J. R. Statist. Soc. B (1998)
60, *Part* 2, *pp.* 351–363

Curve registration

J. O. Ramsay† and Xiaochun Li

McGill University, Montreal, Canada

[Received March 1996. Revised April 1997]

Summary. Functional data analysis involves the extension of familiar statistical procedures such as principal components analysis, linear modelling and canonical correlation analysis to data where the raw observation x_i is a function. An essential preliminary to a functional data analysis is often the registration or alignment of salient curve features by suitable monotone transformations h_i of the argument t, so that the actual analyses are carried out on the values $x_i\{h_i(t)\}$. This is referred to as dynamic time warping in the engineering literature. In effect, this conceptualizes variation among functions as being composed of two aspects: horizontal and vertical, or domain and range. A nonparametric function estimation technique is described for identifying the smooth monotone transformations h_i and is illustrated by data analyses. A second-order linear stochastic differential equation is proposed to model these components of variation.

Keywords: Dynamic time warping; Geometric Brownian motion; Monotone functions; Spline; Stochastic time; Time warping

1. Introduction

Techniques in functional data analysis (Ramsay and Silverman, 1997) can be employed to study the variation in a sample of functions x_i, $i = 1, \ldots, N$, and their derivatives. In practice these functions are often a consequence of a preliminary smoothing process applied to discrete data, and in others the entire functions may be immediately available by on-line recording techniques.

Fig. 1 illustrates a problem that can frustrate even the simplest analyses of replicated curves. 10 estimates of the acceleration in height show individually the salient features of growth in children: the large deceleration during infancy is followed by a rather complex but small acceleration phase during late childhood, and then the dramatic acceleration–deceleration pulses of the pubertal growth spurt finally give way to zero acceleration in adulthood. The timing of these salient features obviously varies from child to child. Ignoring this timing variation in computing a cross-sectional mean function (the bold broken curve in Fig. 1) can result in an estimate of average acceleration that does not resemble any of the observed curves: the mean curve has less variation during puberty than any single curve, and the duration of the mean pubertal growth spurt is rather larger than for any individual curve.

Fig. 2 displays a similar problem for mean temperature records of two Canadian cities; the marine climate of St John's, Newfoundland, is associated with rather later seasons than is the continental climate of Edmonton, Alberta. Before studying other ways in which the two curves differ, we need to consider how their seasons can be compared on the same timescale.

Fig. 3(a) presents a particularly common registration problem. In an experiment described

†*Address for correspondence*: Department of Psychology, Stewart Biological Sciences Building, McGill University, 1205 Dr Penfield Avenue, Montreal, Quebec, H3A 1B1, Canada.
E-mail: ramsay@psych.mcgill.ca

 1369–7412/98/60351

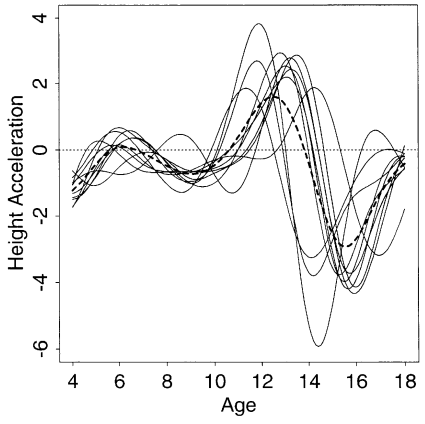

Fig. 1. 10 height acceleration curves (in centimetres per year squared) for boys estimated by Ramsay, Bock and Gasser (1995): - - - -, cross-sectional mean, illustrating the fact that averaging unregistered curves can result in an average that does not resemble any sample curve

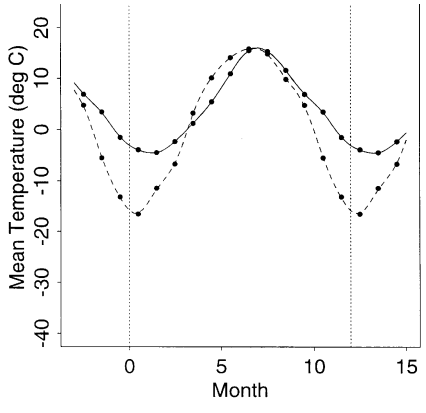

Fig. 2. Mean temperature records for Edmonton, Alberta (- - - -), and St John's, Newfoundland (———)

in Ramsay, Wang and Flanagan (1995), the force exerted by the thumb and forefinger was recorded during 20 brief pinches applied to a force meter, with a background force of about 2 N applied before and after the pinch. The starting time for each record was arbitrary, so it was essential to find a common timescale to combine information across the records.

