Biometrika (2002), **89,** 1, *pp.* 145–158

Printed in Great Britain

Estimating and depicting the structure of a distribution of random functions

BY PETER HALL

Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia

peter.hall@anu.edu.au

AND NANCY E. HECKMAN

Department of Statistics, University of British Columbia, Vancouver BC V6T 1Z2, Canada

heckman@stat.ubc.ca

SUMMARY

We suggest a nonparametric approach to making inference about the structure of distributions in a potentially infinite-dimensional space, for example a function space, and displaying information about that structure. It is suggested that the simplest way of presenting the structure is through modes and density ascent lines, the latter being the projections into the sample space of the curves of steepest ascent up the surface of a functional-data density. Modes are always points in the sample space, and ascent lines are always one-parameter structures, even when the sample space is determined by an infinite number of parameters. They are therefore relatively easily depicted. Our methodology is based on a functional form of an iterative data-sharpening algorithm.

Some key words: Bandwidth; Cluster analysis; Functional data analysis; Gaussian process; Generalised Fourier expansion; Karhunen–Loève expansion; Kernel methods; Line of steepest ascent; Mode; Nonparametric density estimation; Tree diagram.

1. INTRODUCTION

In conventional statistical problems, involving either one- or two-dimensional data, some of the relationships among data values can be assessed visually. They often provide important information about the processes that might have generated the data. For example, the number of clusters is a guide to the number of components in the sampled population, and the distance of a data value from the centre of its cluster may indicate the strength of its connection to one component rather than another.

For functional data, however, these relationships are not nearly so obvious from the data, and standard structure-displaying methods often do not overcome this difficulty. In particular, conventional perspective mesh plots are generally not helpful when the data are more than two-dimensional. In this paper we argue that, for functional data, the simplest way of depicting structure is through modes and density ascent lines, the latter being the projections into the sample space of the curves of steepest ascent up the surface of a functional-data density. The key to our methodology is a functional form of an iterative data-sharpening algorithm, introduced in a finite-dimensional setting by Choi &

Hall (1999a). In that approach, data are transformed so that they move towards the mode of a nonparametric density estimator.

One reason for focusing on modes and density-ascent lines is that modes are always points in the sample space, and ascent lines are always one-parameter structures, even when the sample space is determined by an infinite number of parameters. This makes the depiction of distributional features relatively straightforward. By way of comparison, contours and ridge lines of a functional-data density have co-dimension one and so are generally infinite-parameter structures. This makes them as difficult to depict as the density itself. Also, the modes of a functional-data density are directly related to cluster analysis, and in fact our techniques lead to simple and particularly effective methods for classification of functional data. Densities in infinite-dimensional spaces, and density contours and ridge lines there, are discussed in § 4.

In § 2 we suggest a general method for constructing a functional-data density estimator, based on a measure of distance in the function space. There is a very wide choice. For example, if one wished to identify directions in which the distribution was relatively highly diffuse then one could use the inverse of a measure of covariance to describe distance. However, to reduce the reader's workload we shall employ only existing distance measures. For example, when the data are curves we shall use L_2 distance, or its square, between two data or between their derivatives. The latter approach can reveal important information about differences in shape as opposed to differences in location. In this context a rank correlation measure of distance, proposed by Heckman & Zamar (2000), is also useful.

Our modified form of the data-sharpening algorithm reduces the step length of the motion to zero, at least conceptually; in implementation, the step length is very small but positive. The modification also involves shifting the 'points', which are now functions, all the way to the mode, and measuring the distance, in the function space, through which they are moved. This technique enables us to work throughout, both conceptually and statistically, with the original data. In particular, it is not necessary to select a basis and a dimension in order to convert the original, functional-data problem into a more conventional one of multivariate analysis.

Since the density of functional data cannot be viewed directly, the mathematical accuracy of statistical approximations to the density is not as important as it would be in a low-dimensional problem. As a result, choice of the bandwidth cannot be meaningfully motivated by traditional criteria, such as minimisation of mean squared error. We suggest instead that it be based on criteria that are relatively easily interpreted, such as the fineness of data classification that a given bandwidth produces; as bandwidth is decreased the number of clusters into which the distribution is divided by its modes is increased in an approximately monotone way. We use this as the basis for describing the effects of different levels of smoothing on properties of density estimators.

If the distance measure can be expressed through an inner product then it may be used as the basis for a tree diagram, depicting directed relationships among data values. These techniques are related to those based on minimal spanning trees for spatial data (Florek et al., 1951; Friedman & Rafsky, 1981, 1983), and to density-based versions of those techniques suggested by Cheng et al. (2002). Our methods are illustrated by application to real data in § 3.

There is a significant literature on direct comparison of data curves, as distinct from comparison via the structure of their distribution. It includes work of Ramsay et al. (1995) on comparison of growth curves, as well as methods suggested by Kneip & Gasser (1992),

Wang & Gasser (1997) and Ramsay & Li (1998) for aligning data curves. Methods for correlation analysis and principal components analysis for curve data are also related, not least because they can help to find low-dimensional approximations to problems that we treat in the full, infinite-dimensional forms. In this context we mention work of Leurgans et al. (1993), Pezzulli & Silverman (1993) and Silverman (1995). Nonparametric density and mode estimation via dimension reduction were proposed by Gasser et al. (1998). Early nonparametric approaches to functional data analysis include that of Rice & Silverman (1991). These methodologies and many others are discussed at length by Ramsay & Silverman (1997). Techniques for analysing structure, including clustering, for complex datasets have been discussed by, for example, Hartigan (1975), Friedman & Rafsky (1983), Hall et al. (1992), Tierney (1990), Swayne et al. (1991) and Cheng et al. (2002).

