2

Tools for exploring functional data

2.1 Introduction

This chapter reviews topics that are notational and conceptual background to our main development of functional data analysis beginning in Chapter 3.

Our notation will already be familiar to many readers, but some will welcome a review, and others will encounter the notation that we use here for the first time. We have tried hard to avoid using notation other than what is familiar to statisticians and routine in calculus courses.

We will draw rather heavily on your expertise in matrix analysis and multivariate statistics, and you may want to consult Section A.7, which reviews some matrix algebra tools that we will need within framework of the multivariate linear model. This brief account is relevant here because, in fact, most of our functional data analyses and models will be converted to equivalent matrix formulations through the device of representing functions by basis function expansions, a topic that comes up in the next chapter. Also discussed in the Appendix are matrix decompositions, projections, and the constrained maximization of quadratic forms.

After some remarks on notation in Section 2.2, we consider the basic anatomy of a function in Section 2.4. What features in a function might be of interest? How are functions different from vectors? How do we quantify the amount of information that is needed to specify a function? What does it mean to say that a function is "smooth"?

2.2 Some notation

2.2.1 Scalars, vectors, functions and matrices

The reader should be warned that we try to use notation that brings out the basic structure of what is being done, and that this may entail the use of conventions that are at first sight a little unfamiliar. For example, we do not usually bother to distinguish in our notation between scalar quantities (numbers) and functions. This means that a single symbol x can refer to a scalar or to a function. The nervous reader should be assured that this convention is only used to clarify, rather than confuse, the discussion! In general, the context should always make clear when a symbol refers to a scalar or function. This emphasizes our guiding intuition that a function is to be considered as single unitary entity. The perhaps more familiar notation $x(t)$ refers to the *value* of function x at argument value t rather to the entire function.

On the other hand, in this edition we adhere to the usual practice of showing vectors as boldface lower case letters such as $\mathbf{x}$, and matrices in boldface upper case. We always use the notation $\mathbf{x}'$ for the transpose of a vector $\mathbf{x}$. We need matrix algebra at every turn, and it seems better not to ask readers used to bold symbols to do without this device.

If x is a vector or function, its elements or values x_i or $x(t)$ are usually scalars, but sometimes it is appropriate for the individual x_i or $x(t)$ to be a vector, and then we use boldface. Also, it is handy to use the notation $x(\mathbf{t})$ to denote the vector containing the values of function x at each of the argument values in vector $\mathbf{t}$.

It is often clearer to use longer strings of letters in a distinctive font to denote quantities more evocatively than standard notation allows. For example, we use names such as

- `Temp` for a temperature record,
- `Knee` for a knee angle
- `LMSSE` for a squared error fitting criterion for a linear model, and
- `RSQ` for a squared correlation measure.

2.2.2 Derivatives and integrals

Our notation for the derivative of order m of a function x is $D^m x$; this produces cleaner formulas than $d^m x/dt^m$. It stresses that differentiation is an *operator* that acts on a function x to produce another function Dx. Of course, $D^0 x$ refers to x itself. The superscript method works neatly when we consider derivatives of derivatives, and also when we use $D^{-1}x$ to refer to the indefinite integral of x, since $D^1 D^{-1} x = D^0 x = x$ as expected. We also use operators that act on functions in other ways, and it is convenient to use a consistent notation.

The definite integral $\int_a^b x(t)\,dt$ will often be shortened to $\int x$ when the context makes clear both the limits of integration a and b and the variable t over which the integration takes place.

2.2.3 Inner products

Inner product notation for functions, as in

$$\langle x, y\rangle = \int x(t)y(t)\,dt, \tag{2.1}$$

was used much more frequently in the first edition than in this. We found that many readers had difficulty coping with the notation, and we also found that we could do without it nearly everywhere. Nevertheless, inner product notation is a powerful tool, and if a reader wishes to learn more, the Appendix offers a summary and some illustrations. We will use rather more frequently the notation $\|x\|$ for the *norm* of x, a measure of its size. The most common type of norm, called the L_2 norm, is related to the inner product through the relation

$$\|x\|^2 = \langle x, x\rangle = \int x^2(t)\,dt\ .$$

The Appendix contains additional material on inner product notation.

2.2.4 Functions of functions

Functions are often themselves arguments for other functions. For example, in Chapter 7 we will consider a nonlinear transformation $h(t)$ of argument t that maps t on to the same interval that it occupies. That is, for example, time is transformed nonlinearly into time. We then need the function whose values are $x[h(t)]$, which we can indicate by x^*. In this case, we use the *functional composition* notation $x^* = x \circ h$. The function value $x^*(t)$ is indicated by $(x \circ h)(t)$.

Moreover, in the same chapter, we will use the *inverse* function which results from solving the relation $h(g) = t$ for g given t. This function, having values $g(t)$, is denoted by h^{-1}. This does not mean, of course, the reciprocal of h, which we simply indicate as $1/h$ on the rare occasion that we need it. In fact, the functional compositions $h \circ h^{-1}$ and $h^{-1} \circ h$ satisfy

$$(h \circ h^{-1})(t) = (h^{-1} \circ h)(t) = t$$

and, in functional composition sense, therefore h and h^{-1} cancel one another.

Another type of function transforms one function into another; that is, takes an entire function as its argument rather than a function value. The most important example is the transform D that transforms function x into its derivative Dx. The indefinite integral is another example, and as are the arithmetic operations applied to functions. We call such functional transformations *operations* or *operators*.

