ESTIMATING AND DEPICTING THE STRUCTURE OF A DISTRIBUTION OF RANDOM FUNCTIONS

Peter Hall[1] and Nancy E. Heckman[1,2]

July 26, 2000

ABSTRACT. 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. Our methodology is based on nonparametric density estimation, and draws inference about the slope of the density. The latter step is implemented in a purely iterative way, using only elementary operations of addition and multiplication, and does not require any differentiation or dimension-reduction. Nevertheless it leads in a very simple and reliable manner to "curves" of steepest ascent up the "surface" defined by an estimate of the density of a potentially infinite-dimensional distribution. The projections of these curves into the sample space are always one-dimensional, or more properly one-parameter, structures, and so can be displayed visually even when the sample space is a class of functions. Also, the modes to which the sample space projections lead are themselves points in the sample space, and so can be easily depicted. By way of contrast, the contour lines and ridges of the density of a random function are generally infinite-parameter structures, and so are very difficult to display. As a result, traditional approaches to representing structure are not beneficial in infinite-parameter problems. Likewise, since the density and its estimators usually cannot be illustrated in a meaningful way then traditional methods for bandwidth choice, based on optimising a mathematical measure of performance, do not have much relevance. We suggest instead a technique based on cluster analysis in function space.

KEYWORDS. 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.

SHORT TITLE. Distribution of functions.

[1] Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia.

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

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. One reason is that modes are always points in the sample space, and density-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 1 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.

We suggest a general method for constructing a functional-data density estimator, based on a measure of distance in the function space. Distance can be defined in a variety of ways, depending on which features of functions are of greatest importance. For example, if the data are curves then the L_2 distance between the curves' derivatives, rather than between the functions themselves, may 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 and Zamar (2000), is also useful.

Our approach 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, then 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 easy to interpret, such as the fineness of data classification that a given bandwidth produces. Specifically, 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.

The key to our methodology is a functional form of an iterative data-sharpening algorithm, introduced by Choi and Hall (1999). In that approach, multivariate or infinite-dimensional data are transformed in such a way that they move a specified distance towards the mode of a nonparametric density estimator. Our modification reduces the step length of this 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.

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, Block and Gasser (1995) on comparison of growth curves, as well as methods suggested by Kneip and Gasser (1992), Wang and Gasser (1997) and Ramsay and Li (1998) for aligning data curves. Methods for correlation- 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, Moyeed and Silverman (1993), Pezzulli and Silverman (1993) and Silverman (1995). Nonparametric density and mode estimation via dimension reduction were proposed by Gasser, Hall and Presnell (1998). Early nonparametric approaches to functional data analysis include that of Rice and Silverman (1991). These methodologies and many others are discussed at length by Ramsay and Silverman (1997). Techniques for analysing structure, including clustering, for complex data sets have been discussed by, for example, Hartigan (1975), Friedman and Rafsky (1983), Hall, Qian and Titterington (1992), Tierney (1990), Swayne, Cook and Buja (1991) and Cheng, Hall and Hartigan (2000).

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 probability space $\mathcal{S}$. Let $d(x, y)$ be a measure of the distance between elements x and y of $\mathcal{S}$. In the present section we shall take $\mathcal{S}$ to be a vector space that is closed under multiplication of its elements by real numbers; thus, $\mathcal{S}$ will be a module, in the mathematical sense of that term. The majority of our numerical examples in section 3, and our theoretical account in section 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. In particular, if X denotes a generic datum X_i, if $h > 0$ and $x \in \mathcal{S}$, and if $p_h(x)$ equals the probability that $d(x, X) \leq h$, then we ask that the order relation $p_h(x) < p_h(y)$, for functions x and y in the sample space, be preserved for all sufficiently small h. It is only the order relation that is important to us, not the actual size of $p_h(x)$, and so our estimators of p_h do not require normalisation. Definitions of density in infinite-dimensional, or more properly infinite-parameter, function spaces will be discussed in sections 4.1–2.

Let K be a nonnegative and nonincreasing function on the positive half-line, let h be a bandwidth, and define $K_h(x, y) = K\{d(x, y)/h\}$. An approximation to $p_h(x)$ is given by

$$\hat{p}_h(x) = \sum_{i=1}^{n} K_h(x, X_i)\,, \tag{2.1}$$

up to a positive multiplier that does not depend on x. Define the transformation

$$\widehat{T}(x) = \frac{\sum_i X_i\, K_h(x, X_i)}{\sum_i K_h(x, X_i)}\,, \tag{2.2}$$

denoting an average of all data vectors that lie in a neighbourhood of $x \in \mathcal{S}$. It is a Nadaraya-Watson estimator in which the data are taken as both explanatory and response variables. The transformation $\widehat{T}$ suffers virtually no numerical difficulties, 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 $\|\cdot\|$ is defined on $\mathcal{S}$, then $\|\widehat{T}(x)\| \leq \sup_i \|X_i\|$ for all x.