Earlier applications of data sharpening methods to the problem of mode or ridge estimation include those of Fwu et al. (1981), S. B. Boswell in an unpublished 1983 Rice University Ph.D. dissertation and Jones & Stewart (1997). In the more general context of density or intensity estimation, and ridge estimation, contributions of Samiuddin & El-Sayyad (1990), Jones & Signorini (1997) and Choi & Hall (1999a, b), should be mentioned.

2. Methodology

2·1. *Modes, ascent lines and clusters*

Suppose a sample $\mathcal{X} = \{X_1, \ldots, X_n\}$ is drawn from a distribution on the sample space $\mathcal{S}$, and let $d(x, y)$ be a measure of the distance between elements x and y of $\mathcal{S}$. We shall often take d to be the square of L_2 distance, in which case $d^{\frac{1}{2}}$ is a metric. In most cases only the interpretation of distance suffers if d is not a metric, even if d is asymmetric. An exception is the case of tree diagrams, where symmetry is relatively important. In the present section we shall take $\mathcal{S}$ to be a vector space. Our numerical examples in § 3, and our theoretical account in § 4, will address the case where $\mathcal{S}$ is a space of functions.

Central to our definitions of modes, ascent lines and clusters will be the assumption that the sampling distribution admits a measure of density with respect to d. Let X denote a generic datum X_i, suppose $h > 0$ and $x \in \mathcal{S}$, and write $p_h(x)$ for either the probability that $d(x, X) \leqslant h$ or a smooth approximation to the probability such as the kernel-based one given in the next paragraph. Then we ask that either

(a) $p_h(x) < p_h(y)$ for all sufficiently small h or
(b) $p_h(x) > p_h(y)$ for all sufficiently small h or
(c) $p_h(x)/p_h(y) \to 1$ as $h \to 0$.

Thus, only relative sizes of $p_h(x)$ are important to us, not the actual sizes, and so our estimators of p_h do not require normalisation. The definition of a density in infinite-dimensional, or more properly infinite-parameter, function spaces will be discussed in § 4·1.

Let K be a nonnegative and nonincreasing function on the positive half-line, decreasing to 0 as $u \to \infty$. We shall generally take $K(u)$ to be either e^{-u} or $e^{-u^2/2}$, although compactly supported kernels may also be employed. Let h be a bandwidth, and define $W_h(x, y) = K\{d(x, y)/h\}$. An empirical version of $p_h(x)$ is given by

$$\hat{p}_h(x) = \sum_{i=1}^{n} W_h(x, X_i), \tag{2·1}$$

up to a positive multiplier that does not depend on x. Indeed, we may take $p_h(x)$ to be

the kernel-smoothed probability $E\{\hat{p}_h(x)\}$. Define the transformation

$$\hat{T}_h(x) = \frac{\sum_i X_i W_h(x, X_i)}{\sum_i W_h(x, X_i)}, \tag{2·2}$$

denoting an average of all data vectors that lie in a neighbourhood of $x \in \mathscr{S}$. It is a Nadaraya–Watson estimator in which the data are taken as both explanatory and response variables. The transformation $\hat{T}_h$ suffers from virtually no numerical difficulty, even in infinite-dimensional problems. It requires no implicit solution of equations, and employs only elementary operations, such as addition and multiplication. If we use a kernel, such as the Gaussian, with infinite support, and if a norm $\|.\|$ is defined on $\mathscr{S}$, then $\|\hat{T}_h(x)\| \leqslant \sup_i \|X_i\|$ for all x.

Let $\tilde{p}_h$ be the version of $\hat{p}_h$ computed using $\tilde{K}$, defined by $\tilde{K}(u) = \int_{v>u} K(v)\, dv$, instead of K. In particular, $\tilde{p}_h \equiv \hat{p}_h$ if $K(u) = e^{-u}$. We shall show in § 4·2 that, when the distance between functions x and y is defined by $d(x, y) = \int (x-y)^2$, $\hat{T}_h$ takes x to a point $\hat{T}_h(x)$ which is in a region of generally higher empirical density, as measured by $\tilde{p}_h$, than the region containing x. Moreover it will be proved that $\hat{v}_h(x) = \hat{T}_h(x) - x$ is a vector in the direction of the projection into $\mathscr{S}$ of a generally upwards-moving trajectory on the density surface $\tilde{\mathscr{T}}$, defined by $s = \tilde{p}_h(x)$.

To construct mode estimators and clusters, take a small positive quantity ε and, for each $X_i \in \mathscr{X}$, trace out the sequence of points $X_{i0}, X_{i1}, \ldots$ defined iteratively by $X_{i0} = X_i$ and $X_{i,j+1} = X_{ij} + \varepsilon \hat{v}_h(X_{ij})$, for $j \geqslant 0$. In the theoretical limit as $\varepsilon \to 0$, this sequence becomes a density ascent line, $\hat{\mathscr{L}}_i$ say, that represents the projection into $\mathscr{S}$ of a trajectory drawn from $(X_i, \hat{p}_h(X_i))$ to a local maximum of a density surface.

It will be shown in § 4·2 that $\hat{\mathscr{L}}_i$ may be interpreted as the projection into $\mathscr{S}$ of a line of steepest ascent up a density surface computed using the kernel $\tilde{K}$. The context of this result is that described two paragraphs above. Therefore, density ascent lines $\hat{\mathscr{L}}_i$ can be explicitly related to steepest ascent curves up density surfaces. Each line starts at a data value and ends at a mode or perhaps at a saddlepoint. Of course, the latter points will not themselves be data values. If we want the modes to be local maxima of the density surface defined by $s = \hat{p}_h(x)$, where $\hat{p}_h$ is given by (2·1), then we should replace K in the definition at (2·2) by $-K'$. This result implies that, if distance is measured by the square root of $\int (x-y)^2$, and if we take K to be proportional to a Normal density, then the kernels used to define the density estimator $\hat{p}_h$ and its associated transformation $\hat{T}_h$ are identical.

