7

Time Warping Handwriting and Weather Records

7.1 Introduction

In Chapter 6 we encountered what is almost always a fact of life in functional data. Curves vary in two ways: vertically, so that certain oscillations and levels are larger in some curves than others; and horizontally, so that the timings or locations of prominent features in curves vary from curve to curve. We call these two types of variation *amplitude* and *phase*, respectively. You might want to glance back at Figure 6.9 to see a schematic diagram illustrating this concept.

We now look more closely at amplitude and phase variation in the context of two rather different sets of data. The first is a sample of the printing (by hand) of the characters "fda." Each observation is a series of strokes separated by gaps where the pen is lifted off the paper, along with the clock times associated with these events. The timing of strokes and cusps varies from sample to sample, and we consider how to register these curves by transforming time so that, as nearly as possible, each stroke occurs at the same time for all curves. The aim of registration is to yield a sample of curves that vary only in terms of amplitude. The phase variation does not disappear, though; it is captured in the time transformations that we estimate for each curve.

Our second example is a single long time series, daily temperature measurements for the 34 years spanning 1961 through 1994. Naturally these data have a strong annual pattern, but one has only to appeal to personal experience to know that winter, for example, arrives late in some years

and early in others. Therefore, we want to speed up and slow down time within each year so that the seasons will change at the same time across all years. We do this for many reasons, among them to get a better estimate of the average annual temperature curve, and to get tighter estimates of long-term trends such as might be associated with global warming.

We reserve the discussion of the more technical aspects of just how registration is achieved to Section 7.6, but it will first be helpful to spell out more formally a model for how curves vary.

7.2 Formulating the registration problem

Curve registration can be expressed formally as follows. We have a sample of N functions x_i. Each curve is defined over an interval, and the length of the interval may vary from curve to curve. For simplicity, let us assume a common origin but a variable end point, and make the intervals $[0, T_i]$.

A basic form of registration is to preprocess each curve by rescaling its argument range to a common standard interval $[0, T_0]$. This standard time interval $[0, T_0]$ may, for example, be the average interval $[0, \bar{T}]$. Although we assume the existence of a standard interval, we do not require that the data have necessarily been scaled to fit this interval.

Now let $h_i(t)$ be a transformation of time t for curve i, which we call a *time warping function*. The argument t varies over $[0, T_0]$. The values of $h_i(t)$, however, range over the curve i's interval $[0, T_i]$, and satisfy the constraint $h_i(0) = 0$ and $h_i(T_0) = T_i$. Thus the time warping function maps the standard interval $[0, T_0]$ to the interval on which the function x_i lives.

The fact that the timings of events retain the same order regardless of the time scale implies that the time warping function h_i should be strictly increasing, so that $h_i(t_1) > h_i(t_2)$ if and only if $t_1 > t_2$. In fact, $h_i(t)$ is just a growth curve of the kind that we studied in Chapter 6. We can think of clock time t as growing linearly with a constant velocity of one second per second. We can think of curve i's "system time" as evolving at a rate that can change slightly from one clock unit to another. We show that printing is running ahead of itself at some times, and late at others; winter comes early some years, and late at others.

This strict monotonicity condition ensures that the function h_i is *invertible*, so that for each y in the interval $[0, T_i]$ there is a unique t for which $h(t) = y$. We use the notation h_i^{-1} to denote the inverse function,[1] for which $h_i^{-1}(y) = h_i^{-1}[h_i(t)] = t$. The invertibility of h_i means that it defines a one-to-one correspondence between the time points on the two different time scales.

[1]Not to be confused with the reciprocal of h, a concept which we do not use in this discussion.

Let $x_0(t)$ be a fixed function defined over $[0, T_0]$ that provides a 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 x_0. The following is a model for two functions $x_0(t)$ and $x_i(t)$ differing primarily in terms of phase variation,

$$x_i[h_i(t)] = x_0(t) + \epsilon_i(t) \ , \tag{7.1}$$

where the residual or disturbance function ϵ is small relative to x_i and roughly centered about 0. Because we assume that ϵ is small relative to x_i, this model postulates that major differences in shape between target function x_0 and specific function x_i are due only to phase variation. Having identified the N warping functions $h_i(t)$, we can then calculate the *registered functions* $x_i[h_i(t)]$. Methods for fitting the model (7.1) are developed later in this chapter.