We shall show in section 4.4 that $\widehat{T}$ takes x to a point $\widehat{T}(x)$ which is in a region of generally higher empirical density, as measured by an estimator $\tilde{p}_h$ similar to $\hat{p}_h$, than the region containing x. Therefore, $\hat{v}(x) = \widehat{T}(x) - x$ is a vector in the direction of the projection into $\mathcal{S}$ of a generally upwards-moving trajectory on the density surface $\widehat{\mathcal{T}}$, defined by $s = \tilde{p}_h(x)$. It will be shown in section 4.4 that $\hat{v}(x)$ is in fact a tangent to a curve of steepest ascent on the density surface.

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

In numerical work we found that it is possible for this algorithm to stop at a point which, to the numerical accuracy of the computations, appears to be a saddlepoint. However the algorithm can be made to continue to a mode by the expedient of reducing the bandwidth slightly, and so we may assume without loss of generality that $\hat{x}^0$ is a mode. Our example in section 3, particularly the analysis of precipitation curves corresponding to weather stations at Vancouver and Victoria, will illustrate the way information about relatively flat parts of functional-data density surfaces can be obtained through density ascent lines via small perturbations of the bandwidth.

Properties of our algorithm will be addressed in section 4.4. We shall show there that, when $\mathcal{S}$ is a space of functions and the distance between x and y is defined by $d(x,y) = \int (x-y)^2$, $\widehat{\mathcal{L}}_i$ may be interpreted as the projection into $\mathcal{S}$ of a line of steepest ascent up a density surface that is computed in the same manner as $\widehat{\mathcal{T}}$ but with a slightly different kernel, $\widetilde{K}(u) = \int_{v>u} K(v)\, dv$. Therefore, density ascent lines $\widehat{\mathcal{L}}_i$ can be explicitly related to steepest ascent curves up density surfaces. The lines start at data values and end at modes. If we want these 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 in standard L_2 terms, 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 $\widehat{T}$ are identical.

Each mode estimate defines a cluster in the dataset, equal to the set of those data X_i whose density ascent lines lead to the given mode estimate.

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 parametrised 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

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

The relationship between a data function X_i and the corresponding modal function $\hat{x}^0$, at the opposite end of $\widehat{\mathcal{L}}_i$ 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 $(\cdot,\cdot)$, with $d(x,y) = (x-y, x-y)$. Then, within a given cluster $\mathcal{C}$, a tree diagram depicting directed relationships among data values may be constructed as follows. Link $X_i \in \mathcal{C}$ to X_j, in that direction, if among all $X_j \in \mathcal{C}$ such that $\hat{p}_h(X_j) > \hat{p}_h(X_i)$, X_j minimises

$$D(X_i, X_j) = (1 + \lambda)\, d(X_i, X_j) - \lambda\, (X_j - X_i, u_i)^2 \,, \tag{2.3}$$

where $\lambda \geq 0$ and $u_i = v_i / d(0, v_i)^{1/2}$ is an element of $\mathcal{S}$ with unit length. If there exists no point $X_j \in \mathcal{C}$ 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 $\mathcal{C}$ at which $\hat{p}_h$ is maximised, and may, without essential loss of generality, be identified with the mode that corresponds to $\mathcal{C}$.

In (2.3), λ denotes a positive constant that weights the penalty term, i.e. $d(X_i, X_j) - (X_j - X_i, u_i)^2 \geq 0$, that is added to the simple distance measure $d(X_i, X_j)$ when defining $D(X_i, X_j)$. Increasing λ results in greater emphasis being given to linkages that reflect upwards movement on the surface $\widehat{\mathcal{T}}$. These techniques are related to those based on minimal spanning trees for spatial data (e.g. Florek *et al.*, 1951; Friedman and Rafsky, 1981, 1983), and to density-based versions of those techniques suggested by Cheng, Hall and Hartigan (2000).

3. NUMERICAL EXAMPLES

3.1. Preliminaries. Throughout our numerical work we use the standard Gaussian kernel. 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 daily average precipitation at 35 Canadian weather stations between 1961 and 1994. Day 1 is January 1. The data are available from J.O. Ramsay's website, `www.psych.mcgill.ca`. The raw averages are particularly noisy, and so we smooth them using the algorithm `smooth.spline` in Splus, choosing the smoothing parameter by generalised cross-validation. This produces the smooth curves in Figure 1. To ensure periodicity, each dataset is replicated three times, in the obvious manner, prior to smoothing.

The high-amplitude curve in Figure 1 corresponds to the Prince Rupert weather station. When using L_2 to measure distance either between the curves or between their derivatives, and unless the bandwidth is exceptionally large, the Prince Rupert curve consistently appears as a mode. In this respect at least it is the analogue of an isolated outlier in a Euclidean dataset, where an inordinately large bandwidth must be used to connect it to the rest of the data when constructing a density estimate. Therefore, in all subsequent analyses it is removed, leaving just 34 weather stations.

3.2. Density surface defined by L_2 distance on functions. Figure 2 shows the original 34 data curves, and the modal functions, resulting from four L_2 analyses using bandwidths equivalent to the 30th, 20th, 18th and 15th percentiles of the inter-function distances, equal respectively to $h = 16.7$, 12.1, 11.5 and 10.4. As outlined in section 2, the modal functions are obtained iteratively as limits of sequences $X_{i,j+1} = X_{ij} + \epsilon\, \hat{v}(X_{ij})$, $j \geq 0$, where $X_{i0} = X_i$ and $\epsilon > 0$ is chosen small. Here we have taken $\epsilon = 0.1$.