In numerical work we found that, as a result of step length not being infinitesimal, it is possible for the algorithm tracing a density ascent line to stop before it reaches its mode, $\hat{x}^0$ say. It can be made to continue, and terminate at a mode, by the expedient of reducing the value of ε slightly. Each mode estimate $\hat{x}^0$ defines a cluster in the dataset, equal to the set of those data X_i whose density ascent lines lead to $\hat{x}^0$. Our example in § 3, the analysis of precipitation curves using L_2 distance, will illustrate the way information about relatively flat parts of functional-data density surfaces can be obtained through density ascent lines.

A particularly attractive feature of density ascent lines is the fact that, no matter how complex the sample space, they are sets determined by a single parameter. For a univariate density, the density ascent line starting at a real number x is an interval with x as one of its endpoints and a mode of the density, or density estimator, as the other. The line has an identical definition in the case of functional data, except that the interval is replaced

by a path in the function space. Any point on the path may be parameterised by the integral of the infinitesimal distances through which the function has moved, up to that point. As a numerical approximation to this quantity, and in a functional-data setting, the length of the path that leads to the function X_{ij} may be taken equal to

$$S_j = \varepsilon \sum_{k=0}^{j-1} \left\{ \int \hat{v}_h(X_{ik})^2 \right\}^{\frac{1}{2}}.$$

The relationship between a data function X_i and the corresponding modal function $\hat{x}^0$, at the opposite end of $\hat{\mathscr{L}}_i$ from X_i, may be illustrated by plotting the continuum of density ascent lines that lead from X_i to $\hat{x}^0$. This depiction can be particularly revealing as a dynamic graphic, indicating both the manner and rate at which different functional-data features change as functions of S_j. We suggest that the time dimension in the dynamic graphic be taken proportional to S_j, so that features change more rapidly as the density ascent line moves to regions of the sample space that correspond to more steeply graded parts of the functional-data density. On paper, the depiction of functional change along a density ascent line can be represented as a discrete sequence of plots at points that are equally spaced in terms of the distance measure S_j.

2·2. *Tree diagrams*

Suppose the distance function d is defined in terms of an inner product $(.,.)$, with $d(x, y) = (x - y, x - y)$. Let $\hat{\mathscr{T}}$ denote the surface represented by the equation $s = \hat{p}_h(x)$. Within a given cluster $\mathscr{C}$, a tree diagram depicting directed relationships among data values may be constructed as follows.

Let λ denote a positive constant. We shall use it to weight a penalty term,

$$d(X_i, X_j) - (X_j - X_i, u_i)^2 \geqslant 0,$$

that is added to the simple distance measure $d(X_i, X_j)$ when defining its penalised form, $D(X_i, X_j)$. Increasing λ will result in greater emphasis being given to linkages that reflect upwards movement on the surface $\hat{\mathscr{T}}$. We shall link $X_i \in \mathscr{C}$ to X_j, in that direction, if, among all $X_j \in \mathscr{C}$ such that $\hat{p}_h(X_j) > \hat{p}_h(X_i)$, X_j minimises

$$D(X_i, X_j) = (1 + \lambda) d(X_i, X_j) - \lambda (X_j - X_i, u_i)^2,$$

where $u_i = \hat{v}_h(X_i)/d(0, \hat{v}_h(X_i))^{\frac{1}{2}}$ is an element of $\mathscr{S}$ with unit length. If there exists no point $X_j \in \mathscr{C}$ such that $\hat{p}_h(X_j) > \hat{p}_h(X_i)$, then no linkage is drawn away from X_i to another point; X_i represents the element of $\mathscr{C}$ at which $\hat{p}_h$ is maximised, and may, without essential loss of generality, be identified with the mode that corresponds to $\mathscr{C}$.

3. Numerical examples

3·1. *Preliminaries*

Throughout our numerical work we use $K(u) = e^{-u^2/2}$. When taking d to be L_2 distance, employing this kernel is of course equivalent to using $K(u) = e^{-u}$ when d is the square of L_2 distance, modulo a change of bandwidth. For the sake of brevity we do not show plots of function sequences represented by density ascent lines, or plot the measure S_j of distance along the steepest-ascent curve. Information in these graphs has been incorporated into our discussion, however. In particular, plots of S_j against j are invaluable for showing where the density surface is steep and where there are 'shoulders' in the surface.

We apply our methods to a dataset of 35 curves representing the natural logarithm of the daily precipitation at 35 Canadian weather stations averaged over the years 1961 to 1994. Day 1 is 1 January. The data are available from J. O. Ramsay's website, `www.psych.mcgill.ca`. The logarithms of the averages are particularly noisy, and so we smooth the logarithms using the algorithm `smooth.spline` in S-Plus, choosing the smoothing parameter by generalised crossvalidation. This produces the smooth curves in Fig. 1. To ensure periodicity, each dataset is replicated three times, in the obvious manner, prior to smoothing.

