised when the correlation of the bumps used to form the KDE (here 0.88) agrees closely with the correlation of the modes in the true density (here 0.9).
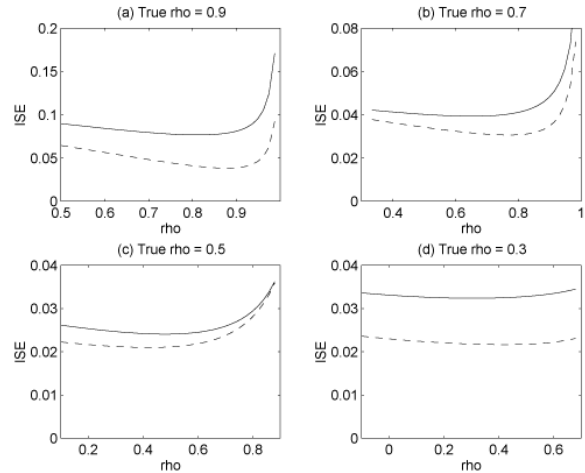
Figure 6. Typical graphs of ISE against r for samples taken from four different bimodal NMDs. Solid lines: univariate NS rule, dashed lines: bivariate 2-stage DPI rule.

On the other hand, it is clear from figure 6a that the ISE curve rises only slowly to the left of the minimum. This suggests that for this sample there is a relatively wide range of values of r (or h_3) which give rise to a "small" error. This is more noticeably the case when the univariate normal scale rule is used to select h_1 and h_2 (the solid curve in figure 6a). Also notice that, as might be expected, h selection via the bivariate 2-stage DPI rule gives rise to lower errors than h selection via the univariate normal scale rule. Finally, note that the simulation described here was repeated with a sample size of $n = 62$ (the same as the size of the Kea lead isotope field data), however no significant difference in results was apparent.

The same procedure outlined above was repeated by taking 100 samples (of size $n = 100$) from three further bimodal NMDs where the two normal densities within each mixture both have correlation $r_2 = 0.7$, $r_3 = 0.5$ and $r_4 = 0.3$ respectively. Figure 7 shows histograms which illustrate how the minimising value of r, $\hat{\rho}$ was distributed for these simulations. For completeness the case for NMD_1 (where the two densities within the mixture have correlation $r_1 = 0.9$) is also included. In each case the values of h_1 and h_2 were chosen automatically via the 2-stage DPI rule. Figure 6b-d shows typical graphs of ISE against r.
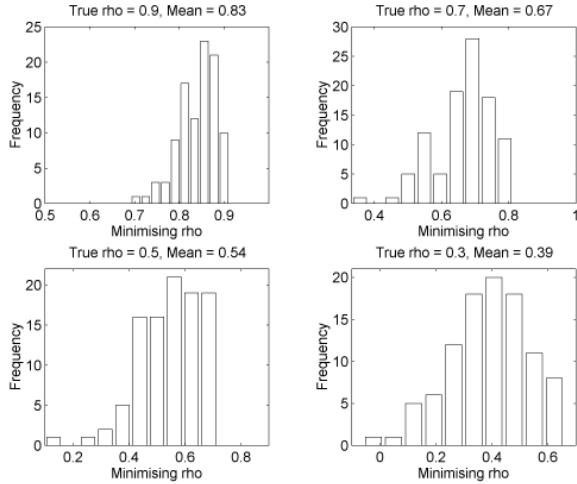
Figure 7. Histograms illustrating the distribution of $\hat{\rho}$ based upon 100 samples of size $n = 100$ from four different NMDs.

Based upon figures 6 and 7 we can make three observations.

1. The minimising value of r is clearly distributed around the correlation of the underlying densities within the NMD (the "true r" value).
2. As the true correlation decreases (i.e. the true value of r gets closer to zero) the distribution of the minimising value of r becomes more widely spread.
3. As the true correlation decreases we find that a large range of values of r give rise to a small error (the curves of ISE against r in figure 6 are very flat). Furthermore, the size of the error is generally smaller than for cases where the true correlation is high.

Wand (1992) has shown that in order to optimally estimate a bivariate normal density, one should use kernel "bumps" with the same covariance structure as the density itself. Since the densities used here are mixtures of two bivariate normal pdfs *with the same covariance structure* it is also natural to use kernel "bumps" with this covariance structure. However, what should our strategy be if a dataset consists of sub-groups with different covariance structure? (An example of this would be a sample from the NMD in the bottom left of figure 4.) An intuitive approach would be to use differently oriented kernel "bumps" within each sub-group. Unfortunately this approach assumes that we know to which sub-group each data point should be assigned, which is a difficult problem in its own right, and may indeed have been the point of the data analysis in the first place!

Another approach would be to use a single overall "bump" shape based upon the covariance structure of the dataset as a whole. This approach is dismissed as "inappropriate" by Wand and Jones (1994) who argue that the overall covariance structure can be quite different from the structure of individual sub-groups within the dataset. As a simple illustration of this fact we note that the average value of the correlation coefficient for the 100 samples from NMD_1 discussed earlier is $\hat{\rho} = 0.63$. By contrast, the true correlation coefficient of the two bivariate normal pdfs which make up the underlying density is 0.9.

3.2 An archaeological example

Figure 8 shows four KDEs based upon the rim diameter and overall height of $n = 60$ Bronze Age Italian cups. These data are a subset of material originally published by Lukesh and Howe (1978).