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 (e.g. Kostrowicki and Piela, 1991; Moré and Wu, 1997), it is unclear whether 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.

In particular, reducing h beyond the values used to prepare Figure 2 results in a larger number of modes, up to 34 for sufficiently small bandwidths. Taking ϵ too large when defining the sequence $\{X_{ij},\, j \geq 0\}$ gives noticeably different modal functions, and in some cases a different number of modes, implying that in such cases ϵ operates as a smoothing parameter. No such effects were observed for small ϵ, however, and in the case of the precipitation data it was usually sufficient to take $\epsilon < 0.2$.

The modal curves change only slightly as the bandwidth decreases. This property was observed in simulation studies 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 Figures 3–5 for the cases of 2–4 modes, respectively. The first cluster, containing just 17 curves, is the same in the 2- and 3-mode analyses. In the 4-mode case one curve, from the Thunder Bay weather station, moves from the first cluster to the second cluster. Interestingly, the cluster groups are very stable, not changing with h. For instance, values of h from 11.6 to 14.5 yield 2 modes, with the clusters remaining the same throughout.

The curves in the second cluster in Figure 3 do not have similar shapes. They correspond to similar annual precipitation levels, however, and this seems to be the effective discriminator that is operating there. Indeed, the average annual precipitation ranges for the stations in clusters 1 and 2, modulo the smoothing that is used to define the curves, are 2.14–4.06 and 0.39–1.93 mm, respectively.

As h is slightly decreased, one curve from the second cluster in the bimodal case splits off to form a third cluster; see Figure 4. This corresponds to the Victoria weather station, and it can fairly be regarded as an outlier in the second cluster in the bimodal case. Its maximum is larger and its minimum smaller than those of all other curves in that cluster, and in common with only one other curve, that for Vancouver, it has a marked V shape. In fact, as h is decreased still further the Victoria weather-station curve is joined by its counterpart for Vancouver. Thus, in the 4-mode analysis the second cluster is more uniform in shape. This uniformity, along with the cluster's relatively high average annual rainfall, now make this cluster more suitable for the Thunder Bay weather station.

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 on the raw functions, and also for the rank-based distance discussed below, modal functions change very little as bandwidth is reduced. The first split of the data into two modes places the markedly V-shaped curves for Vancouver and Victoria into the second cluster. With further reduction in the bandwidth the cluster splits into two. This split reflects the fact that relatively small differences between the steep gradients of the Vancouver and Victoria curves exceed the differences among the more shallow gradients of other curves. As h is further reduced, a fourth cluster is formed by the curve for Edmonton splitting off from the first cluster. This curve is shaped like an inverted V, and it has steepest gradient among curves of this type. Thus, when four clusters are formed there are three clusters of one curve each, and one cluster of 31 curves.

The rank distance between two curves is defined to equal the square root of 1 minus the rank correlation between the two functions. See Heckman and Zamar (2000). For bandwidths that produce up to four modes by this measure, the first modal curve is similar to its counterparts for L_2 distance on the raw functions; compare Figures 2 and 6. Modal curves for subsequent splits differ significantly from their L_2 distance counterparts, however, since rank distance focuses more sharply on shape. In particular, the cluster corresponding to the second mode consists of essentially convex curves, and its modal curve is essentially convex. This second mode is similar to the second mode obtained using L_2 distance between the derivatives. The cluster for the third mode is a singleton, equal to the irregularly shaped curve for the Kamloops weather station; and the cluster for the fourth mode consists of essentially concave curves, with an essentially concave modal curve. As with other distances, the cluster groups are very stable. When $h = 0.83$, we have two modes, $h = 0.25$ three modes and when $h = 0.21$, corresponding to the 7th percentile of interfunction distances, there are just four modes. At this bandwidth the four clusters contain respectively 20, 8, 1 and 5 curves.

3.4. Paths of the ascent lines. As noted in section 2.1, a plot of S_j versus j gives information about how the sequence X_{ij}, $j > 0$, moves to a mode, and therefore provides information about the surface represented by $s = \hat{p}_h(x)$. When analysing the precipitation data using either L_2 distance between derivatives or rank-based distances, each sequence moves with fairly constant acceleration until it slows down as it reaches the mode. This indicates that the density surface is quite smooth, with no pronounced shoulders.

However, when using L_2 distance between functions we often see a fast initial rate, slowing down to near-zero change, and then alternate acceleration and deceleration as the iterations converge to the final mode. The curve that is situated at a resting point of the trajectory is similar to a modal curve from an analysis with smaller bandwidth. For instance, the trajectories of six of the data curves in cluster 2 of Figure 3 have "resting points" in the 1-mode and 2-mode analyses, and in the 1-mode analysis the trajectories of these data curves rest momentarily at a curve that is similar to, but flatter than, the second modal function. In the 2-mode analysis, the trajectories of five of the six data curves rest at a curve that is slightly flatter than the fourth modal curve. The trajectory of the sixth data curve, namely Victoria, rests at the third modal curve. Together these results show that the density surface defined by conventional L_2 distance is far more bumpy, or unsmooth, than in the more shape-oriented approaches discussed in the previous paragraph.