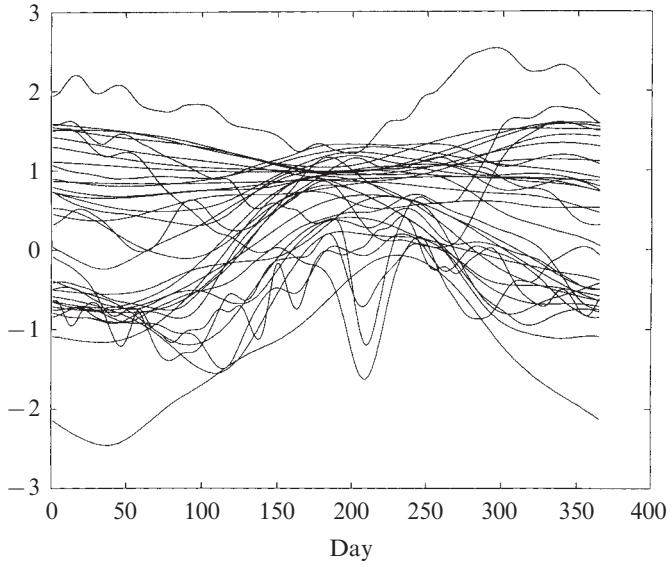

Fig. 1. Smoothed log precipitation curves of 35 Canadian weather stations averaged over years 1961–1994; day 1 is 1 January.

3·2. *Density surface defined by L_2 distance on functions*

Figure 2 shows the modal functions resulting from four L_2 analyses using bandwidths $h = 14$, 10, 8 and 6, equivalent to the 39th, 26th, 19th and 12th percentiles of the interfunction distances. As outlined in § 2, the modal functions are obtained iteratively as limits of sequences $X_{i,j+1} = X_{ij} + \varepsilon \hat{v}_h(X_{ij})$, for $j \geqslant 0$, where $X_{i0} = X_i$ and $\varepsilon > 0$ is chosen small. Here we have taken $\varepsilon = 0{\cdot}1$.

In kernel estimation of a univariate density, Silverman (1981) showed that the number of modes is a decreasing function of the bandwidth, provided one uses a normal kernel. While properties of smoothing with a normal kernel in the multivariate case have been studied (Kostrowicki & Piela, 1991; Moré & Wu, 1997), it is unclear whether or not Silverman's result generalises. Silverman's (1981) derivation is based on Schoenberg's (1950) result about variation-diminishing transformations, and its multivariate version does not hold for integral transforms obtained by convolution with the Gaussian kernel. For example, in the bivariate case there exist smooth functions that have only a finite number of turning points, but are such that their convolution with a Gaussian density has a continuum of turning points. Nevertheless, in all the examples we considered we found that reducing the bandwidth never reduced the number of modes.

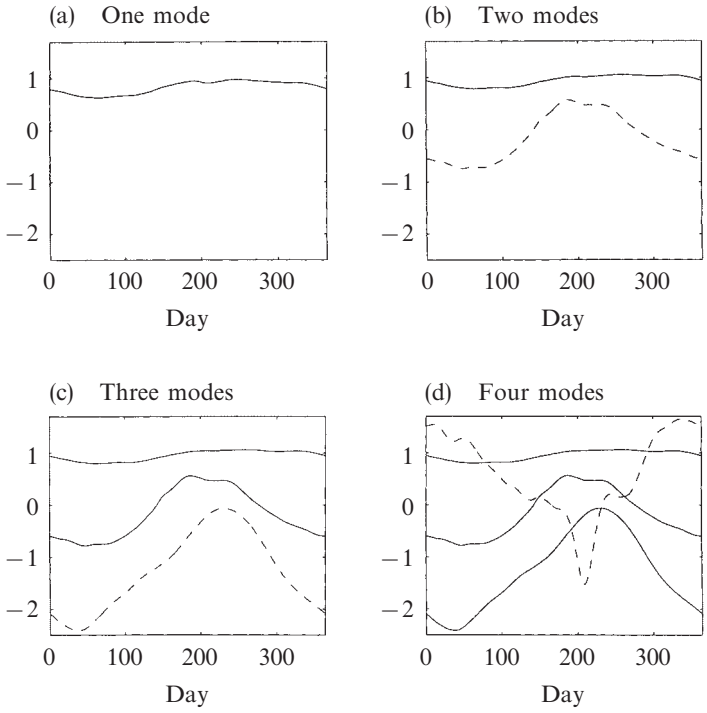

Fig. 2. Weather station data: Modal curves for four L_2 analyses using different bandwidths. In each panel the new modal function is indicated by a dotted line. The other modal functions change slightly from panel to panel, because different bandwidths are used, but the changes are slight. Therefore, the first, ..., fourth modal functions are easily recognised; the first k of these are represented in the kth panel.

In particular, reducing h beyond the values used to prepare Fig. 2 results in a larger number of modes, up to 35 for sufficiently small bandwidths. The value of ε seems only to affect convergence of the algorithm. If ε is chosen to be large, the algorithm either converges very slowly or not at all. However, upon convergence the modal curves are basically unchanged by the choice of ε.

The modal curves change only slightly as the bandwidth decreases. This property was observed in simulated data too, and is a particularly helpful feature of our method when it is used for classification: the modes that represent clusters change only very slowly as the number of clusters is altered. This means that 'old' clusters are easily recognised even if some of their members have deserted them for other modes.

The precipitation curves whose density ascent lines converge to a given mode may be regarded as being in the same cluster, and are depicted in Fig. 3 for the case of four modes. The first cluster contains 19 curves, the second cluster contains 14 curves, the third cluster consists only of the Resolute weather station and the fourth cluster contains only the Victoria weather station. In the two- and three-mode analyses, Victoria was part of the first cluster. Resolute was part of the second cluster in the two-mode analysis, but split off to form its own cluster in the three-mode analysis. Interestingly, the cluster groups are very stable, not changing with h. For instance, values of h from 9 to 13 yield two modes, with the clusters remaining the same throughout.

