CONCEPT OF DENSITY FOR FUNCTIONAL DATA

AURORE DELAIGLE
U MELBOURNE & U BRISTOL

PETER HALL
U MELBOURNE & UC DAVIS

Concept of Density in Functional Data Analysis

The notion of probability density for a random function is becoming increasingly important in functional data analysis.

For example, it underpins discussion of the mode of the distribution of a random function, addressed by Gasser, Hall and Presnell (1998), Hall and Heckman (2002) and Dabo-Niang, Ferraty and Vieu (2004 [×2], 2006).

Nonparametric or structure-free methods for curve estimation, from functional data, involve the concept of density, not least because they generally are based on estimators of Nadaraya-Watson type, which require division by an estimator of a small-ball probability. See, for example, Ferraty, Goïa and Vieu (2002 [×2], 2007 [×2]), Ferraty and Vieu (2002, 2003, 2004, 2006 [×2]) and Niang (2002).

Outline of Results and Methodologies

We shall argue that the concept of a density for functional data generally cannot be well defined, at least not in a conventional sense.

Functional data are often described through analysis in the space determined by the eigenfunctions obtained by principal component representation of the function's covariance. We shall show that, when considering the data from this viewpoint, it is possible to define a meaningful concept of density for a specific scale or resolution level, which is intrinsically linked to a particular dimension.

The challenge is to determine that dimension. We shall give an argument which leads directly from scale to dimension, through a simple approximation to a small-ball probability at a given scale.

Outline of Results and Methodologies – 2

Once we have this approximation it is feasible to by-pass scale altogether, and develop a simple approximation to density for any specific dimension, founded only on the eigenvalue sequence and on the sequence of densities of principal component scores.

The density approximation also suggests a simple and appealing definition of mode, and leads directly to an empirical approximation to density for a given dimension.

Our empirical methods involve estimating the densities of principal component scores, using approximations to those scores based on estimators of eigenvalues and eigenfunctions.

This problem is itself of intrinsic interest, not least because principal component score densities reveal interesting shape differences.

Quick Summary of PCA for Functional Data

Let X be a random function supported on a compact interval $\mathcal{I}$. If the covariance function of X is positive definite, it admits a spectral decomposition:

$$K(s,t) \equiv \operatorname{cov}\{X(s), X(t)\} = \sum_{j=1}^{\infty} \theta_j \, \psi_j(s) \, \psi_j(t) \,,$$

where the expansion converges in mean square, and $\theta_1 \geq \theta_2 \geq \ldots$ are the eigenvalues, with respective orthonormal eigenvectors ψ_j, of the linear operator with kernel K.

The functions $\psi_1, \psi_2, \ldots$ form a basis for the space of all square-integrable functions on $\mathcal{I}$, and, in particular we can write, for X and any square-integrable function x on $\mathcal{I}$,

$$X = \sum_{j=1}^{\infty} \theta_j^{1/2} \, X_j \, \psi_j \,, \quad x = \sum_{j=1}^{\infty} \theta_j^{1/2} \, x_j \, \psi_j \,,$$

where the quantities $X_j = \theta_j^{-1/2} \int_{\mathcal{I}} X \, \psi_j$ and $x_j = \theta_j^{-1/2} \int_{\mathcal{I}} x \, \psi_j$ are the principal component scores (sometimes referred to as the principal components) corresponding to functions X and x.

Quick Summary of PCA for Functional Data – 2

The X_j's are always uncorrelated (this follows from orthogonality of the ψ_j's), and we shall assume that they are independent. This is exactly correct if X is a Gaussian process, and it is almost always assumed to be the case in empirical or numerical work.

In such cases, as here, independence is often interpreted pragmatically; it captures the main features of a population, allows relatively penetrating theoretical analysis, and motivates simple, practical methodology.

Why Functional Data Generally Cannot Have a Probability Density

If X is a random vector of finite length then we generally define the probability density, $f(x)$, of X at the point x, as the limit as h decreases to zero of the probability that X lies in the ball of radius h centred at x, divided by the Lebesgue measure of that ball.