What does $h(t)$ mean? Let's assume that the ice breaks up on the St. Lawrence River at Montreal on the average on April 7th, day 97 for nonleap years. But in 1975 spring is late and the ice goes out on April 14th, or day 104. We want, therefore, that $h_{1975}(97) = 104$, so that $x_{1975}[h_{1975}(97)] = x_{1975}(104)$, and therefore that, from a clock perspective, the ice is breaking up simultaneously in both the standard year and in 1975 when time is running a week late. In effect, in this case, the warping function speeds up time to compensate for its being tardy in 1975.

On the other hand, imagine that in the same year the leaves on Mont Royal in the city change color on September 15th (day 258) instead of September 30 (day 303) as is normal. Then $h_{1975}(303) = 258$, and the warping function is slowing down system time at a point when it is running ahead to conform to clock time. Thus, $h(t) > t$ corresponds to a process running slow, and $h(t) < t$ to one running fast.

In most of the examples we consider, the target function x_0 is not given. Instead we have to construct it from the data. Typically, we begin by mapping each interval linearly to the standard interval $[0, T_0]$, and set x_0 initially to be the sample mean $\overline{x}$ of the functions x_i after this scaling. We then register the individual functions to $\overline{x}$, and update the estimate of x_0 to be the mean of the *registered* functions. We now update the warping functions by registering the individual functions to this new estimate of x_0. In principle, it is possible to iterate the process of updating x_0 then reestimating the warping functions, but this is rarely necessary in practice.

The functions $x_i(t)$ that we are discussing here may be derivatives as, for example, the velocity curves in Chapter 6. It can be better to register derivatives instead of the original functions because derivatives tend to oscillate more, and therefore have more distinctive features to align. In addition, in phenomena such as human growth, features in the derivative are the true aspects of interest in the problem.

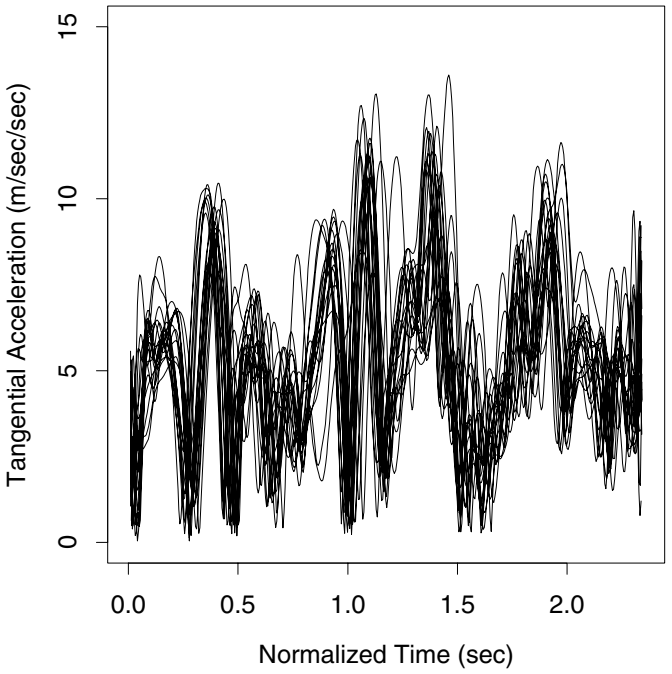

Figure 7.1. The tangential acceleration (7.2) on the X–Y plane for 20 samples of the printing of the characters "fda" by a single individual.

7.3 Registering the printing data

These data are recordings of the X-, Y-, and Z-coordinates 200 times per second of the tip of the pen during the printing by hand of the characters "fda." In the experiment, there were a number of subjects, and each repeated the printing 20 times. Because this is printing instead of cursive writing, the vertical Z-coordinate is important.