The values of h_1 and h_2 used to form the KDEs in figure 8 were calculated using the 2-stage bivariate DPI rule. (This could be expected to over-smooth the KDEs as here the underlying density has two modes with a reasonable amount of between-mode separation. However, when other h selection rules are applied to this dataset, the resulting values of h_1 and h_2 are not significantly different to those used here.) Proceeding clockwise from the top left sub-figure, the values of h_3 are -0.5, 0, 0.75 and 0.5 respectively. These values correspond to r = -0.52, 0, 0.78 and 0.52 respectively. It can be seen from figure 8 that although a wide range of h_3 values have been used, the overall appearance of the KDE does not change by much. In particular, the overall conclusion that there are two major groups within the dataset is independent of the value of h_3 (or r) used.
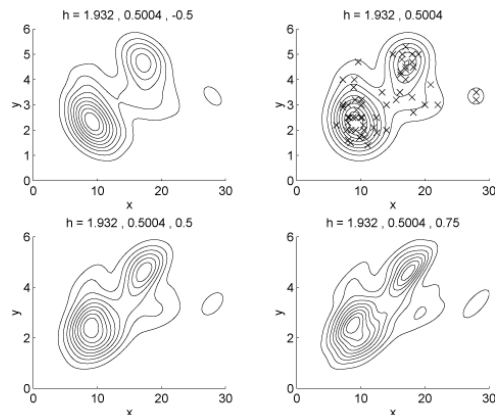
Figure 8. Four KDEs exhibiting little variation in structure as h_3 is changed.

Clearly this dataset consists of two sub-groups without any strong correlation (since the sub-groups are relatively circular). For this reason we can use a wide range of values of h_3 without significantly altering the resulting KDE. In fact one could argue that there is no need to use the third smoothing parameter in this case and that two smoothing parameters (illustrated by the top right sub-figure) are adequate.

4 Conclusions

Since there are currently no automatic data-based methods of selecting the smoothing parameter h_3, we are forced to rely upon more ad-hoc approaches. Fortunately, if a dataset has structure along the lines of the Kea lead isotope field data, or samples from NMD_1 (where sub-groups within the dataset share some natural orientation) then it is a relatively straightforward matter to choose an appropriate value of h_3 "by eye". Bearing this in mind, let us now reconsider our earlier example of the Kea lead isotope field data. We asked which of the KDEs in figure 2 is the most appropriate model of the true underlying density. Clearly this dataset has a strong natural orientation and we feel that the KDE in the top right of figure 2 best reflects this. The value of h_3 chosen has the effect of lining up the kernel "bumps" in the same direction as the natural orientation of the dataset. Note that the possible modality of the underlying density only becomes apparent upon making this choice of smoothing parameter.

On the other hand, if the dataset has no natural orientation, or consists of sub-groups with no natural orientation, then it is usually the case that a wide range of values of h_3 should give satisfactory results (see section 3.2). Indeed it may be more sensible to use only two smoothing parameters in these cases.

Finally, if a dataset consists of sub-groups with different natural orientations (for example, samples from the bottom left NMD of figure 4) then the situation is less clear. Unless a categorisation of the dataset exists which makes it possible to use different values of h_3 between sub-groups, then it is possibly best to use only two smoothing parameters in these cases.

Bibliography

Baxter, M J, and Beardah, C C, 1995 Graphical Presentation of Results from Principal Components Analysis, in *Computer Applications and Quantitative Methods in Archaeology 1994* (eds J Huggett and N Ryan), British Archaeological Reports, International Series 600, 63-67

Baxter, M J, and Beardah, C C, 1996 Beyond the Histogram: Improved Approaches to Simple Data Display in Archaeology Using Kernel Density Estimates, *Archeologia e Calcolatori,* 7, 397-408

Baxter, M J, Beardah, C C, and Wright, R V S, 1997 Some Archaeological Applications of Kernel Density Estimates, *Journal of Archaeological Science,* 24, 347-354

Binford, L R, 1978 Dimensional Analysis of Behaviour and Site Structure: Learning from an Eskimo Hunting Stand, *American Antiquity*, 34, 330-361

Bowman, A, and Foster, P, 1993 Density Based Exploration of Bivariate Data, *Statistics and Computing*, 3, 171-7

Jones, M C, Marron, J S and Sheather, S J, 1996 Progress in Data-Based Bandwidth Selection for Kernel Density Estimation, *Computational Statistics*, 11, 337-381

Lukesh, S S, and Howe, S, 1978 Protoapennine vs. Subapennine: Mathematical Distinction Between Two Ceramic Phases, *Journal of Field Archaeology*, 5, 339-47

Sain, S R, Baggerly, K A and Scott, D W, 1994 Cross-validation of Multivariate Densities, *Journal of the American Statistical Association*, 89, 807-817

Sheather, S J, and Jones, M C, 1991 A Reliable Data-based Bandwidth Selection Method for Kernel Density Estimation, *Journal of the Royal Statistical Society*, Series B, 53, 683-690

Silverman, B, 1986 *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London

Stos-Gale, Z A, Gale, N H, and Annets, N, 1996 Lead Isotope Data from the Isotrace Laboratory, Oxford: Archaeometry data base 3, ores from the Aegean, part 1, *Archaeometry,* 38, 381-390

Wand, M P, 1992 Error analysis for General Multivariate Kernel Estimators, *Journal of Nonparametric Statistics*, 2, 1-15

Wand, M P, and Jones, M C, 1993 Comparison of Smoothing Parameterizations in Bivariate Kernel Density Estimation, *Journal of the American Statistical Association*, 88, 520-528

Wand, M P, and Jones, M C, 1994 Multivariate Plug-in Bandwidth Selection, *Computational Statistics*, 9, 97-116

Wand, M P, and Jones, M C, 1995 *Kernel Smoothing*, Chapman and Hall, London

Contact details
Christian C. Beardah
Dept of Mathematics, Statistics and Operational Research
The Nottingham Trent University
Nottingham NG11 8NS
UK
e-mail: c.beardah@maths.ntu.ac.uk