These examples illustrate that the rigid metric of physical time may not be directly relevant to the internal dynamics of many real life systems. Rather, there can be a sort of physiological or meteorological timescale that relates non-linearly to physical time and varies from case to case. Human growth is, ignoring external factors, largely a consequence of a complex sequence of hormonal events that do not happen at the same rate from child to child and also have a variable rate over the growth of a specific child. Weather is driven by ocean currents, reflectance changes for land surfaces and other factors that are timed differently for different spatial locations. And finally muscle contractions do not build up and release at exactly the same rate from one pinch to another.

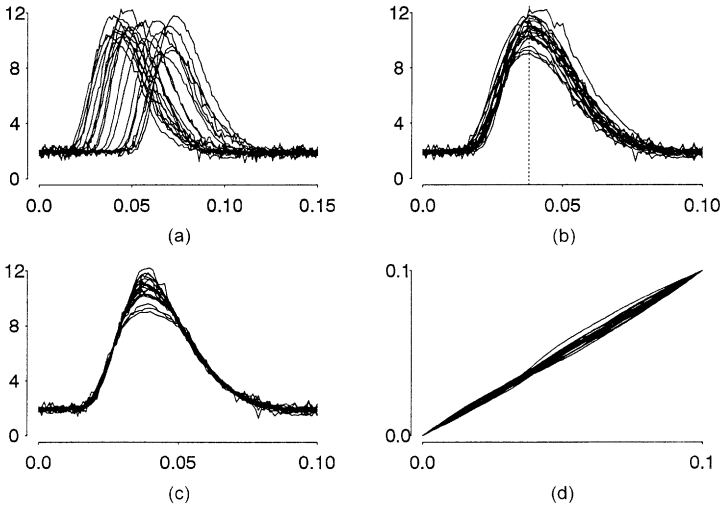

Fig. 3. (a) 20 records of force (newtons) exerted by the thumb and forefinger with a maintained background force of 2 N (the starting time (seconds) of each record is arbitrary); (b) these records with the times of maximum force (the vertical dotted line) being aligned; (c) completely registered force functions; (d) the time-warping functions that register them

Put more abstractly, $x_i(t_j)$, the values of two or more functions, may differ because of two types of variation. The first is the more familiar *range variation* or vertical variation due to the fact that two functions x_1 and x_2 may simply differ at points of time at which they can be compared. But they may also exhibit *domain variation* in that x_1 and x_2 should not be compared at a fixed time t, but at times t_1 and t_2 at which the two cases are essentially in comparable states. For example, the intensity of the pubertal growth spurts of two children should be compared at their respective ages of peak velocity defined by $D^2x_1(t_1) = D^2x_2(t_2) = 0$, rather than at any fixed age.

The problem of transforming the arguments of curves to align various salient features is described in a very large literature in many different fields. The problem is referred to in this paper and by Silverman (1995) as *curve registration*, the engineering literature tends to the evocative term *time warping* (Sakoe and Chiba, 1978; Wang and Gasser, 1995) and the process of registering curves for computing average curves is called *structural averaging* by Kneip and Gasser (1988, 1992). Registering outcomes over surfaces and volumes is especially important in medical imaging (Bookstein, 1991).

1.1. Formulation of the curve registration problem

The curve registration problem can be expressed formally as follows. Let N functions x_i be defined on closed real intervals that can be taken without loss of generality as $[0, T_i]$. These functions may be vector valued, as would be the case, for example, if they indicated positions in two- or three-dimensional space or simultaneous growth in several aspects of the skeleton. The upper boundaries may either vary randomly or be fixed. In practice the boundaries of the interval are usually defined by marker events such as birth and a fixed adult age for the growth data, or by arbitrary values such as midnight on December 31st for the weather data. Or it may be that the interval is simply sufficiently large to include all of the curves of interest plus some tail behaviour of little concern. In the event that the functions are periodic with

known period, it will be assumed that each x_i is extended beyond $[0, T_i]$ if there is a need to use information beyond the interval. Thus, for periodic data we can also permit the study of the sampled functions over the intervals $[0, T_i] + \delta$ for any δ.

Let $h_i(t)$ be a transformation of time t for case i with domain either $[0, T_0]$ for non-periodic data or $[0, T_0] + \delta$ for periodic data. The fact that the timings of events remain in the same order regardless of the timescale entails that h_i, the time-warping function, should be strictly increasing, i.e. $h_i(t_1) > h_i(t_2)$ for $t_1 > t_2$. That is, the function h_i is invertible so that for the same event the time points on two different timescales correspond to each other uniquely. Let y be a fixed function defined over $[0, T_0]$ to provide a sort of template for the individual curves x_i in the sense that after registration the features of x_i will be aligned in some sense to those of y. We can propose, for example, the model

$$y(t) = x_i\{h_i(t)\} + \epsilon_i(t) \tag{1}$$

or

$$y = x_i \circ h_i + \epsilon_i,$$

where ϵ is small relative to x_i and roughly centred on 0. If, alternatively, the template y is defined by discrete values y_j, $j = 1, \ldots, n$, then our model becomes

$$y_j = x_i\{h_i(t_j)\} + \epsilon_{ij}. \tag{2}$$

The registration task is to estimate the time-warping functions h_i so that the de-warped components x_i can be studied separately, along with possible analyses of the functions h_i as well.