3.5. Tree diagrams. In Figure 7 we give an example of a tree diagram for the thirteen weather stations in cluster 2 of Figure 5. The diagram is constructed using $\lambda = 0.3$. 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 Arvida, Bagotville, Ottawa and Montreal.

4. THEORETICAL PROPERTIES

4.1. Density of X. In section 4 we shall develop properties of our methods in the case where elements of $\mathcal{S}$ are functions defined on a region $\mathcal{R}$ in a Euclidean space, and distance between functions is measured using the square of the L_2 metric. Thus, $d(f, g) = \int (f - g)^2$. The case where the square root is taken can be recovered from our results. Of course, we may interpret f and g as the derivatives of the actual data functions, or as some other functional of them; they need not represent the raw data. In the present subsection we construct densities in finite dimensional function spaces and for general Gaussian process, and in the next we extend the results to Gaussian mixtures.

Let $\{\psi_1, \psi_2, \ldots, \psi_k\}$ be a set of orthornomal functions on $\mathcal{R}$ and suppose that the random function X can be represented as $X = \sum_i Z_i \psi_i$, where the random vector $(Z_1, \ldots, Z_k)$ has density f. Write $\bar{0}$ for the zero function and recall that $p_h(x) = P\{d(x, X) \leq h\}$. For $x = \sum_i z_i \psi_i$ it is straightforward to prove that under regularity conditions on f, $p_h(x)/p_h(\bar{0})$ converges to $f(z_1, \ldots, z_k)$ as $h \to 0$. We therefore define $p(x)$ to be $f(z_1, \ldots, z_k)$.

Now we restrict ourselves to the case where X is a Gaussian process. Let $\alpha = (\alpha_1, \alpha_2, \ldots)$ and $y = (y_1, y_2, \ldots)$ be infinite sequences of real numbers, with $\alpha_i \geq 0$ for $i \geq 1$. Define $\langle y \rangle = \sum_i y_i^2$, the square of the Euclidean norm. We shall assume that $\langle \alpha \rangle < \infty$. Given an orthonormal sequence of functions $\psi_1, \psi_2, \ldots$ on $\mathcal{R}$, let $\mathcal{S}_\alpha$ denote the class of functions x on $\mathcal{R}$ that can be expressed as

$$x = \sum_{i=1}^{\infty} y_i \, \alpha_i \, \psi_i \,, \tag{4.1}$$

with $\sum_i y_i^2 \alpha_i^2 < \infty$. In the work below, $\mathcal{S}_\alpha$ will play the role of the sample space, $\mathcal{S}$, introduced in a more general setting in section 2.1. Suppose too that the random function X can be represented as

$$X = \sum_{i=1}^{\infty} Y_i \, \alpha_i \, \psi_i \,, \tag{4.2}$$

where the random variables $Y_1, Y_2, \ldots$ are independent and have a common standard Normal distribution. Thus, X is a general zero-mean Gaussian process on $\mathcal{R}$, and (4.2) is its Karhunen-Loève expansion.

To indicate the relationship between the function x and the vector y we shall write x_y to denote the right-hand side of (4.1) for a given y, and $y(x) = (y_1(x), y_2(x), \ldots)$ for the version of y such that (4.1) holds for a given x. In this notation, define

$$\lfloor x \rfloor_\alpha = \langle y(x) \rangle = \sum_{i=1}^{\infty} \left\{ \int_{\mathcal{R}} x(u) \, \psi_i(u) \, du \right\}^2 \alpha_i^{-2} \,, \tag{4.3}$$

taking $0/0 = 0$.

We define $p(x)$ to be $\exp(-\frac{1}{2} \lfloor x \rfloor_\alpha)$. Next we relate this definition to p_h and $\hat{p}_h$. Note that this agrees with our definition of p in the finite-dimensional setting. When all but a finite number of the α_i's vanish, it is straightforward to prove that, under mild regularity conditions on the kernel, both

$$\rho_1(x) = \frac{p_h(x)}{p_h(\bar{0})} \quad \text{and} \quad \rho_2(x) = \frac{E\{\hat{p}_h(x)\}}{E\{\hat{p}_h(\bar{0})\}} \tag{4.4}$$

converge to $p(x)$ as $h \to 0$, uniformly in x. Call these properties P_1. The first ratio at (4.4) describes density based on the distance measure d, and the second can be interpreted as the bias of a standard multivariate kernel density estimator. Properties P_1 argue that p be interpreted as being proportional to the density of X.

The denominators in the ratios at (4.4) are merely convenient normalising constants, and their only important property is that they do not depend on x. The discussion in the remainder of this section would not alter in any important ways if they were omitted.