The curves within each of the first and second cluster in Fig. 3 do not all have the same

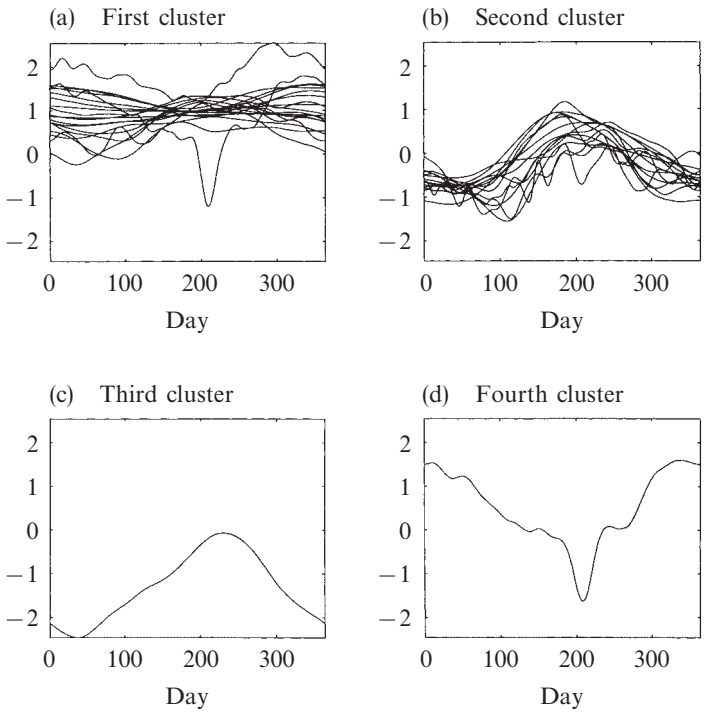

Fig. 3. Weather station data: Four mode case. Shown in the kth panel are the data curves whose density ascent curves converge to the kth modal function, for the four modal functions in Fig. 2(d).

shapes. The curves correspond to similar annual precipitation levels, however, and this seems to be the effective discriminator that is operating there. Indeed, the ranges of values of logarithms of annual averages of daily precipitations for the stations in the first and second clusters of Fig. 3, modulo the smoothing that is used to define the curves, are 0·42 to 1·86 and $-0{\cdot}52$ to 0·03 log mm, respectively. The average for Resolute, third cluster, is $-1{\cdot}31$ and the average for Victoria, fourth cluster, is 0·49.

Thus our analysis identifies the two stations, Victoria and Resolute, as possible outliers relative to the other data. Resolute is easily distinguishable in Fig. 1, being the lowest curve, with average of the smoothed logarithm of precipitation much lower than the other stations. Victoria is also easily seen as an outlier in the first cluster: its maximum is larger and its minimum smaller, and, in common with only one other curve, that for Vancouver, it has a marked 'V' shape. The average of Victoria's daily logarithm of precipitation is lower than that of Vancouver's, being 0·49 compared to 0·91. It is possible, however, that four clusters are too many for this dataset, and that Victoria should be included in the same cluster as Vancouver, which it most resembles in terms of shape of precipitation curve.

3·3. *Density surface defined by L_2 distance on derivatives or by rank distance*

For the weather station data, L_2 distance on function derivatives performs well at identifying curves that have unusually steep gradients. It ignores differences in average annual rainfall.

As in the case of L_2 distance applied directly to the functions, and also for the rank-based distance discussed below, modal functions change very little as bandwidth is reduced.

The first modal function is relatively flat, but with precipitation in summer slightly elevated. The second modal function shows low precipitation in late summer/early autumn and high precipitation in the winter. The cluster corresponding to this mode contains only two stations, Vancouver and Victoria, which are distinguished by their V-shaped precipitation curves. The third modal function is opposite to the second, showing low levels of precipitation in the winter and high levels in the summer. The fourth modal function is that for Resolute, which has highest precipitation in the autumn.

The rank distance between two curves is defined to be $\sqrt{(1-\rho)}$, where ρ denotes the rank correlation between the two functions; see Heckman & Zamar (2000). For bandwidths that produce up to four modes by this measure, the first, second and third modal curves are similar to their counterparts for L_2 distance on the derivatives. However, the rank-based analysis focuses more on the qualitative similarity of shapes of the curves. For instance, in the two-mode rank-based analysis, the cluster corresponding to the second mode includes more than just Vancouver and Victoria: the second cluster contains all nine stations with high winter precipitation and low summer precipitation. Decreasing h to the three-mode case results in five stations moving from the first group to form a cluster. These five stations also form the third cluster in the derivative based analysis, with low precipitation in the winter and high in the summer. Decreasing h to the four-mode case results in a singleton for the fourth mode, Kamloops, whose curve is irregularly shaped.

In the three- and four-mode analyses, convergence was very slow, indicating that the surface is somewhat flat near the third and fourth modes. In fact, we had to reduce the value of ε from 0·1 to 0·05 to guarantee convergence of our algorithm. Using such a small value of ε for all iterations and all weather stations slows the algorithm. However, to speed convergence, we can decrease the value of ε for some stations and also decrease ε after sufficiently many iterations.

3·4. *Paths of the ascent lines*

As noted in § 2·1, a plot of S_j versus j gives information about how the sequence X_{ij}, for $j>0$, moves to a mode and therefore provides information about the surface represented by $s=\hat{p}_h(x)$. In the analysis of the precipitation data these plots are usually convex, with the associated sequence of iterated curves moving with fairly constant deceleration until the mode is reached. This indicates that, along this path, the density surface is quite smooth, with no pronounced shoulder. However, sometimes the plots have flat spots, indicating that the sequence is resting at a shoulder of the density surface. We see this in two situations, one where the initial part of the sequence does not change much and another when the initial change is rapid followed by little change and then followed by fast convergence to the mode. The former situation applies to unusual curves, ones that are, in some sense, outliers, not at first being attracted to a mode. The latter situation applies to groups of curves, typically those that would form a cluster if one used a smaller value of h. In this case, the curve that is situated at a resting point of the trajectory is usually either similar to a modal curve from an analysis with smaller bandwidth or lies between two modal curves. For instance, in the one-mode analysis using L_2 distance, the trajectories of the data curves in the second, third and fourth clusters of Fig. 3 rest momentarily at a curve that is similar to, but flatter than, the second modal function. Thus we infer that there is a shoulder in the unimodal density, situated near the second modal function.