This problem can be seen to complement the usual nonparametric regression problem, where h is the identity function, i.e. $h(t) = t$, and x is to be estimated. Alternatively, if we suppose that $x \circ h = h$, the registration task becomes a monotone nonparametric regression problem.

1.2. Marker registration and fitting criteria

Marker registration is often used in engineering, biology, physiology and other fields. It is the process of aligning curves by identifying the timing of certain salient features in the curves, of which the zero of acceleration during the pubertal growth spurt and optimal temperature timings are examples. Using this strategy, curves are aligned by transforming time so that marker events occur at the same values of the transformed times. Comparisons between marker timings can also be made by using corresponding transformed times. Sakoe and Chiba (1978) estimated the values of h at marker timings by minimizing the sum of weighted distances of two speech patterns at the marker timings and imposing monotonicity and continuity on h. They solved for the discrete values of h by using a dynamic programming algorithm. Kneip and Gasser (1988, 1992) described marker registration in detail from a statistical perspective. However, marker registration can present some problems: marker events may be missing from certain curves, and marker timing estimates can often be difficult to obtain. These issues are discussed for human growth curves by Ramsay, Bock and Gasser (1995).

As an alternative to marker registration, Silverman (1995) developed a technique for curve registration that does not require explication of markers. He optimized a global fitting criterion with respect to a restricted parametric family of transformations of time shifts and applied this approach to estimating a shift in time for each of the temperature functions in

35 Canadian weather-stations. He also incorporated this shift into a principal components analysis of the variation among curves, thus explicitly partitioning variation into range and domain components.

This paper goes beyond Silverman's method by using an arbitrarily flexible yet computationally convenient smooth monotone transformation family developed by Ramsay (1998) and thus presents a nonparametric curve registration approach. The smoothness of the transformation is controlled by a penalty term. This approach can be applied to a broad range of applications: not only can it be applied to the case that curves differ from each other in the time domain by a constant time shift, but also to the case that curves differ from each other in the time domain by a variable time shift and scaling factor.

2. Smooth monotone transformations

Suppose that a function h has an integrable second derivative in addition to being strictly increasing. Then every such function can be described by the homogeneous linear differential equation

$$D^2h = w\,Dh \tag{3}$$

because a strictly monotone function has a non-zero derivative, and hence the weight function w is simply D^2h/Dh, or the *relative curvature* of h. This equation, subject to the requirement that $h(0) = 0$ and $h(T_0) = T_i$, has the solution

$$h(t) = C_1\{D^{-1}\exp(D^{-1}w)\}(t) = C_1(M\,D^{-1}w)(t) \tag{4}$$

where D^{-1} is the partial integration operator and $C_1 = T_i/D^{-1}\exp\{D^{-1}w(T_0)\}$. The integration–rectification operator, $M = D^{-1}\exp$, which in this case maps a differentiable integrable function $D^{-1}w$ into a twice-differentiable monotone function, may be called the *monotonization operator*. When w is constant, $h(t) = (C_1/w)\exp(wt)$, so that an exponential function has constant relative curvature. A straight line is implied by $w = 0$.

The relative curvature w can also be seen as the rate of the local percentage change in Dh. The Taylor expansion of Dh at t_0 yields

$$Dh(t) \approx Dh(t_0)\{1 + w(t_0)(t - t_0)\}.$$

Thus $w(t_0)$ is approximately the proportional change in Dh per unit time at $t = t_0$.

Just like using log- or exp-functions to eliminate the need for imposing positivity in many situations, using this monotone family eliminates the need for imposing monotonicity on the time transformation functions h by allowing us to estimate the unconstrained function w.

2.1. The stochastic time model

The smooth monotone family is also useful in introducing the concept of *stochastic time*. A random stochastic time process can provide a statistical model for some types of registration problem and permits the simulation of random warping functions.

The following stochastic differential equation captures the concept of the time continuum having a stochastic character and thus varying in some systematic way from record to record:

$$\mathrm{d}\{Dh(t)\} = Dh(t)\{w(t)\,\mathrm{d}t + \mathrm{d}z(t)\} \tag{5}$$

where z is a stochastic process, w is a fixed relative acceleration function and $\mathrm{d}z(t)$ is the deviation from this function at time t. As a specific example, let $z = B$ be Brownian motion

with parameters 0 and σ^2, so that dB represents Gaussian white noise with mean 0 and variance σ^2. The solution to equation (5) in the Ito sense is then

$$h(t) = C_0 + C_1 D^{-1} \exp\{D^{-1}w(t) + B(t) - \sigma^2 t/2\} = C_0 + C_1 M(w + dB)(t) \quad (6)$$

where C_0 and C_1 are constants (Øksendal, 1995). The additional term $\sigma^2 t/2$ arises because B is not differentiable, and equation (5) cannot be treated as a differential equation in the conventional sense.