Again, suppose a finite number of α_i's are nonzero and let $x = \sum_i z_i \psi_i$. We have already discussed the case where $z_i = 0$ whenever $\alpha_i = 0$, in the finite-dimensional setting above. Suppose there exists an $\alpha_i \neq 0$ for which $z_i = 0$. Then $\exp(-\frac{1}{2}\lfloor x \rfloor_\alpha) = 0$. In view of the representation of X at (4.2), however, such functions x have probability zero of arising and so should be ascribed zero density. Therefore our definition of $p(x)$ is appropriate for all $x \in \mathcal{S}_\alpha$.

Next we treat the infinite-dimensional Gaussian case, characterised by an infinite sequence of nonzero constants α_i. By considering the legitimacy of increasingly high- but finite-dimensional approximations to this context, we see that if a density may be defined in the infinite-dimensional case it should continue to be given by $p(x) = \exp(-\frac{1}{2}\lfloor x \rfloor_\alpha)$.

Note, however, that $p(x) > 0$ if the expansion of x at (4.1) has only a finite number of nonzero terms, whereas no such vector x could arise, with positive probability, from the expansion (4.2) in the infinite-dimensional case. Contradictions such as this cannot be resolved in terms of the very limited amount of smoothing that is available from forming the probability $p_h(x) = P\{d(x, X) \leq h\}$, since finite subsequences $(Y_{i+1}, Y_{i+2}, \ldots, Y_{i+j})$ have the same likelihood no matter what the value of i. The contradictions can be removed, however, by averaging over nearby choices of the bandwidth.

We shall consider only one approach to averaging, chosen to simplify our technical argument. In place of ρ_1 at (4.4) we define, for any $t > 0$,

$$\rho_3(x) = \frac{\int_{v>0} p_{hv}(x)\, e^{-tv}\, dv}{\int_{v>0} p_{hv}(\bar{0})\, e^{-tv}\, dv},$$

and for simplicity we define the kernel estimator $\hat{p}_h$ using the exponential kernel, $K(u) = \exp(-u)$ for $u > 0$. Again, the denominator in the definition of $\rho_3(x)$ is merely a convenient normalising constant, its only important property being that it does not depend on x. Recall that our distance measure d equals the square of

L_2 distance; as a result, using the exponential kernel in the definitions of ρ_3 and $\hat{p}_h$ is equivalent to using a Gaussian kernel when L_2 distance is employed.

We claim that for each $t > 0$ in the definition of $\rho_3(x)$, and with $p(x) = \exp(-\frac{1}{2}\lfloor x \rfloor_\alpha)$, both $\rho_2(x)$ and $\rho_3(x)$ converge to $p(x)$, uniformly in x, as $h \to 0$. Call these properties P_2. In an infinite-dimensional setting they are the analogues of P_1, and they argue that p be interpreted as the density of the distribution of X. As in the finite-dimensional case the fact that $p(x) = 0$ when $x \notin \mathcal{S}_\alpha$ gives a correct definition of density.

We conclude this section by outlining a derivation of P_2. If $x_y \in \mathcal{S}_\alpha$ then the variable

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

has Laplace transform

$$\chi_y(t) = E\{\exp(-tZ_y)\} = \prod_{i=1}^{\infty} \left\{ \left(1 + 2t\alpha_i^2\right)^{-1/2} \exp\left(- \frac{t\alpha_i^2 y_i^2}{1 + 2t\alpha_i^2} \right) \right\}. \tag{4.5}$$

Now,

$$\chi_y(t) \prod_{i=1}^{\infty} (1 + 2t\alpha_i^2)^{1/2} \to p(x)\,, \tag{4.6}$$

uniformly in y, as $t \to \infty$. Write f_y for the density of Z_y, let $\vec{0}$ denote the vector $(0, 0, \ldots)$, let L be a bounded, nonnegative, piecewise-continuous function, and let λ denote its Laplace transform. The Laplace transform of

$$\ell_y(v) = \int_0^{hv} L\{v - (u/h)\}\, f_y(u)\, du = h \int_0^{v} L(v - u)\, f_y(hu)\, du$$

equals $\lambda(t)\, \chi_y(t/h)$. From this property and (4.6), which may be written equivalently as $\chi_y(t)/\chi_{\vec{0}}(t) \to \exp(-\frac{1}{2}\lfloor x \rfloor_\alpha)$ as $t \to \infty$, we deduce that at each argument t of the respective Laplace transforms,

$$\frac{\text{Laplace transform of } \ell_y(\cdot)}{\text{Laplace transform of } \ell_{\vec{0}}(\cdot)} \to p(x)\,, \tag{4.7}$$

uniformly in y, as $h \to 0$.

Taking $L(u) = I(u \geq 0)$ or $L(u) = M(-u)$, for a nonnegative function M defined on the whole real line, we have, respectively,

$$\ell_y(v) = P\{d(x, X) \leq hv\} \quad \text{or} \quad \ell_y(v) = E\{\hat{p}_h(x)\}\,,$$

the later constructed using the kernel $K(u) = K_v(u) \equiv M(u - v)$. The Laplace transform of $\ell_y(\cdot)$ is then equal to the numerator in the definition of $\rho_3(x)$, and to a constant multiple of the numerator in the definition of $\rho_2(x)$, respectively, with the kernel in the latter case being $K(u) = \exp(-tu)$ and the constant not depending on y. Property P_2 follows from these results and (4.7).