For example, in Euclidean space of dimension r,

$$f(x) = \lim_{h \downarrow 0} \left(h^r \, v_r\right)^{-1} P(\|X - x\| \leq h)\,, \tag{1}$$

where $\| \cdot \|$ denotes Euclidean distance in $\mathbb{R}^r$, and v_r represents the content of the r-dimensional unit sphere.

It might be expected that a formula analogous to (1), with the divisor $h^r \, v_r$ replaced by a different function of h, would be appropriate for estimating the probability density of a random function X. However, in general it is not.

Why Functional Data Generally Cannot Have a Probability Density – 2

To appreciate why, let X be a random function and x a fixed function, and let $\|X - x\|$ denote the L_2 distance between X and x. If there were to exist a function, $\alpha(h)$ say, such that the probability density

$$f(x) = \lim_{h \downarrow 0} \{\alpha(h)\}^{-1} P(\|X - x\| \leq h)$$

were well defined, then for all x we would have

$$\log f(x) = \lim_{h \downarrow 0} \Big[- \log \big\{\alpha(h)\big\} + \log P(\|X - x\| \leq h) \Big] ,$$

and thus

$$\log P(\|X - x\| \leq h) = C_1 + \log f(x) + o(1) , \qquad (2)$$

where $C_1 = \log \big\{\alpha(h)\big\}$ does not depend on x, and $f(x)$ does not depend on h.

Why Functional Data Generally Cannot Have a Probability Density – 3

However, it can be shown that

$$\log P(\|X - x\| \leq h) = C_1(r, \theta) + \sum_{j=1}^{r} \log f_j(x_j) + o(r)\,, \tag{3}$$

where $r = r(h)$ diverges to infinity as h decreases to zero, f_j is the density of the jth principal component score, x_j is the version of that score for the function x, and both r and the constant C_1 depend on h and on the infinite eigenvalue sequence, θ say. (Neither r nor C_1 depends on x or on the distributions of the principal component scores.)

The term $\sum_{j \leq r} \log f_j(x_j)$ in (3) typically diverges at rate r as r is increased, and in particular it is generally not equal to $o(r)$. It therefore dominates the x component in equation (3), and so (2) cannot hold.

Why Functional Data Generally Cannot Have a Probability Density – 4

For example, in the simple case where the random function X is a Gaussian process, (3) becomes:

$$\log P(\|X - x\| \leq h) = C_2(r, \theta) - \tfrac{1}{2} \sum_{j=1}^{r} x_j^2 + o(r)\,, \tag{4}$$

where $C_2(r, \theta)$ denotes another constant and $x_1, x_2, \ldots$ is a realisation of a sequence of independent standard normal random variables.

Clearly, the term $\frac{1}{2} \sum_{j \leq r} x_j^2$ is generally not $o(r)$, even if the sequence of x_j's is bounded. Therefore, it dominates the part of (4) that depends on x, which is thus not generally bounded.

We could replace the term $\frac{1}{2} \sum_{j \leq r} x_j^2$ by its expected value, $\frac{1}{2}\, r$, in the case where the x_j's are taken to be normal $\mathrm{N}(0, 1)$, but that would not lead to an expansion that depended on x in a useful and meaningful way. These issues prevent a general, rigorous definition of a probability density for functional data.

Log-Densities

Recall formula (3):

$$\log P(\|X - x\| \leq h) = C_1(r, \theta) + \sum_{j=1}^{r} \log f_j(x_j) + o(r)\,,$$

where $r = r(h)$ diverges to infinity as h decreases to zero, f_j is the probability density of the jth principal component score, x_j is the version of that score for the function x, and both r and the constant C_1 depend on h and on the infinite eigenvalue sequence, θ say. (Neither r nor C_1 depends on x or on the distributions of the principal component scores.)