The stochastic function $\exp(B)$ is often called *geometric Brownian motion* (Øksendal, 1995). Since Brownian motion is an independent additive increments process, its exponential is an independent multiplicative factor process. The observed rate function Dh is the deterministic function $\exp\{D^{-1}w(t) - \sigma^2 t/2\}$ perturbed multiplicatively by $\exp(B)$. This can be envisaged as a clock that is running fast or slow from instant to instant, constantly undergoing a percentage change in rate in a memoryless chaotic manner.

If there is no drift ($w = 0$), under appropriate initial conditions we have $E(h) = t$. (See Appendix A for details.) This tells us that the central location for these individual warped times coincides with the diagonal line $h(t) = t$, i.e. the solution to $D^2h/Dh = 0$.

The stochastic time model is of interest in modelling why curves need alignment in situations where the terminating event occurs at a random time T_i. Many situations involve fixed termination times, often as a consequence of a preliminary normalization of the time interval, and thus the constraint $h_i(T) = T$ holds. For these cases, a Brownian bridge model would be more appropriate, but we shall not pursue this topic any further.

3. Estimation of warping function *h*

3.1. Estimation of h for a fixed target y

Let y be a fixed function in the same class as the sample functions x_i. Dropping the subscript for the moment, consider the problem of estimating the time-warping function h that minimizes a measure of the fit F_λ of $x \circ h$ to target function y.

In this paper we minimize the penalized squared error criterion

$$F_\lambda(y, x|h) = \int \|y(t) - x\{h(t)\}\|^2 \, dt + \lambda \int w^2(t) \, dt, \quad (7)$$

for h in the smooth monotone family defined in equation (4). Thus h is estimated by estimating its relative curvature w. If y is observed discretely, $\int \|\cdot\|^2 \, dt$ is replaced by a sum of squared errors. For a given value of λ, if Dh is close to 0 or equivalently if h is very flat, a heavy penalty is put on D^2h via $w = D^2h/Dh$ to ensure that h is not so wiggly as to deviate from monotonicity; in contrast, if Dh is large, this same λ effectively gives a light penalty on D^2h, paying less attention to the curvature of the transformation function h. Though penalizing D^2h ensures smoothness in h, it does not ensure monotonicity, whereas penalizing w yields both smoothness and monotonicity.

Larger values of smoothing parameter λ shrink the relative curvature $w = D^2h/Dh$ to 0, and therefore shrink $h(t)$ to t. Moreover, since the relative curvature measure w is scale free, appropriate values of λ tend not to vary much from one application to another. We find, for example, that λ-values of 10^{-4}, 10^{-3} and 10^{-2} have worked well over a range of applications.

In the analyses reported in this paper, the function w is represented by a linear combination of B-spline bases

$$w(x) = \sum_{k=0}^{K} c_k B_k(x). \tag{8}$$

The B-spline bases are of a specified order and defined by a breakpoint sequence ξ_l, $l = 1, \ldots, L$. The definition (4) of h involves two partial integrals, and although the use of quadrature schemes even as simple as the trapezoidal rule is quite practical for computational purposes it would be desirable in many problems to have an explicit expression for h. Accordingly Ramsay (1998) used order 1 B-spline bases for w, since these permit the expression of h in a closed form.

3.2. The Procrustes fitting criterion

The Procrustes fitting process, used in many multivariate data analysis problems, involves the alternation between using the data to define a target for defining a particular transformation of each observation and estimating the transformations themselves. In the applications, the cross-sectional average $\bar{x}^{(0)}(t)$ of the unregistered curves is used as the initial target y for the estimation of each sample warping function $h_i^{(1)}$. If the curves have obvious landmarks, they may also be aligned before computing $\bar{x}^{(0)}(t)$.

Once these warping functions have been estimated, an updated cross-sectional average

$$y(t) = \bar{x}^{(1)}(t) = N^{-1} \sum_i x_i\{h_i^{(1)}(t)\}$$

can be computed and used as a target for computing revised warping functions. Our experience indicates, however, that there is seldom any need for this revision, since the change in the h_i from the first to the second iteration tends to be negligible.