The registration problem is illustrated by plotting the magnitude of the *tangential acceleration* vector,

$$TA(t) = [X''(t) + Y''(t)]^{1/2} \tag{7.2}$$

on the X–Y plane for each curve for one of our subjects. Tangential acceleration is an important property in the study of the dynamics of printing. To simplify the plot, the time taken to draw each record in Figure 7.1 was first normalized to the average time, 2.3 seconds. We see that the timings of the acceleration peaks vary noticeably from replication to replication.

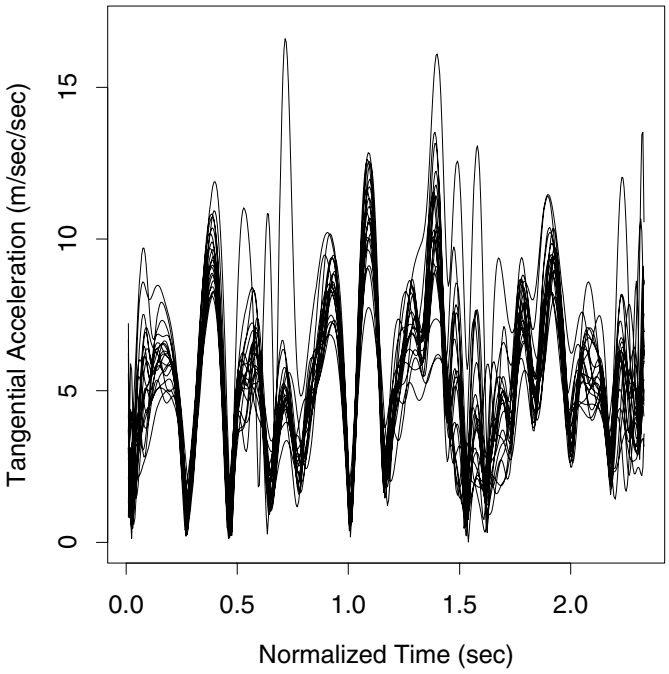

Figure 7.2. The tangential acceleration curves for the registered printing samples.

The registered results are shown in Figure 7.2, and we see that the acceleration peaks are now much more cleanly aligned. Moreover, when we look at the mean tangential acceleration calculated before and after registration, as shown in Figure 7.3, we see that the registration has also improved the amount of detail in the mean function. The peaks are higher, more sharply defined, the valleys are closer to zero, and some small peaks emerge that were washed out in the unregistered mean function.

We return to these data in Chapter 11, where we consider whether we can identify someone by using a differential equation that describes that person's printing.

7.4 Registering the weather data

Functional data often come to us as a single long time series spanning many days, months, years, or other time units. The variation in data such as these is usually multilevel in nature. There is usually a clear annual, diurnal, or other cycle over the basic time unit called the *season* of the data, combined

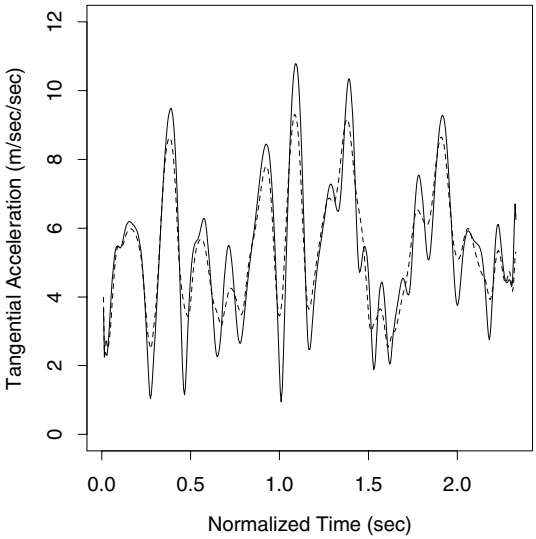

Figure 7.3. The mean tangential acceleration curve for the registered printing samples is plotted as a solid line, and the mean for the unregistered data as a dashed line.

with longer-term trends that span many time units. Moreover, the seasonal cycle may also show some evolution over the time spanned by the series.