The log-density,

$$\ell(x \,|\, r) = \frac{1}{r} \sum_{j=1}^{r} \log f_j(x_j)\,,$$

captures the variation, with x, of the logarithm of the small-ball probability $p(x \,|\, h) = P(\|X - x\| \leq h)$, up to and including terms of order r, and in particular gives rise to a remainder of strictly smaller order than r.

Log-Densities – 2

In other words, the log-density function describes the main differences in the sizes of small-ball probabilities for different values of x.

Moreover, up to terms that are negligible relative to those captured by the log-density $\log p(x \,|\, h)$, $\ell(x \,|\, r)$ is a monotone increasing function of $p(x \,|\, h)$.

While $\ell(x \,|\, r)$ cannot, in general, be employed to compare densities for different random function distributions, it can be used as the basis for comparing density at different points x for the same random function distribution and for dimension r.

ESTIMATING LOG-DENSITIES

The first step is to estimate the eigenvalues θ_j and eigenfunctions ψ_j. Starting from independent data $X_{[1]}, \ldots, X_{[n]}$ on X, compute

$$\widehat{K}(s,t) = \frac{1}{n} \sum_{i=1}^{n} \{X_{[i]}(s) - \bar{X}(s)\} \{X_{[i]}(t) - \bar{X}(t)\} = \sum_{j=1}^{\infty} \hat{\theta}_j \, \hat{\psi}_j(s) \, \hat{\psi}_j(t) \, , \qquad (5)$$

where $\bar{X} = n^{-1} \sum_i X_{[i]}$ and the terms are ordered so that $\hat{\theta}_1 \geq \hat{\theta}_2 \geq \ldots$.

(We use square-bracketed subscripts so as not to confuse the ith data value $X_{[i]}$ with the ith principal component score, X_i, of X. The expansion (5) is the empirical analogue of a theoretical one given earlier: $K(s,t) = \sum_j \theta_j \, \psi_j(s) \, \psi_j(t)$.)

We interpret $\hat{\theta}_j$ and $\hat{\psi}_j$ as estimators of the eigenvalues θ_j and eigenfunctions ψ_j, respectively.

ESTIMATING LOG-DENSITIES – 2

Next we calculate approximations

$$\hat{X}_{[ij]} = \hat{\theta}_j^{-1/2} \int_{\mathcal{I}} (X_{[i]} - \bar{X})\, \hat{\psi}_j\,, \quad \hat{x}_j = \hat{\theta}_j^{-1/2} \int_{\mathcal{I}} (x - \bar{X})\, \hat{\psi}_j$$

to the principal components $X_{[ij]} = \theta_j^{-1/2} \int_{\mathcal{I}} (X_{[i]} - EX_{[i]})\, \psi_j$ and to $x_j = \theta_j^{-1/2} \int_{\mathcal{I}} (x - EX)\, \psi_j$.

An estimator $\hat{f}_j$ of the probability density function f_j of $\theta_j^{-1/2}\, (X_j - EX_j)$ can be computed using standard kernel methods:

$$\hat{f}_j(u) = \frac{1}{nh} \sum_{i=1}^{n} W\left(\frac{\hat{X}_{[ij]} - u}{h} \right),$$

where h denotes a bandwidth and W is a kernel function.

Estimating Log-Densities – 3

Our estimator of the log-density $\ell(x \,|\, r) = r^{-1} \sum_{1 \le j \le r} \log f_j(x_j)$ is

$$\hat{\ell}(\hat{x} \,|\, r) = \frac{1}{r} \sum_{j=1}^{r} \log \hat{f}_j(\hat{x}_j)\,.$$

An attractive feature of $\hat{\ell}(\hat{x} \,|\, r)$ is the ease with which it can be computed for a range of values of r.

The value of h, for any given density estimator $\hat{f}_j$, can be chosen using standard methods for random data, reflecting the fact that $\hat{X}_{[ij]}$, for $1 \le i \le n$, is an approximation to the independent sequence $\theta_j^{-1/2}\,\{X_{[ij]} - E(X_{[ij]})\}$, $1 \le i \le n$.