3.3. Extension of the fitting criterion

More generally, instead of using criterion (7) in the estimation of the warping function h_i, we may want to minimize

$$F_\lambda(y, x|h) = \sum_{j=0}^{m} \int \alpha_j(t) \|D^j y(t) - D^j x\{h(t)\}\|_j^2 \, \mathrm{d}t + \lambda \int w^2(t) \, \mathrm{d}t \tag{9}$$

where $\alpha_j(t)$ are weight functions,

$$\|D^j y(t) - D^j x\{h(t)\}\|_j^2 = (D^j y(t) - D^j x\{h(t)\})' \mathbf{W}_j (D^j y(t) - D^j x\{h(t)\}) \tag{10}$$

and the $\mathbf{W}_j$ are weight matrices. This loss function incorporates several potentially useful aspects. It is possible, for example, that curve registration should take place at the level of some derivative $D^j y$ and $D^j x \circ h$ rather than at the level of the functions themselves. For example registering the acceleration functions for the growth curves has turned out to be more illuminating. More generally, it may be profitable to use a mixture of derivatives defined by the α_j if we are interested in function behaviour at several levels. The weight matrices $\mathbf{W}_j$ also allow for more general weightings of the elements of the functions when they are vector valued. Finally, the weight functions α_j also permit unequal weighting of fit to the target over time. It may happen, for example, that we are primarily interested in registering the curves over some central portion of the interval.

The derivative of F_λ with respect to coefficient vector $\mathbf{c}$ (as in equation (8)) is

$$\frac{\partial F_\lambda(y, x|h)}{\partial \mathbf{c}} = -2\sum_{j=0}^{m} \int \alpha_j(t)\, \frac{\partial h(t)}{\partial \mathbf{c}} \left(\frac{\partial D^j x(h)}{\partial h}\right)' \mathbf{W}_j(D^j y(t) - D^j x\{h(t)\})\, \mathrm{d}t$$

$$+ \lambda \int \left\{\frac{\partial w(t)}{\partial \mathbf{c}}\right\}^2 \mathrm{d}t. \tag{11}$$

The derivative $\partial D^j x(h)/\partial h$ must be estimated with care. It must match $D^j x$ in the sense that, if $D^j x$ has been estimated by a smoothing technique, then $\partial D^j x(h)/\partial h$ must be the derivative of the smoother, evaluated at the values of function h. This can be achieved by using higher order polynomial or spline smoothing techniques. An S-PLUS module `Pspline` for function and derivative estimation by spline smoothing penalizing the norm of the derivative of order m by the first author is available by anonymous file transfer protocol at `statlib.stat.cmu.edu` or from the World Wide Web at `www.stat.cmu.edu`. Our experience indicates that if a derivative of order j is required the penalty should use the derivative of order $m = j + 2$.

4. Illustrations and applications

4.1. A test problem

The first illustration of the technique illustrates the capacity of the curve registration technique to recapture known warping functions h_i given a fixed target function y and a set of functions $y \circ h_i^{-1}$, where $(h_i^{-1} \circ h_i)(t) = t$. The target was

$$y(t) = \sin(t^2/\pi), \qquad t = 0, 2\pi/100, 4\pi/100, \ldots, 2\pi, \tag{12}$$

and each inverse warping function h_i^{-1} was an approximate realization of the stochastic time function (6), with $w = 0$ and $\sigma = 4$. These inverse warping functions were achieved for each curve by generating 1001 values from $N(0, 16)$, and using the trapezoidal rule to approximate the partial integrals at the 101 argument values. Fig. 4(a) displays 20 sample curves $y \circ h_i^{-1}$, along with the target function y, and Fig. 4(b) shows the corresponding 20 inverse warping functions h_i^{-1}. Although the curvature in the warping functions is mostly fairly mild, the consequences for the $y \circ h_i^{-1}$ can be rather severe.

Each of these sample functions was registered by estimating the $\hat{h}_i$ that minimizes criterion (7). These estimates were in class (8) where the B-splines were of order 1 with six break point values $\xi_k = 0, 2\pi/5, 4\pi/5, 6\pi/5, 8\pi/5, 2\pi$; thus each estimate was defined by five parameters c_k. The smoothing parameter used was $\lambda = 0.0001$. Fig. 5(a) shows the 20 registered functions, and Fig. 5(b) indicates the quality of the recovery of the true warping functions h_i by the estimates $\hat{h}_i$ for three sample curves.

4.2. A cautionary example

The next example is intended to illustrate that too much flexibility in h can lead to serious distortion of a registered curve if $\epsilon(t)$ in model (1) is not centred approximately at 0. The single function to be aligned was simply twice the target y used in the previous example, so no alignment is really required. Fig. 6 shows what happens when 21 equally spaced break point values were used and $\lambda = 0.001$. The technique attempts to minimize the discrepancy between the target and the curve by compressing time ($Dh(t) < 1$) in regions where both curves are either increasing or decreasing and expanding time ($Dh(t) > 1$) at the peaks. It is essential to

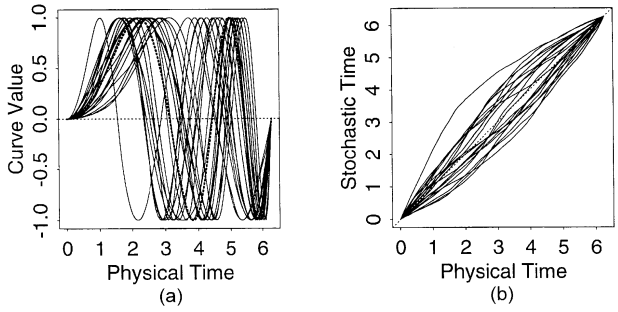

Fig. 4. (a) 20 sample curves, each produced by applying the target function indicated by the dotted curve to the values of a random monotone function h_i^{-1}; (b) these 20 monotone functions