The data in this example are 12,410 daily temperatures at Montreal over the 34 years from 1961 to 1994 (in leap years temperatures for February 28 and 29 were averaged). Because these are 24-hour averages, the actual daily lows and highs were more extreme. The minimum and maximum temperatures recorded in this period were $-30°$C and $30°$C, respectively. All our analyses are conducted on the entire series, but we do not *plot* the results for the entire time interval, since this is too much detail to put in a graph. Figure 7.4 focuses on 1989, when a severe cold snap came at Christmas, and was followed by a strong thaw.

We now smooth the temperatures in two ways. We smooth merely to remove the day-to-day variation, which from our perspective is too short-range to be interesting, although we are reluctant to call it error or noise. When we are done, we are left with an estimate of the smooth part of temperature variation. We achieved this by using 500 B-spline basis functions of order 6. The knots were equally spaced, and occurred at about every 25 days. This smooth, which we denote $x(t)$, is shown in Figure 7.4 as a solid line.

The second smooth $x_0(t)$ is designed to estimate the strictly periodic component of the sequence. This was achieved by expanding the series in

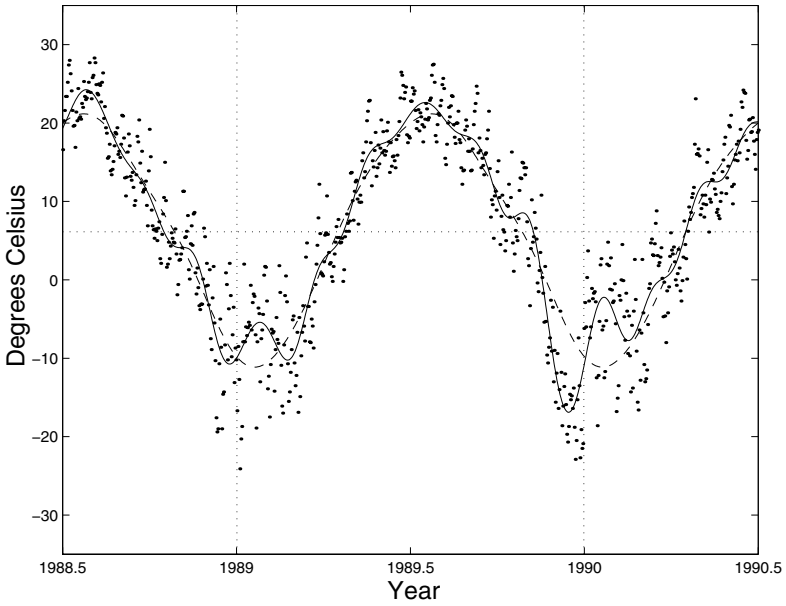

Figure 7.4. Temperature data for Montreal from mid-1988 to mid-1990. Daily mean temperatures are plotted as points, a smooth of the data as a solid line, and a strictly periodic trend as a dashed line. The horizontal dashed line indicates the mean temperature over the 34 years of data.

terms of nine Fourier basis functions with base period 365.25 days. In signal analysis jargon, we applied a high-pass filter. Now the standard deviation of the residuals from this trend was 4.74°C, which is necessarily higher than the unconstrained B-spline smooth, but we were surprised at how small the increase actually was. This periodic trend is shown as the dashed line in Figure 7.4.

We now subtract the strictly periodic curve $x_0(t)$ from the smooth curve $x(t)$ to highlight trends and events unexplained by seasonal variation. The result is shown in Figure 7.5, and the standard deviation of these differences is 2.15°C. We see the cold snap of 1989 as the strongest negative spike, and we also see a number of episodes where the smooth trend is either above or below zero for comparatively long periods. The temperature was higher than average for a long period after 1990, for example.

Some of this longer-term trend can be viewed as phase variation, due to the early or late arrival of some seasons. For example, the cold snap of 1989 would not have been so dramatic if it had come around January 15, 1990, when temperatures approaching −30°C happen more often, and indeed were seen a year earlier. We need to remove our estimate of the phase variation to get a better sense of just how extreme this event was.