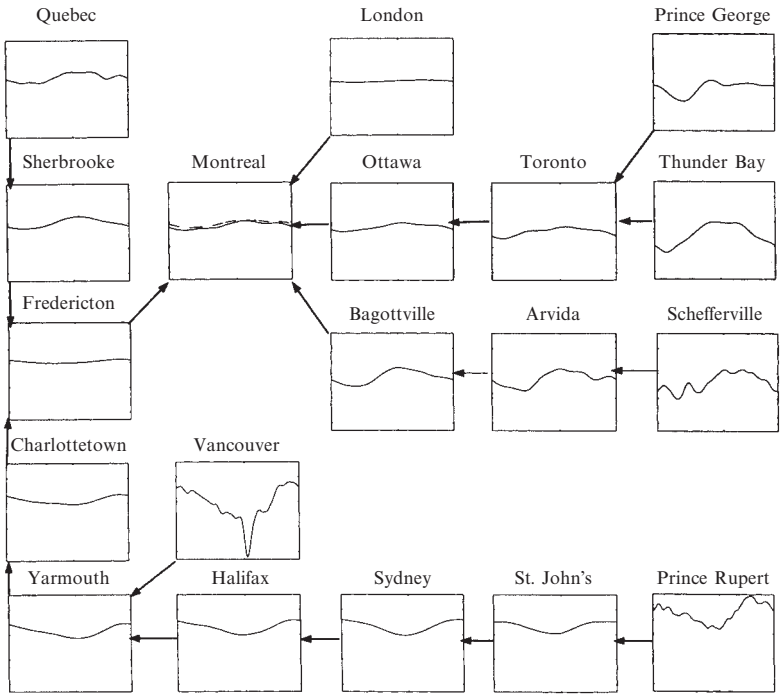

Fig. 4. Weather station data: Tree diagram for the 19 Canadian weather stations in first cluster of Fig. 3.

3·5. *Tree diagrams*

In Fig. 4 we give an example of a tree diagram for the 19 weather stations in the first cluster of Fig. 3. The diagram is constructed using $\lambda = 0{\cdot}3$. The value of $\hat{p}_h(X_i)$ is maximised for the Montreal weather station, and so Montreal forms the endpoint of the tree. The second modal curve is shown as a dashed line in the plot for Montreal. It can be seen that the data curves become similar to the Montreal curve as one progresses through the tree. Note, for instance, the sequence Schefferville, Arvida, Bagottville and Montreal.

3·6. *Simulation study*

Here we summarise a simulation study showing that the real structure of a distribution of random curves can be reliably recovered using our algorithm. For this purpose we shall view our method purely as a classification rule, and compare it with k-means clustering (Hartigan, 1975), as implemented in S-Plus. It will be seen that k-means clustering performs a little better at classification, but the techniques produce similar results and of course our approach provides more explicit information about the shape of the density surface.

Each of 300 datasets of size $n = 50$ was generated from the model

$$X(t) = A\cos(2\pi t) + (B + \mu)\sin(2\pi t),$$

with A and B being independent standard Normal random variables. For 25 of the 50 curves we took $\mu = 2$, and for the other 25 curves, $\mu = -2$. Each curve was observed at 100 equally-spaced points $t \in [0, 1]$. When applying k-means clustering we set the number of clusters equal to two. When implementing our algorithm we used Euclidean distance and, for each dataset, determined the bandwidth h that yielded two clusters. In 93% of

datasets the same h produced the two desired clusters; the remaining datasets required individual analysis. In almost all cases our algorithm converged when $\varepsilon = 0{\cdot}15$, but in a few instances it was necessary to decrease ε to 0·10 or 0·05.

For each algorithm the modal number of misclassifications was one, with 35% of samples having just one misclassification when using k-means clustering, and 33% when using our approach. The median number of misclassifications was also one; the proportion of cases where the number of misclassifications was no more than one was 60% in the case of k-means clustering and 52% in the case of our own approach. The 75th percentiles were two in each case, with the percentages of two or fewer misclassifications being 75% and 83% in the cases of k-means clustering and our technique, respectively. More broadly, the distribution of error rate was skewed slightly to the right for our method, relatively to that for k-means clustering.

4. Theoretical properties

4·1. *Density of X*

In this section we suggest a definition of a density p of the distribution of a Gaussian process X. It is shown that the simpler kernel estimator $\hat{p}_h$, suitably normalised, converges uniformly to p. Results for a mixture of Gaussian processes are given in a longer version of this paper, obtainable from the authors.