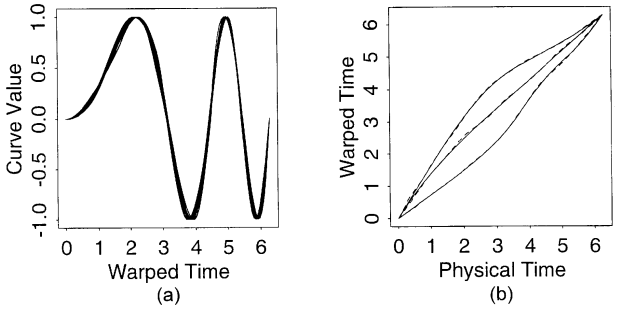

Fig. 5. (a) 20 registered sample curves; (b) three estimated warping functions (———) and the corresponding true functions used to generate the sample curves (- - - -)

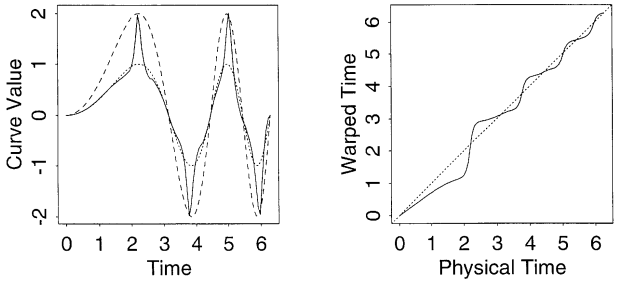

Fig. 6. The full curve is a consequence of registering the broken curve to the target indicated by the dotted line by using 21 break points and $\lambda = 0.001$: the warping function must be more heavily regularized by increasing λ or by reducing the number of break points.

impose more regularity on w to avoid these effects. The distortion was minimal for this example for $\lambda = 10$, or with $\lambda = 0.001$ and only 11 break points.

4.3. Registering the pinch force records

Each pinch force record shown in Fig. 3(a) contains enough of the background 2-N force record to contain all of the build-up and release of force. Our first step was a simple marker

registration, using the time of maximum force as the marker. The initial portion of each record was clipped so that the time of maximum force occurred at that of the earliest recorded maximum, and the final portion of each record was then clipped so that each record contained 0.1 s or 151 force values. The warping function h in the first step is, then, $h_i(t) = t - \delta_i$. Fig. 3(b) displays the result.

This simple procedure failed to align the curves completely, in part because the noise in the data meant that the time of maximum force was also rather noisy, and in part because the shapes of the curves varied to some extent. In the next step we applied the Procrustes registration process by computing the cross-sectional mean of the marker-aligned force records. This provided the target values y_j in model (2), and these still contained a certain amount of noise.

Because the minimization of the fitting criterion (7) requires that the functions x_i and their derivatives Dx_i at any value of t, our next step was to use polynomial spline smoothing to estimate these functions. We used the `Pspline` module described earlier, penalizing the third derivative so that the first-derivative estimate would be smooth. Our final choice of smoothing parameter was slightly higher than that indicated by the minimum generalized cross-validation criterion to provide reasonable smoothness in each Dx_i. The spline smoothing process assures that Dx_i is the actual derivative of x_i.

Function w was expanded in terms of order 1 B-splines defined by the break value sequence 0, 0.01, 0.02, . . ., 0.1, and we liked the results that we obtained with smoothing parameter $\lambda = 1.0$. Fig. 3(c) shows that a substantial improvement in registration was achieved over marker alignment. The warping functions h_i are given in Fig. 3(d). The Newton–Raphson iterations used to minimize the fitting criterion with respect to the coefficients defining w_i all converged in four iterations or fewer. We judged that carrying out another iteration of the Procustes procedure by re-registering with respect to the mean of these registered curves did not result in any interesting change in the estimated mean.

4.4. Registering the height acceleration functions

The 10 acceleration functions in Fig. 1 were registered by using the cross-sectional mean shown in Fig. 1 as a target. The break values ξ_k defining the order 1 B-splines were 4, 7, 10, 12, 14, 16 and 18 years, and the curves were registered over the interval [4, 18] using criterion (7) with $\lambda = 0.01$. Two Procrustes iterations produced the results displayed in Fig. 7. Fig. 7(a) displays the 10 warping functions h_i, and Fig. 7(b) shows the curves $x_i \circ h_i$. Fig. 1 compares the unregistered and registered cross-sectional means. We see that the differences are substantial, and moreover that the mean of the registered functions tends to resemble much more closely most of the sample curves displayed in Fig. 1.