4.2. Density of a mixture of Gaussian processes. Suppose there are $k \geq 1$ components in the Gaussian mixture, which is constructed by taking X equal to

$$X^{(j)} = \mu^{(j)} + \sum_{i=1}^{\infty} Y_i^{(j)} \, \alpha_i^{(j)} \, \psi_i^{(j)} \,,$$

with probability $\pi^{(j)}$ for $1 \leq j \leq k$, where (a) the variables $Y_i^{(j)}$ for $i \geq 1$ and $1 \leq j \leq k$ are all independent Normal $N(0,1)$, independent too of the process of choosing among the $X^{(j)}$'s, (b) the components of the weight vectors $\alpha = \alpha^{(j)} = (\alpha_1^{(j)}, \alpha_2^{(j)}, \ldots)$ are nonnegative and each satisfies $\langle \alpha^{(j)} \rangle < \infty$, (c) for each j the functions $\psi_1^{(j)}, \psi_2^{(j)}, \ldots$ form a complete orthonormal sequence on a subset $\mathcal{R}$ of Euclidean space, and (d) for each j the function $\mu^{(j)}$ is representable as

$$\mu^{(j)} = \sum_{i=1}^{\infty} u_i^{(j)} \, \alpha_i^{(j)} \, \psi_i^{(j)} \,,$$

with $\sum_i (u_i^{(j)} \, \alpha_i^{(j)})^2 < \infty$. This prescription represents X as a general Gaussian-process mixture; in particular we do not require the orthonormal sequences in Karhunen-Loève expansions of the component processes $X^{(j)}$ to be identical.

Define $\alpha^{(j_1)}/\alpha^{(j_2)}$ to equal 1 if it is of the form $0/0$, and to equal infinity if it has the form of a nonzero quantity divided by zero. Define infinity multiplied by zero to equal infinity. Assume there exists a "lightest" weight vector $\alpha^{(j_0)}$, in the sense that, using the just-mentioned conventions, the quantities

$$c^{(j)} = \prod_{i=1}^{\infty} \frac{\alpha_i^{(j_0)}}{\alpha_i^{(j)}} \,, \quad 1 \leq j \leq k \,,$$

are all well-defined and finite. Put

$$p(x) = \sum_{j=1}^{k} c^{(j)} \, \pi^{(j)} \, \exp\big(-\tfrac{1}{2} \lfloor x - \mu^{(j)} \rfloor_{\alpha^{(j)}} \big) \,. \tag{4.8}$$

The properties P_1 and P_2 continue to hold for this choice of p, provided an appropriate choice is made for the denominators in the definitions of $\rho_j(x)$. Of course, the new denominators do not depend on x. Therefore $p(x)$ may be regarded as a density of the process X.

Just as in the context of finite-dimensional mixtures, there are obviously possibilities for degeneracy where $p(x) = 0$ for all x. Moreover, in some settings where p does not vanish, the formula at (4.8) may ascribe zero mass to the jth component of the mixture merely because it can exhibit, with positive probability, a basis element $\psi_i^{(j)}$ that is always orthogonal to each basis element in the dominant component with index j_0. Again this reflects abnormalities that also arise in finite-dimensional multivariate problems.

4.3. Properties of density estimates. In a finite-dimensional setting the convergence rate of $\hat{p}_h$ to p is polynomially fast as a function of n. We should of course define the convergence rate in relative terms, for example as the rate at which

$$\Delta(x) = \frac{\hat{p}_h(x)}{E\{\hat{p}_h(\bar{0})\}} - p(x)$$

converges to zero. If all but k of the α_i's vanish, and we take h to be of size $n^{-2/(k+4)}$, then $\Delta(x) = O_p(n^{-4/(k+4)})$. This convergence rate exactly reflects that of a conventional k-variate density estimator, and may be derived using an argument that is familiar in that setting.

The rate in infinite-dimensional settings is slower than the inverse of any power of n. As we noted in section 1, numerical accuracy of density estimators is often not a major issue in functional data analysis, and so this theoretical result is not necessarily of direct practical bearing. Moreover, we may deduce from the nature of low-dimensional approximations that even the result's mathematical implications may not be serious unless, in the Gaussian case, a substantial amount of emphasis in the expansion (4.2) of X is given to a large number of terms, i.e. the α_i's converge slowly to zero. When low-dimensional approximations provide adequate information about properties of the distribution of X, the mathematical approximation to p afforded by $\hat{p}_h$ is often good, even in moderately-sized samples; see Scott and Wand (1991) for closely related results in more conventional settings.

Nevertheless, the rate of convergence of density estimators is generally a topic of methodological interest, and so we conclude this section with an example of properties in the infinite-dimensional case. Assume the context of section 4.1, that

$\alpha_i \sim ci^{-\beta}$ as $i \to \infty$ where $c > 0$ and $\beta > \frac{1}{2}$, and that $x \in \mathcal{S}_\alpha$ has the property

$$\sum_{i=1}^{\infty} y_i(x)^2 \, \alpha_i^{-2} < \infty \,. \tag{4.9}$$

Defining x^α to be the function $\sum_i y_i(x)\, \psi_i$ we see that $\lfloor x^\alpha \rfloor_\alpha$ equals the left-hand side of (4.9). Take $K(u) = e^{-u}$ for $u > 0$. Then it may be shown that the optimal rate of convergence of $\Delta(x)$ to zero equals $O_p\{(\log n)^{-2\beta}\}$, and is achieved with a bandwidth equal to a sufficiently large constant multiple of $(\log n)^{-2\beta}$. Call these properties P_3. Slower convergence rates, without requiring (4.9), may be derived by similar arguments.