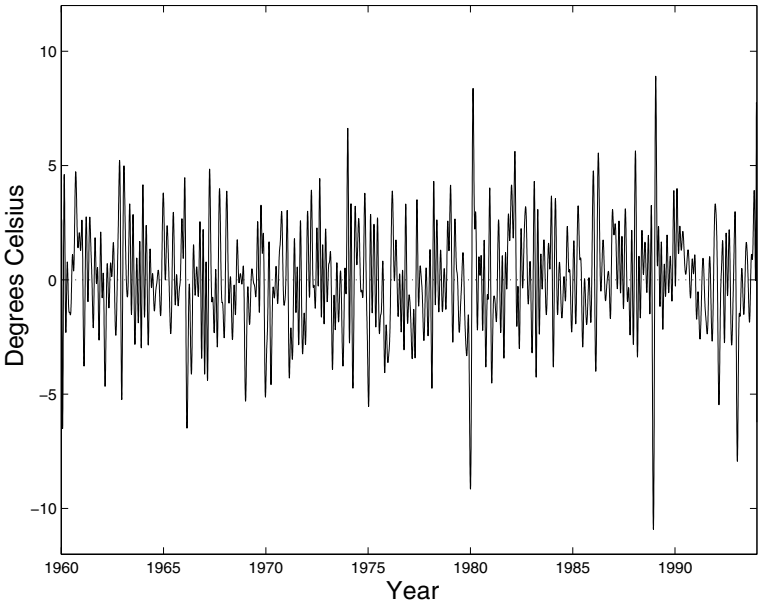

Figure 7.5. The difference between the smooth trend and the strictly periodic trend for the Montreal temperature data.

Figure 7.6 shows what happens around 1989 when we register the smooth trend $x(t)$ to the strictly periodic target $x_0(t)$. We used 140 B-spline basis functions of order 5 to define the relative acceleration function $w(t)$ defining time warping function $h(t)$ as described in Section 7.6, yielding a spacing between knots of three months. This seemed to give enough flexibility to capture some of the within-year phase variation, but not enough to distort fine features in the curves. Now we see that the cold snap at Christmas 1989 is positioned after registration in January 1990. The standard deviation of the differences between the registered temperature curve and the strictly periodic has now dropped to 1.73°C. We can now estimate the proportion of the variation of the unconstrained smooth around the strictly periodic smooth due to phase variation by the squared multiple correlation $R^2 = (2.15^2 - 1.73^2)/2.15^2 = 0.35$. Thus, about a third of the smooth variation in temperature is due to phase.

To get some idea of how much shift in time is required to achieve the results in Figure 7.6, we can plot the difference between the warped and actual time functions $h(t) - t$, called the *time deformation function*. This is shown in Figure 7.7, and we see that midwinter in 1989/1990 arrived about 25 days early.

What about global warming? The smaller residuals for the registered data fit by strictly linear trend should help us to detect any long-term linear trend in the data. The slope for the regression of these residuals

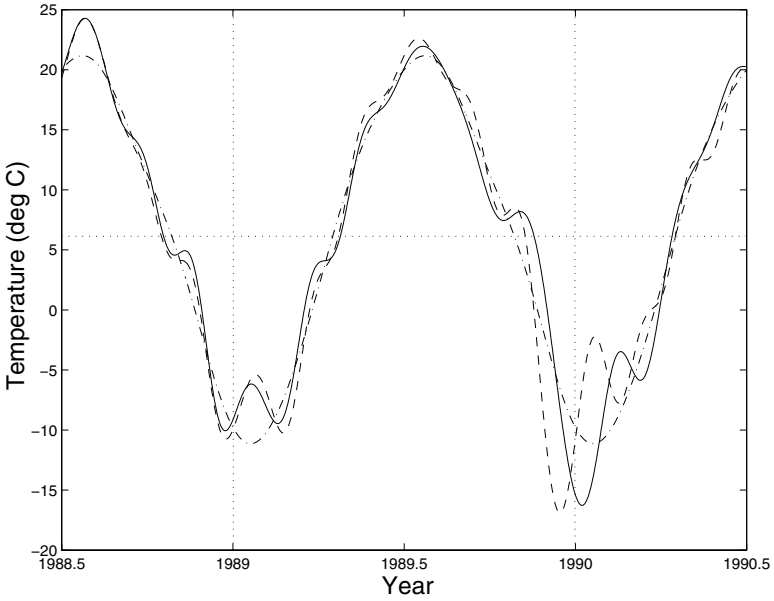

Figure 7.6. Temperature data for Montreal from mid-1988 to mid-1990 registered to the strictly periodic trend. The registered smooth of the data is the solid line, the unregistered smooth is the dashed line, and the strictly periodic trend is the dashed-dotted line. The horizontal dashed line indicates the mean temperature over the 34 years of data.