5. Discussion

The main objective of this paper was to display the capacity of the smooth monotone function family (4) to render curves similar in shape by a non-linear transformation or warping of the argument. Our experience with functional data suggests that this step, although often overlooked, ought to be a routine part of a functional data analysis that combines information across curves. Ramsay and Silverman (1997) offer further discussion and illustration of this problem.

Model (1) and criterion (7) which goes naturally with it have their limitations, however, as

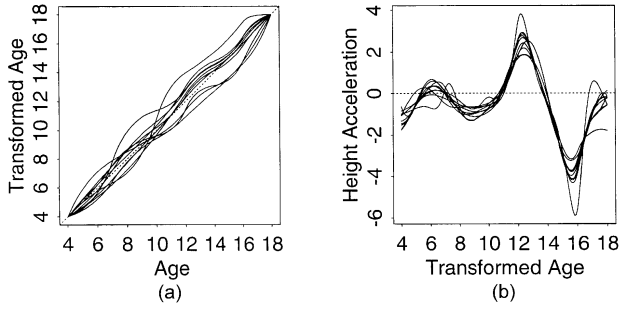

Fig. 7. (a) Estimated time-warping functions h_i for the 10 height acceleration curves in Fig. 1; (b) the registered curves

the cautionary example above indicates. Kneip and Gasser (1992) considered the more general model

$$y(t) = a_i(t)\, x_i\{h_i(t)\} + \epsilon_i(t) \tag{13}$$

in which the amplitude modulation function a_i allows for scale variation among curves, such as seen in Fig. 6 and to some extent in Fig. 3. We are investigating alternative fitting criteria that will work better than equation (7) for this situation.

It must be recognized that the curve registration problem tends to be underdetermined, especially when the smooth functions x_i are themselves to be estimated to some extent from noisy data. The underdetermination is especially evident in formulation (13), where there is no hope of deriving stable estimates of both a_i and h_i unless it can be assumed that their variation is small relative to that of x_i. Even then, there are situations, such as $x_i \circ h_i \approx h_i$, where there will be a trade-off among the shapes of a_i, x_i and h_i that cannot be resolved by the data. Because of this, the continuous control over the departure of h_i from linearity offered by the penalization of $\|w^2\|$ is an essential feature of a practical registration technique in our view. The same applies to the departure of a_i from constancy, suggesting the differential equation formulation for a_i

$$Da_i = v_i a_i.$$

The problem of choosing the smoothing parameter λ is on the one hand a little easier here because of the invariance of w with respect to changes of scale in t. But on the other hand the tendency to underidentification means that one should keep w as small as possible consistent with obtaining a reasonable degree of registration. We have not found any way of making this decision automatically, and we are not even sure that it would be desirable to do so, since the amount of registration required for a problem can depend on various other considerations such as what subsequent analyses are involved.

Finally, the registration of two- and three-dimensional images, where the argument t is multidimensional, is a much more difficult problem, and one in great need of a useful solution, especially in fields such as neuroimaging. It is our hope that there will be ways of generalizing these results that will prove profitable in this wider domain.

Acknowledgement

This research was supported by grant A320 from the Natural Sciences and Engineering Research Council of Canada and was completed while the second author was a Postdoctoral

Fellow at the Centre de Recherches Mathématiques and the Institut des Sciences Mathématiques, Université de Montréal, Montréal, Canada

Appendix A: Moment calculation of the stochastic time model

Let $X_1 = h$, $X_2 = Dh$ and $\mathbf{X} = (X_1, X_2)'$. Then the second-order stochastic differential equation (5) can be written as a system of two first-order equations:

$$\begin{aligned} \mathrm{d}X_1(t) &= X_2(t)\,\mathrm{d}t; \\ \mathrm{d}X_2(t) &= X_2(t)\{w(t)\,\mathrm{d}t + \mathrm{d}z(t)\}. \end{aligned} \tag{14}$$

Let $p(\mathbf{x}, t|\mathbf{x}_0)$ be the conditional density function, given the initial condition $\mathbf{x}(0) = \mathbf{x}_0$. This is an Ito-type equation and consequently satisfies the Fokker–Planck diffusion equation

$$Dp(\mathbf{x}, t|\mathbf{x}_0, 0) = -x_2\,\frac{\partial p}{\partial x_1} - w(t)\frac{\partial (x_2 p)}{\partial x_2} + \frac{\sigma^2}{2}\frac{\partial^2 (x_2^2 p)}{\partial x_2^2}, \tag{15}$$

where Dp is the time derivative $\partial p/\partial t$, and the initial distribution is $p(\mathbf{x}, 0) = 1$ for $\mathbf{x} = \mathbf{x}_0$ and $p(\mathbf{x}, 0) = 0$ otherwise. The analytical solution of this equation is not obvious, and its numerical solution can also involve considerable difficulty. Fortunately, however, the moment functions of $\mathbf{X}$ can be calculated relatively easily to yield some information about the process. The moment functions

$$m_{kl}(t) = E\{X_1^k(t) X_2^l(t)\}$$

satisfy

$$Dm_{kl}(t) = k\, m_{k-1,l+1}(t) + l\left\{w(t) + \frac{\sigma^2}{2}(l-1)\right\} m_{kl}(t). \tag{16}$$