Australian Rainfall Data

We applied our density estimation methods to Australian rainfall data. The data consist of rainfall measurements at each of 204 Australian weather stations, and the function $X(t)$ represents the rainfall at time t, where t denotes the period that has passed, in a given year, at the time of measurement.

Rainfall at time t was averaged over the years for which the station had been operating (40 to 100 years in each case), with the aid of a local polynomial smoother passed through discrete observations. We scaled t so that it took its values in $[0, 365]$.

The following figure depicts the yearly rainfall curves. At the top we show those stations (usually located in the north) which exhibit a "tropical" pattern, i.e. those where most rain fell in mid to late summer; and at the bottom we show the stations (usually in the south) where the majority of rain came in cooler months.

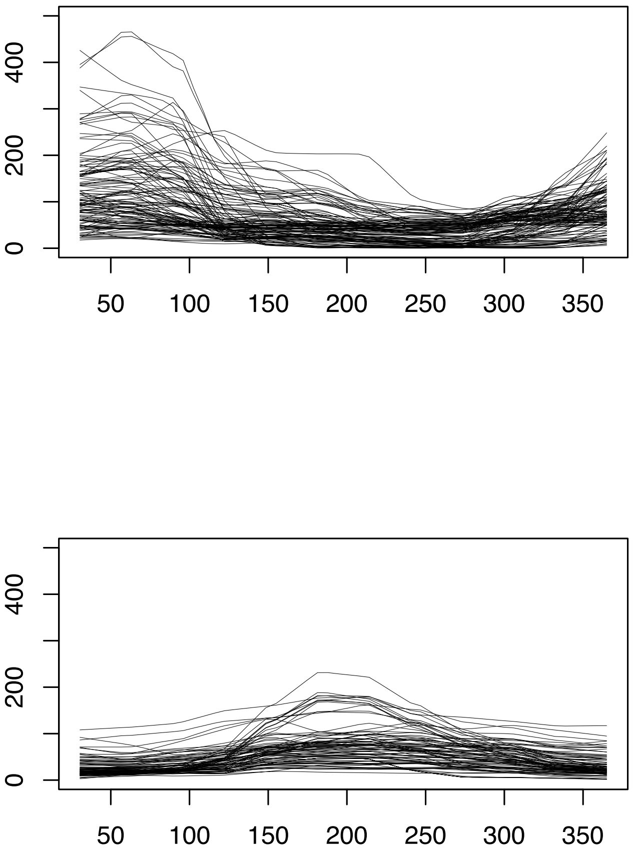

Figure 1: Australian rainfall data at weather stations which get the most rainfall during the summer months (left) or during the winter months (right).

LOG-DENSITY ESTIMATORS FOR RAINFALL DATA

The log-density estimator $\hat{\ell}(\hat{x} \,|\, r)$ cannot be directly visualised at resolution levels greater than $r > 2$.

To see the effect that increasing r has on $\hat{\ell}(\hat{x} \,|\, r)$, we calculated this density for $r = 1, \ldots, 10$ and for $x = X_{[1]}, \ldots, X_{[n]}$ (i.e. for each data curve), and then, for each r, we classified the n data curves into several groups according to the value of $\hat{\ell}(\hat{X}_{[i]} \,|\, r)$, using (in the figure) colours ranging from blue for the lowest values of $\hat{\ell}(\cdot \,|\, r)$, to yellow for the largest values.

The figure shows the groups of curves obtained for $r = 2$ (left-hand column) and $r = 10$ (right-hand column).

In each column, the first graph shows all the curves and the other graphs show, from top to bottom, groups of curves for indices i that correspond to increasing values of $\hat{\ell}(X_{[i]} \,|\, r)$.