Suppose X is a random function from a space $\mathscr{R}$ to the real line, and has zero mean. We may write the Karhunen–Loève expansion of X as

$$X = \sum_{i=1}^{\infty} Y_i \alpha_i \psi_i,$$

where the random variables $Y_1, Y_2, \ldots$ are independent and have a common standard Normal distribution, each $\alpha_i \geqslant 0$, $\sum_i \alpha_i^2 < \infty$ and $\psi_1, \psi_2, \ldots$ is an orthonormal sequence of functions from $\mathscr{R}$ to the real line. Put $\alpha = (\alpha_1, \alpha_2, \ldots)$, let $y = (y_1, y_2, \ldots)$ be a sequence of real numbers, and define $\langle y \rangle = \sum_i y_i^2$. Let $\mathscr{S}_\alpha$ denote the class of functions x on $\mathscr{R}$ that can be expressed as

$$x = \sum_{i=1}^{\infty} y_i \alpha_i \psi_i, \tag{4·1}$$

with $\sum_i y_i^2 \alpha_i^2 > \infty$. The sample space $\mathscr{S}$, introduced in § 2·1, will here be $\mathscr{S}_\alpha$. Given $x \in \mathscr{S}_\alpha$ define

$$\lfloor x \rfloor_\alpha = \sum_{i=1}^{\infty} \left\{ \int_{\mathscr{R}} x(u) \psi_i(u)\, du \right\}^2 \alpha_i^{-2},$$

where we interpret 0/0 as 0. If x is expressed in terms of y by (4·1) then $\lfloor x \rfloor_\alpha = \langle y \rangle$.

Assume that $K(u) = \exp(-u)$ for $u > 0$, and define

$$p(x) = \exp(-\tfrac{1}{2} \lfloor x \rfloor_\alpha), \quad \rho_h(x) = E\{\hat{p}_h(x)\} / E\{\hat{p}_h(\bar{0})\},$$

where $\bar{0}$ denotes the zero function. The denominator here is merely a normalising constant, and for us its only important property is that it does not depend on x. We may think of ρ_h as the expected value of the density estimator $\hat{p}_h$, divided by the constant of proportionality $E\{\hat{p}_h(\bar{0})\}$.

The function p may be be interpreted as a density of the distribution of X. Note that

$p(x) = 0$ when $x \notin \mathscr{S}_\alpha$. It may be shown that $\rho_h(x)$ converges to $p(x)$, uniformly in x, as $h \to 0$. A proof involves expressing $E\{\hat{p}_h(x)\}$ in terms of the Laplace transform of the random variable

$$d(x_y, X) = \sum_{i=1}^{\infty} (y_i - Y_i)^2 \alpha_i^2,$$

where x_y denotes the value of the right-hand side of (4·1) for a given value of y.

The rate of convergence of $\hat{p}_h / E\{\hat{p}_h(\bar{0})\}$ to ρ may also be explored; it is of course slower than $n^{-\varepsilon}$ for any $\varepsilon > 0$, since the stochastic process X is infinite-dimensional. Details are obtainable from the authors.

4·2. *Properties of $\hat{T}_h$ and $\hat{x}^0$*

We first relate $\hat{v}_h(x) = \hat{T}_h(x) - x$ to the direction of steepest ascent up a density surface. Define distance by $d(x, y) = \int (x - y)^2$, and put $(x, y) = \int xy$. Let $\tilde{p}_h(x)$ be as at (2·1) but with W_h there replaced by the version of this function using $\tilde{K}(u) = \int_{v > u} K(v)\, dv$ instead of K. Note that, for $K(u) \equiv e^{-u}$, $\tilde{K} \equiv K$ and so $\tilde{p}_h \equiv \hat{p}_h$. The formula below shows that $\hat{v}_h(x)$ is in the direction of steepest ascent up the surface $\tilde{\mathscr{T}}$ defined by $s = \tilde{p}_h(x)$: for $g \in \mathscr{S}$,

$$\tilde{p}_h'(x)(g) \equiv \frac{\partial}{\partial \delta} \tilde{p}_h(x + \delta g)\bigg|_{\delta = 0} = 2h^{-1} \hat{p}_h(x)(\hat{v}_h(x), g).$$

This confirms the claims made in § 2·1 that $\hat{T}_h$ takes x to a point with higher empirical density, and that a density ascent line $\hat{\mathscr{L}}_i$ is the projection into $\mathscr{S}$ of a line of steepest ascent up $\tilde{\mathscr{T}}$. Excepting possible pathological cases where the line keeps moving uphill but never actually converges, it will converge to a point $\hat{x}^0$ that is either a mode or a saddlepoint of the surface defined by $s = \hat{p}_h(x)$.

In the finite, k-dimensional case, assume that the true density p has two uniformly continuous derivatives of all types, the $k \times k$ matrix of second derivatives is nonsingular at each mode, p has three derivatives in the neighbourhood of each mode and no saddlepoint and $n^{-\{2/(k+4)\}+\delta} \leqslant h \leqslant n^{-\delta}$ for some $\delta > 0$ and all sufficiently large n. Then it may be shown that $\hat{x}^0 - x^0 = O_p(h + n^{-1/2} h^{-(k+2)/4})$. It follows from the latter property that $\hat{x}^0 - x^0 = O_p(n^{-2/(k+6)})$ if h is asymptotic to a constant multiple of $n^{-2/(k+6)}$. A central limit theorem for $\hat{x}^0 - x^0$ may also be proved.

4·3. *Definition of density contours and ridge lines*

In this section we verify the claim made in § 1 that contours and ridge lines of densities of functional data are too high-dimensional to be useful for depicting structure. The contour corresponding to a density p is a level set of p, and so is defined by $\mathscr{T}_s = \{x : p(x) = s\}$ where $0 < s \leqslant \sup p$. To define a ridge line, let $\tilde{p}_h'(x)(.) : \mathscr{S} \to \mathscr{R}$ be the gradient functional defined in § 4·2, and let $\| \tilde{p}_h'(x) \|$ denote its norm. Then $x = x^1 \in \mathscr{T}_s$ is on a ridge line if and only if $\| \tilde{p}_h'(x) \|$ achieves a local minimum there among functions in $\mathscr{T}_s$. Under regularity conditions, for each s there will only be a finite number of such x^1's, and these functions will change smoothly with s. The set of ridge lines is the union over s of the set of such functions x^1.