To derive P_3, assume that $h = h(n) \to 0$ as $n \to \infty$, and note that in view of (4.5) we have as $t \to \infty$,

$$\chi_y(t)/\{\chi_{\bar{0}}(t)\, p(x)\} = \exp\left(\tfrac{1}{2} \sum_{i=1}^{\infty} \frac{y_i^2}{1 + 2t\alpha_i^2}\right) = 1 + \tfrac{1}{4} \lfloor x^\alpha \rfloor_\alpha \, t^{-1} + O(t^{-2})\,.$$

Therefore, result (4.7) may be refined to the following: the ratio of the Laplace transform of $\ell_y(v)$ to that of $\ell_{\bar{0}}(v)$, with each evaluated at $t = 1$, equals $p(x)\,(1 + \frac{1}{4} \lfloor x^\alpha \rfloor_\alpha \, h) + O(h^2)$. Equivalently,

$$E\{\Delta(x)/p(x)\} = \frac{E\{\hat{p}_h(x)\}}{E\{\hat{p}_h(\bar{0})\, p(x)\}} - 1 = \tfrac{1}{4} \lfloor x^\alpha \rfloor_\alpha \, h + O(h^2)\,. \tag{4.10}$$

Also, the variance of $\Delta(x)$ is dominated by

$$\frac{E\{L_h(x, X)\}}{n\, [E\{K_h(x, X)\}]^2}\,, \tag{4.11}$$

where $K = M^{[1]}$, $L = M^{[2]}$, $M^{[r]}(u) = e^{-ru}$ and $M_h(x, X) = M\{d(x, X)/h\}$. Arguing as in the proof of property P_2 in section 4.1 we may show that $E\{K_h(x, X)\}$ and $E\{L_h(x, X)\}$ are asymptotic to $p(x)$ multiplied by, respectively, $B_j \prod_i (1 + 2jh^{-1}\alpha_i^2)^{-1/2}$ for $j = 1, 2$, where the nonzero constants B_j do not depend on x. Now,

$$\sum_{i=1}^{\infty} \log\left(1 + 2jh^{-1}\alpha_i^2\right) \sim C_j h^{-1/(2\beta)}\,,$$

where $0 < C_j < \infty$ and $C_2 = 2^{1/(2\beta)} C_1$. Therefore the ratio at (4.11) equals $n^{-1} \exp(-\xi_n \, D\, h^{-1/(2\beta)})$, where $D > 0$ and $\xi_n \to 1$ as $n \to \infty$. Properties P_3 follow from this result and (4.10).

4.4. Properties of $\widehat{T}$ and $\hat{x}^0$. We first relate $\hat{v}(x) = \widehat{T}(x) - x$ to the direction of steepest ascent up a surface. Let $\tilde{p}_h(x)$ be as at (2.1) but with K_h there replaced by $\widetilde{K}_h$, where $\widetilde{K}_h(u) = \widetilde{K}(u/h)$ and $\widetilde{K}(u) = \int_{v>u} K(v)\, dv$. Note that, for $K(u) \equiv e^{-u}$, $\widetilde{K} \equiv K$ and so $\tilde{p}_h \equiv \hat{p}_h$. The formula below shows that $\hat{v}(x)$ is in the direction of steepest ascent up the surface $\widetilde{\mathcal{T}}$ defined by $s = \tilde{p}_h(x)$: for $g \in \mathcal{S}$,

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

where $(\cdot, \cdot)$ is the inner product on $\mathcal{S}$.

It follows that a density ascent line $\widehat{\mathcal{L}}_i$, defined by the limiting form (as $\epsilon \to 0$) of the algorithm $X_{ij} \mapsto X_{i,j+1}$ introduced in section 2.1, moves steadily higher up the surface $\widetilde{\mathcal{T}}$. Excepting possible pathological cases where the line keeps moving uphill but never actually converges, it will terminate at a point $\hat{x}^0$ that is either a modal function or a saddlepoint of the surface defined by $s = \tilde{p}_h(x)$. In the finite, k-dimensional case, assume 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 saddlepoints, and $n^{-\{2/(k+4)\}+\delta} \leq h \leq n^{-\delta}$ 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.

4.5. Definition of density contours and ridge lines. The contour corresponding to a density p is a level set of p, and so is defined by $\mathcal{T}_s = \{x : p(x) = s\}$ where $0 < s \leq \sup p$. To define a ridge line, let $\tilde{p}_h'(x)(\cdot) : \mathcal{S} \to \mathcal{R}$ be the gradient functional defined in section 4.4, and let $\|\tilde{p}_h'(x)\|$ denote its norm. Then $x = x^1 \in \mathcal{T}_s$ is on a ridge line if and only if $\|\tilde{p}_h'(x)\|$ achieves a local minimum there among functions in $\mathcal{T}_s$. Under regularity conditions, for each s there will only be a finite number of such x^1's and these x^1's 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 2-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, in more than one dimension, a contour and a ridge line are both of co-dimension 1, and so their structures are representable using one less parameter than the entire space.