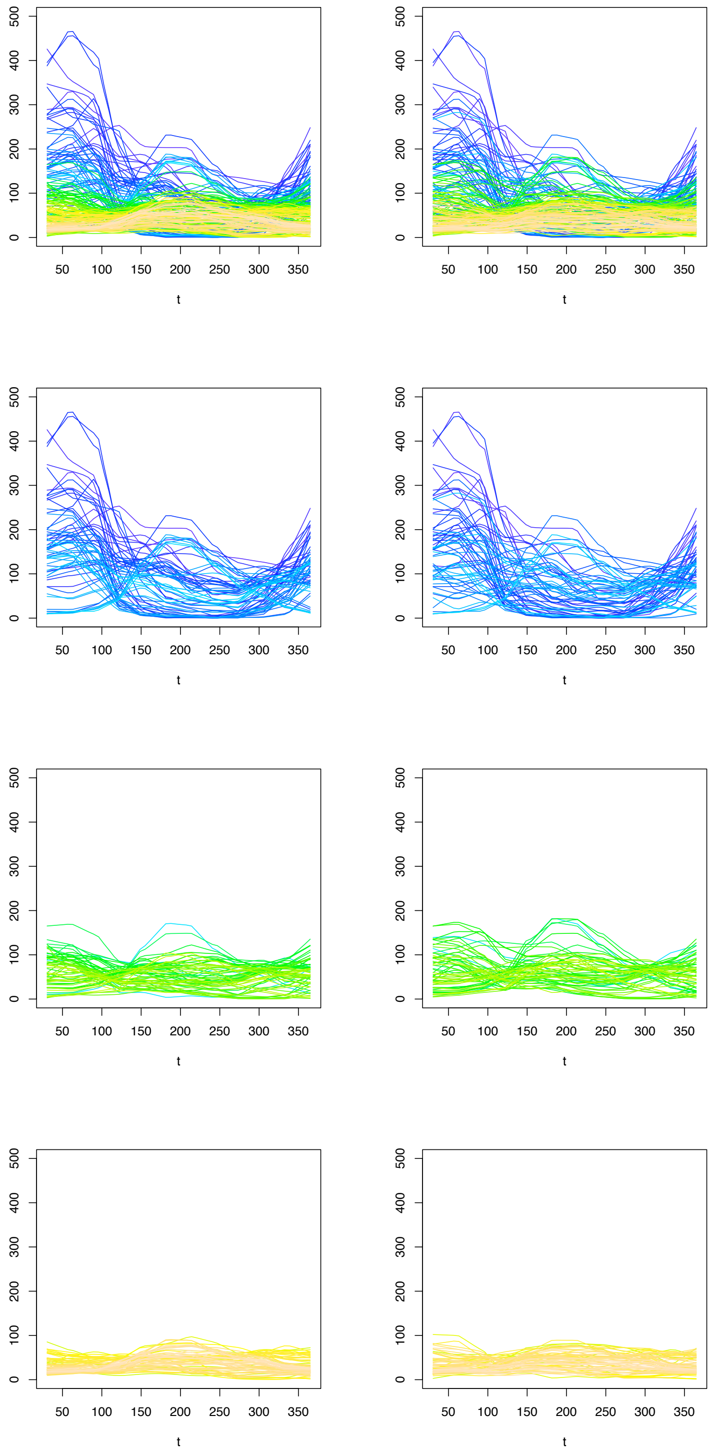

Figure 2: Plots of the 204 rain data curves. In each column, the first graph shows all the curves and the other graphs show, from top to bottom, groups of curves for indices i that correspond to increasing values of $\hat{\ell}(X_{[i]} \,|\, r)$. The left-hand column depicts results when $r = 2$, whereas the right-hand column corresponds to $r = 10$.

Interpretation of Figure

We see that overall the curves of low (respectively, moderate or high) density for $r = 2$ correspond to the curves of low (respectively, moderate or high) density for $r = 10$.

In other words, the density at resolution level $r = 2$ already reflects the main features of the data.

The blue curves roughly correspond to the stations whose rainfall varies the most over the year; these stations are very heterogeneous and thus have low density.

At the other end of the spectrum, the yellow curves correspond to the stations with the flattest yearly rainfall; these stations are quite homogeneous and logically, they have the highest density.

The green curves correspond to a moderate rainfall change over the year and have moderate density values.