In two dimensions, for densities of random two-vectors, ridge lines are generally curves in the plane. They represent the projection into the sample space of ridges on the density surface, where the ridge can now be interpreted in the geographic sense of that term. More generally, it follows from the previous paragraph that a contour and a ridge line are both

of co-dimension one, i.e. their structures require representations involving only one fewer parameter than the entire space. Thus, it is too complex to depict them when the space is infinite-dimensional.

References

Cheng, M.-Y., Hall, P. & Hartigan, J. A. (2002). Estimating gradient trees. *J. Comp. Graph. Statist.* To appear.

Choi, E. & Hall, P. (1999a). Data sharpening as a prelude to density estimation. *Biometrika* **86**, 941–7.

Choi, E. & Hall, P. (1999b). Nonparametric analysis of earthquake point-process data. In *State of the Art in Mathematical Statistics and Probability Theory*, Festschrift for W. R. van Zwet, Ed. M. de Gunst, C. A. J. Klaassen and A. W. van der Vaart, pp. 324–44. Beachwood, OH: Institute of Mathematical Statistics.

Florek, K., Lukaszwicz, J., Perkal, J., Steinhaus, H. & Zubrzycki, S. (1951). Sur la liaison et la division des points d'un ensemble finit. *Colloq. Math.* **2**, 282–5.

Friedman, J. H. & Rafsky, L. C. (1981). Graphics for the multivariate two-sample problem (with Discussion). *J. Am. Statist. Assoc.* **76**, 277–95.

Friedman, J. H. & Rafsky, L. C. (1983). Graph-theoretic measures of multivariate association and prediction. *Ann. Statist.* **11**, 377–91.

Fwu, C., Tapia, R. A. & Thompson, J. R. (1981). The nonparametric estimation of probability densities in ballistics research. In *Proc. Twenty-Sixth Conf. Design of Experiments in Army Research Development and Testing*, Ed. U.S. Army Mathematics Steering Committee, pp. 309–26. Research Triangle Park, NC: U.S. Army Research Office.

Gasser, T., Hall, P. & Presnell, B. (1998). Nonparametric estimation of the mode of a distribution of random curves. *J. R. Statist. Soc.* B **60**, 681–91.

Hall, P., Qian, W. & Titterington, D. M. (1992). Ridge finding from noisy data. *J. Comp. Graph. Statist.* **1**, 197–211.

Hartigan, J. A. (1975). *Clustering Algorithms*. New York: Wiley.

Heckman, N. E. & Zamar, R. H. (2000). Comparing the shapes of regression functions. *Biometrika* **87**, 135–44.

Jones, M. C. & Signorini, D. F. (1997). A comparison of higher-order bias kernel density estimators. *J. Am. Statist. Assoc.* **92**, 1063–73.

Jones, R. H. & Stewart, R. C. (1997). A method for determining significant structures in a cloud of earthquakes. *J. Geophys. Res.* **102**, 8245–54.

Kneip, A. & Gasser, Th. (1992). Statistical tools to analyze data representing a sample of curves. *Ann. Statist.* **20**, 1266–305.

Kostrowicki, J. & Piela, L. (1991). Diffusion equation method of global minimization: performance for standard functions. *J. Optimiz. Theory Appl.* **69**, 269–84.

Leurgans, S. E., Moyeed, R. A. & Silverman, B. W. (1993). Canonical correlation analysis when the data are curves. *J. R. Statist. Soc.* B **55**, 725–40.

Moré, J. J. & Wu, Z. (1997). Global continuation for distance geometry problems. *SIAM J. Optimiz.* **7**, 814–36.

Pezzulli, S. D. & Silverman, B. W. (1993). Some properties of smoothed principal components analysis for functional data. *Comp. Statist.* **8**, 1–16.

Ramsay, J. O. & Li, X. (1998). Curve registration. *J. R. Statist. Soc.* B **60**, 351–63.

Ramsay, J. O. & Silverman, B. W. (1997). *Functional Data Analysis*. New York: Springer.

Ramsay, J. O., Bock, R. D. & Gasser, T. (1995). Comparison of height acceleration curves in the Fels, Zurich, and Berkeley growth data. *Ann. Hum. Biol.* **22**, 413–26.

Rice, J. A. & Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Statist. Soc.* B **53**, 233–43.

Samiuddin, M. & El-Sayyad, G. M. (1990). On nonparametric kernel density estimates. *Biometrika* **77**, 865–74.

Schoenberg, I. J. (1950). On Pólya frequency functions. II: Variation-diminishing integral operators of the convolution type. *Acta Scientarum Mathematicarum Szeged* **12**, 97–106.

Silverman, B. W. (1981). Using kernel density estimates to investigate multi-modality. *J. R. Statist. Soc.* B **43**, 97–9.

Silverman, B. W. (1995). Incorporating parametric effects into functional principal components analysis. *J. R. Statist. Soc.* B **57**, 673–89.

SWAYNE, D. F., COOK, D. & BUJA, A. (1991). XGobi: Interactive dynamic graphics in the X window system with a link to S. In *Proc. Statist. Graph. Sect., Am. Statist. Assoc.*, pp. 1–8. Alexandria, VA: American Statistical Association.

TIERNEY, L. (1990). *LISP-STAT, An Object-Oriented Environment for Statistics and Dynamic Graphics.* New York: Wiley.

WANG, K. & GASSER, T. (1997). Alignment of curves by dynamic time warping. *Ann. Statist.* **25**, 1251–76.

[*Received July* 2000. *Revised April* 2001]