Of particular interest to us are the moment functions $E\{h^k(t)\} = E\{X_1^k(t)\} = m_{k0}(t)$. However, by equation (16), $Dm_{k0}(t) = k\, m_{k-1,1}(t)$, so lower order moment functions must first be calculated. Let

$$\begin{aligned} C_{ij} &= E\{X_1^i(0) X_2^j(0)\}, \\ q_l(t) &= \exp\{l\, D^{-1} w(t) + l(l-1)\sigma^2 t/2\} \end{aligned} \tag{17}$$

with $q_0(t) = 1$. If $m_{k-1,l+1}(t)$ is known, by equation (16) we have

$$m_{kl}(t) = q_l(t)\left\{k\, D^{-1}\left(\frac{m_{k-1,l+1}}{q_l}\right)(t) + C_{kl}\right\}, \tag{18}$$

so the marginal moment function for the rate Dh is

$$m_{0l}(t) = C_{0l}\, q_l(t) = C_{0l}\, \exp\left\{l\, D^{-1} w(t) + l(l-1)\frac{\sigma^2 t}{2}\right\}. \tag{19}$$

Note that the moment function $m_{0l}(t)$ of the stochastic process X_2 (which is described by $\mathrm{d}X_2(t) = X_2(t)\{w(t)\,\mathrm{d}t + \mathrm{d}z(t)\}$) is the product of two independent parts: its moment at the initial state C_{0l} and a time-dependent factor. By equations (18) and (19),

$$m_{1l}(t) = q_l(t)\left\{C_{0,l+1} D^{-1}\left(\frac{q_{l+1}}{q_l}\right)(t) + C_{1l}\right\}, \tag{20}$$

and therefore the first two moments of h are

$$\begin{aligned} E(h) &= m_{10}(t) = C_{01}\, Mw(t) + C_{10}, \\ E(h^2) &= m_{20}(t) = 2\, D^{-1} m_{11}(t) + C_{20}, \end{aligned} \tag{21}$$

where

$$m_{11}(t) = \exp\{D^{-1}w(t)\}\{C_{02}\, M(w + \sigma^2)(t) + C_{11}\}.$$

We are often particularly interested in a process with the initial condition $\mathbf{X}_0 = (0, 1)$, in which case

$$\begin{aligned} C_{0j} &= E\{X_2(0)^j\} = 1, \\ C_{ij} &= 0, \qquad i > 0. \end{aligned}$$

If there is no drift ($w = 0$), we have

$$\begin{aligned} m_{01}(t) &= 1, \\ m_{02}(t) &= \exp(\sigma^2 t). \end{aligned} \tag{22}$$

This tells us that the exponential of a Brownian motion, i.e. the process X_2 defined by $\mathrm{d}X_2(t) = X_2(t)\,\mathrm{d}z(t)$, has a constant mean and an exponentially growing variance $\exp(\sigma^2 t) - 1$. Furthermore,

$$\begin{aligned} E(h) &= t, \\ \mathrm{var}(h) &= \frac{2}{\sigma^4}\exp(\sigma^2 t) - \frac{2t}{\sigma^2} - \frac{2}{\sigma^4} - t^2. \end{aligned} \tag{23}$$

We see that the central location of the random time h coincides with the optimal location when there is no drift, whereas its standard deviation increases exponentially. These equations can be used to compute simple moment estimates of σ^2.

References

Bookstein, F. L. (1991) *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge: Cambridge University Press.

Kneip, A. and Gasser, T. (1988) Convergence and consistency results for self-modeling nonlinear regression. *Ann. Statist.*, **16**, 82–112.

———(1992) Statistical tools to analyze data representing a sample of curves. *Ann. Statist.*, **20**, 1266–1305.

Øksendal, B. (1995) *Stochastic Differential Equations: an Introduction with Applications*. New York: Springer.

Ramsay, J. O. (1998) Estimating smooth monotone functions. *J. R. Statist. Soc.* B, **60**, 365–375.

Ramsay, J. O., Bock, R. D. and Gasser, T. (1995) Comparison of height acceleration curves in the Fels, Zurich, and Berkeley growth data. *Ann. Hum. Biol.*, **22**, 413–426.

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

Ramsay, J. O., Wang, X. and Flanagan, R. (1995) A functional data analysis of the pinch force of human fingers. *Appl. Statist.*, **44**, 17–30.

Sakoe, H. and Chiba, S. (1978) Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Spch Signal Process.*, **26**, 43–49.

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

Wang, K. and Gasser, T. (1995) Alignment of curves by dynamic time warping. Unpublished.