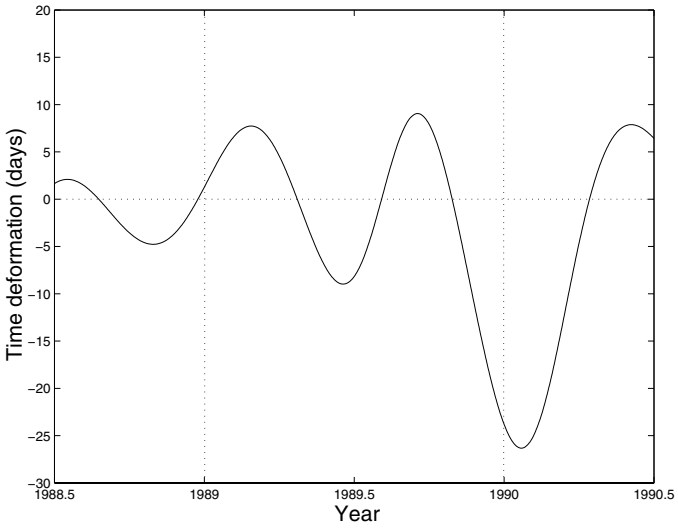

Figure 7.7. The time deformation function $h(t) - t$ for the registration results in Figure 7.6.

on time is 0.0024°C per year, a total of 0.08°C for the 34-year period of observation. The standard error of the regression coefficient, however, is 0.0016°C, and we cannot conclude that this amount of trend is significant.

7.5 What have we seen?

Although we have already seen the registration problem in Chapter 6, the two examples here introduce some new aspects. For the printing data we had to register the three coordinates simultaneously, that is, with a common time warping function $h(t)$. The amount of registration involved was substantially less than for the growth data, but we saw some rather dramatic improvements in the coherence of the tangential amplitude curves in Figure 7.2, and this turns out to be important when we analyze these data later.

Not all functional data involve multiple samples of curves. Rather, a long time series such as the temperature data also contains in a certain sense replicated data. There are 34 repetitions of the annual variation in temperature, and our strictly periodic smooth using the Fourier basis was, in fact, a type of averaging over these repetitions. When we registered the entire series to this periodic template, we discovered that the amount of phase variation was rather substantial, and required in certain years nearly a month of adjustment. Removing phase variation also led to a rather substantial reduction in the total variation of the smooth trend. This discovery seriously challenges most of the methods now used to analyze time series such as this, because they do not provide for phase variation.

7.6 Notes and references

In this section, we generally achieve some simplification of notation by dropping the subscript on the function $x_i(t)$ to be registered as well as the warping function $h_i(t)$.

7.6.1 Continuous registration

We may also register two curves by optimizing some measure of similarity of their shapes, and thus use the entire curves in the process. Put another way, the timings of a fixed set of landmarks provide one way of describing how similar the shapes of two curves are, but we can also choose measures that use the whole curves.