REFERENCES

CHOI, E. AND HALL, P. (1999). Data sharpening as a prelude to density estimation. *Biometrika* **86**, 941–947.

CHENG, M.-Y., HALL, P. AND HARTIGAN, J.A. (2000). Estimating gradient trees. *J. Computat. Graph. Statist.*, to appear.

FLOREK, K., LUKASZWICZ, J., PERKAL, J., STEINHAUS, H. AND ZUBRZYCKI, S. (1951). Sur la liaison et la division des points d'un ensemble finit. *Colloq. Math.* **2**, 282–285.

FRIEDMAN, J.H. AND RAFSKY, L.C. (1981). Graphics for the multivariate two-sample problem. (With discussion.) *J. Amer. Statist. Assoc.* **76**, 277–295.

FRIEDMAN, J.H. AND RAFSKY, L.C. (1983). Graph-theoretic measures of multivariate association and prediction. *Ann. Statist.* **11**, 377–391.

GASSER, T., HALL, P. AND PRESNELL, B. (1998). Nonparametric estimation of the mode of a distribution of random curves. *J. Roy. Statist. Soc. Ser. B* **60**, 681–691.

HALL, P., QIAN, W. AND TITTERINGTON, D.M. (1992). Ridge finding from noisy data. *J. Computat. Graph. Statist.* **1**, 197–211.

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

HECKMAN, N.E. AND ZAMAR, R.H. (2000). Comparing the shapes of regression functions. *Biometrika* **87**, 135–144.

KNEIP, A. AND GASSER, Th. (1992). Statistical tools to analyze data representing a sample of curves. *Ann. Statist.* **20**, 1266–1305.

KOSTROWICKI, J. AND PIELA, L. (1991). Diffusion equation method of global minimization: performance for standard functions. *J. Optim. Theory Appl.* **69**, 269–284.

LEURGANS, S.E., MOYEED, R.A. AND SILVERMAN, B.W. (1993). Canonical correlation analysis when the data are curves. *J. Roy. Statist. Soc. Ser. B* **55**, 725–740.

MORÉ, J.J. AND WU, Z. (1997). Global continuation for distance geometry problems. *SIAM J. Optimiz.* **7**, 814–836.

PEZZULLI, S.D. AND SILVERMAN, B.W. (1993). Some properties of smoothed principal components analysis for functional data. *Comput. Statist.* **8**, 1–16.

RAMSAY, J.O., BLOCK, R.D. AND GASSER, T. (1995). Comparison of height acceleration curves in the Fels, Zurich, and Berkeley growth data. *Ann. Human Biol.* **22**, 413–26.

RAMSAY, J.O. AND LI, X. (1998). Curve registration. *J. Roy. Statist. Soc. Ser. B*

60, 351–363.

RAMSAY, J.O. AND SILVERMAN, B.W. (1997). *Functional Data Analysis.* Springer, New York.

RICE, J.A. AND SILVERMAN, B.W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B* **53**, 233–243.

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

SCOTT, D.W. AND WAND, M.P. (1991). Feasibility of multivariate density estimates. *Biometrika* **78**, 197–205.

SILVERMAN, B.W. (1981). Using kernel density estimates to investigate multimodality. *J. Roy. Statist. Soc. Ser. B* **43**, 97–99.

SILVERMAN, B.W. (1995). Incorporating parametric effects into functional principal components analysis. *J. Roy. Statist. Soc. Ser. B* **57**, 673–689.

SWAYNE, D.F., COOK, D. AND BUJA, A. (1991). XGobi: Interactive dynamic graphics in the X window system with a link to S. In *ASA Proceedings of the Section on Statistical Graphics*, pp.1–8.

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

WANG, K. AND GASSER, Th. (1997). Alignment of curves by dynamic time warping. *Ann. Statist.* **25**, 1251–1276.

Captions for figures

Figure 1: *Smoothed precipitation curves of 35 Canadian weather stations.* The curve for Prince Rupert is represented by the short-dashed line in the upper part of the figure, and was removed from subsequent analyses.

Figure 2: *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, due to a different bandwidth being used, but the the changes are slight. Therefore, the first,..., fourth modal functions are easily recognised; the first k of these are represented in the kth panel.

Figure 3: *Bimodal case.* Shown in the first and second panels are the data curves whose density ascent curves converge to the first and second modal function respectively, in the second panel of Figure 2.

Figure 4: *Trimodal case.* As for Figure 3, but with modal functions 1 through 3.

Figure 5: *Four-mode case.* As for Figure 3, but with modal functions 1 through 4.

Figure 6: *Modal curves for four rank-distance analyses using different bandwidths.* As for Figure 2, but using rank distance.

Figure 7: *Tree diagram for the thirteen weather stations in cluster 2 of Figure 5.*