Silverman (1995) 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

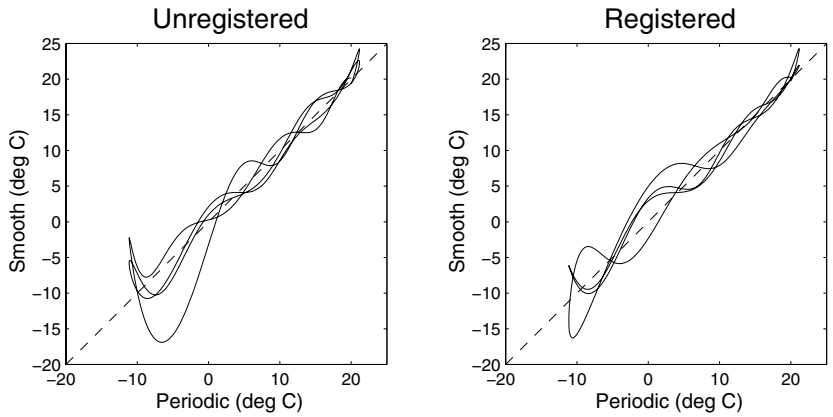

Figure 7.8. In the left panel the values of the unconstrained smooth from mid-1988 to mid-1990 are plotted against the corresponding values of the periodic smooth. In the right panel the registered smooth values are plotted against the periodic smooth values. We see that the values are now closer to the diagonal dashed line.

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. His measure of shape similarity was the total squared error, cast into functional terms as

$$\mathrm{FSSE}(h) = \int_0^{T_0} \{x[h(t)] - x_0(t)\}^2 \, dt \; . \tag{7.3}$$

This measure works well enough provided that the amount of amplitude variation is small, so that the pure phase variation model (7.1) is about right. However, the measure can run into trouble when $x(t)$ and $x_0(t)$ have the same shape but differ in amplitude. Ramsay and Li (1998) offer an example in which it is shown that this criterion has a tendency to "pinch in" the sides of the larger of the two curves in order to make it look more like the smaller.

To evolve an alternative fitting criterion, we could allow a scale factor A, which may depend on i, to yield

$$\mathrm{FSSE}(h, A) = \int_0^{T_0} \{x[h(t)] - Ax_0(t)\}^2 \, dt \; . \tag{7.4}$$

This would be zero if $x_0(t)$ and $x[h(t)]$ differ only by a scale factor, so that $x(t) = Ax_0(t)$ for some positive constant A. This means that the two functions have essentially the same shape, and that the values of $x(t)$ are proportional to those of $x_0(t)$. If the curves are exactly proportional, then

the matrix

$$\begin{bmatrix} \int \{x_0(t)\}^2\,dt & \int x_0(t)x[h(t)]\,dt \\ \int x_0(t)x[h(t)]\,dt & \int \{x[h(t)]\}^2\,dt \end{bmatrix} \tag{7.5}$$

is singular, so only one of its eigenvalues is nonzero. This is also the case if we replace the integrals in the matrix by sums over a mesh of values t_j.

Consider, for example, the relation of the smooth variation in the temperature data to their periodic trend over 1989, shown in the left panel of Figure (7.8). Note the large loop in the lower left of this plot, due to the early arrival of winter in this year. The eigenvalues of the matrix (7.5) are 2.380 and 0.032. The smaller eigenvalue is positive because these two sets of curve values are not proportional to each other.

This line of reasoning suggests that we might choose the warping function $h(t)$ to minimize the logarithm of the smallest eigenvalue of the cross-product matrix (7.5). Denote this quantity by MINEIG(h). In cases like the printing data, where the functions are multivariate, we can form a composite criterion by adding the criterion across functions. The criterion often works even better if we use the first derivative values, or even a higher derivative if it can be estimated stably. This is because derivatives tend to oscillate more rapidly than functions, and also to vary about zero, so that the smallest eigenvalue measure is even more sensitive to whether functions differ only by amplitude variation.

We can see how these two techniques work on an artificial example. Let the target function be $x_0(t) = \sin 2\pi t$, and let the function to be registered be $x(t) = \sqrt{2}(\sin 2\pi t + \cos 2\pi t)$. These two functions have a phase difference of 1/8, and $x(t)$ has a maximum of 2 as compared to the maximum of $x_0(t)$ of 1. Otherwise, the two functions have the same shape. The results are shown in the upper two panels of Figure 7.9, where we see that the registered function is a lateral shift by 0.125 of the unregistered function. In the upper-right panel, we see as expected that $h(t) \approx t$. The problem with the least squares criterion (7.3) can be seen in the bottom two panels. We see that this criterion is minimized in the presence of considerable amplitude differences by pinching in the larger curve over amplitudes where both the smaller and larger curve have values. The resulting warping function is far from diagonal, and even the lateral shift is poorly estimated, with a value of 0.117.

Returning to the registration of the temperature data, the right panel of Figure 7.8 shows that the registered smooth trend is more tightly related to a proportional relationship. The two eigenvalues are now 2.388 and 0.018, and, although the first eigenvalues hardly change at all, the second eigenvalue is now 57% of the corresponding value before registration.

A generalization, investigated by Kneip, Li, MacGibbon, and Ramsay (2000), is to replace the constant A by a smooth positive function $A_i(t)$ which does not vary too quickly. This allows local features of x_i to be

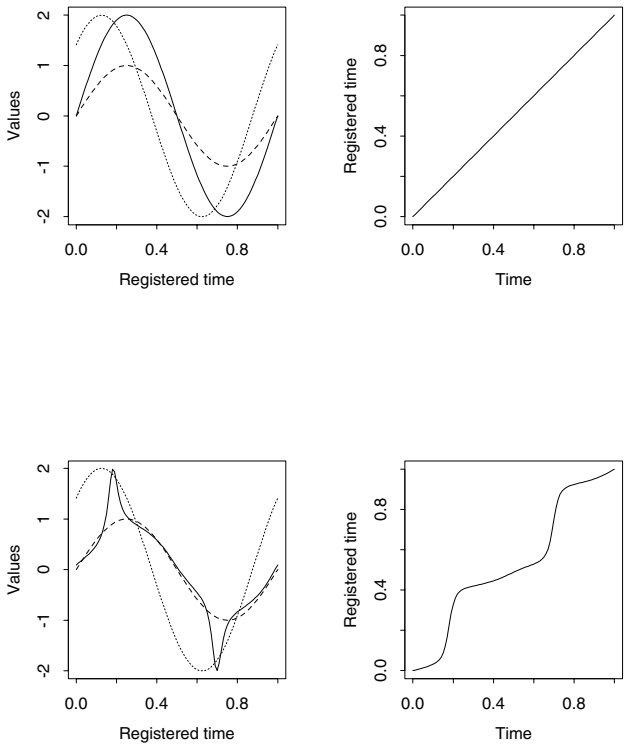

Figure 7.9. The upper two panels show results for an artificial registration problem using the minimum eigenvalue criterion. The dotted curve in the upper-left panel is the curve to be registered to the curve indicated by the dashed line. The solid line is the registered curve. The upper-right panel contains the warping function for this case, $h(t) = t$. The lower panels show the same results using the least squares criterion.

registered to those of x_0 even if the overall scale of variation is not constant across the whole range.

7.6.2 Estimation of the warping function

The software on the Web site associated with this chapter offers a choice between the two fitting criteria defined above: least squared error and minimum smallest eigenvalue of the cross-product matrix. Since the warping function $h(t)$ is strictly increasing, it can be represented using the methodology of Chapter 6 in terms of its relative acceleration $w(t) = h''(t)/h'(t)$. We can then permit a roughness penalty based on the mth derivative of $w(t)$, by minimizing

$$\text{MINEIG}_\lambda(h) = \text{MINEIG}(h) + \lambda \int \{w^{(m)}(t)\}^2 \, dt, \tag{7.6}$$

or the corresponding criterion based on FSSE. In the analyses we have presented, the MINEIG criterion was used. For either criterion, if $m = 0$, larger values of the smoothing parameter λ shrink the relative acceleration w to zero, and therefore shrink $h(t)$ to t. In practice, it is satisfactory to choose the smoothing parameter λ subjectively.

If we need to estimate derivatives of $h(t)$, it may be better to work with higher values of m. This can happen, for example, if we want to use derivatives of the registered functions with respect to t, in which case the chain rule will require the corresponding derivatives of $h(t)$.

Our software represents the function w in terms of a B-spline expansion. Ramsay and Li (1998) use order 1 (piecewise linear) B-splines for w since this permits the expression of h in a closed form and leads to relatively fast computation. Higher-order splines can be used at the expense of some numerical integration.