2.3 Summary statistics for functional data

2.3.1 Functional means and variances

The classical summary statistics for univariate data familiar to students in introductory statistics classes apply equally to functional data. The mean function with values

$$\bar{x}(t) = N^{-1} \sum_{i=1}^{N} x_i(t)$$

is the average of the functions point-wise across replications. Similarly the variance function `var` has values

$$\mathtt{var}_X(t) = (N-1)^{-1} \sum_{i=1}^{N} [x_i(t) - \bar{x}(t)]^2,$$

and the standard deviation function is the square root of the variance function.

Figure 2.1 displays the mean and standard deviation functions for the aligned pinch force data. We see that the mean force looks remarkably like a number of probability density functions well known to statisticians, and in fact the relationship to the lognormal distribution has been explored by Ramsay, Wang and Flanagan (1995). The standard deviation of force seems to be about 8% of the mean force over most of the range of the data.

2.3.2 Covariance and correlation functions

The *covariance function* summarizes the dependence of records across different argument values, and is computed for all t_1 and t_2 by

$$\mathtt{cov}_X(t_1, t_2) = (N-1)^{-1} \sum_{i=1}^{N} \{x_i(t_1) - \bar{x}(t_1)\}\{x_i(t_2) - \bar{x}(t_2)\}.$$

The associated *correlation function* is

$$\mathtt{corr}_X(t_1, t_2) = \frac{\mathtt{cov}_X(t_1, t_2)}{\sqrt{\mathtt{var}_X(t_1)\mathtt{var}_X(t_2)}}.$$

These are the functional analogues of the variance–covariance and correlation matrices, respectively, in multivariate data analysis.

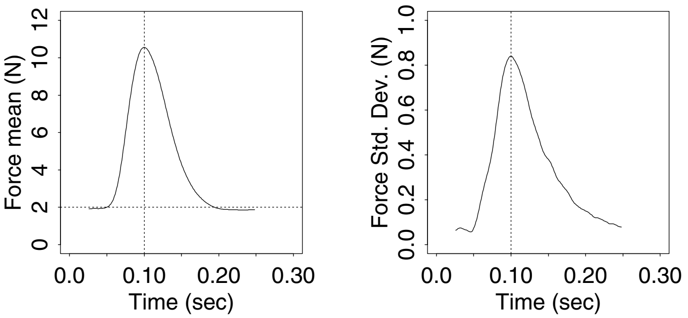

Figure 2.1. The mean and standard deviation functions for the 20 pinch force observations in Figure 1.11 after they were aligned or registered.

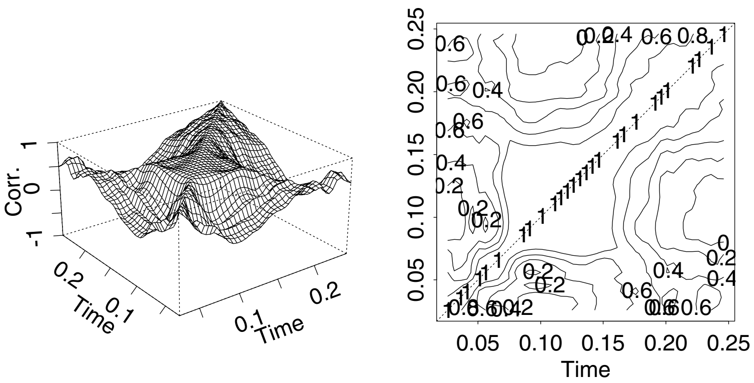

Figure 2.2. The left panel is a perspective plot of the bivariate correlation function values $r(t_1, t_2)$ for the pinch force data. The right panel shows the same surface by contour plotting. Time is measured in seconds.

Figure 2.2 displays the correlation function of the pinch force data, both as a surface over the plane of possible pairs of times (t_1, t_2) and also as a set of level contours.

Our experience with perspective and contour displays of correlation suggests that not everyone encountering them for the first time finds them easy to understand. Here is one strategy: The diagonal running from lower left to upper right in the contour or from front to back in the perspective plot of the surface contains the unit values that are the correlations between identical or very close time values. Directions perpendicular to this ridge of unit correlation indicate how rapidly the correlation falls off as two argument values separate. For example, one might locate a position along the unit ridge associated with argument value t, and then moving perpendicularly from this point shows what happens to the correlation between values at time pair $(t-\delta, t+\delta)$ as the perpendicular distance δ increases. In the case of the pinch force data, we note that the correlation falls off slowly for values on either side of the time 0.1 of maximum force, but declines much more rapidly in the periods before and after the impulse. This suggests a two-phase system, with fairly erratic uncoupled forces in the constant background force phase, but with tightly connected forces during the actual impulse. In fact, it is common to observe low correlations or rapid fall-off when a system is in a resting or ballistic state free from any outside input, but to show strong correlations, either positive and negative, when exogenous influences apply.

2.3.3 Cross-covariance and cross-correlation functions

In the case of the gait data discussed in Section 1.3, we had both hip and knee angles measured through time. In general, if we have pairs of observed functions (x_i, y_i), the way in which these depend on one another can be quantified by the *cross-covariance* function

$$\texttt{cov}_{X,Y}(t_1,t_2) = (N-1)^{-1}\sum_{i=1}^{N}\{x_i(t_1)-\bar{x}(t_1)\}\{y_i(t_2)-\bar{y}(t_2)\}.$$

or the *cross-correlation* function

$$\texttt{corr}_{X,Y}(t_1,t_2) = \frac{\texttt{cov}_{X,Y}(t_1,t_2)}{\sqrt{\texttt{var}_X(t_1)\texttt{var}_Y(t_2)}}.$$

Figure 2.3 displays the correlation and cross-correlation functions for the gait data. In each of the four panels, t_1 is plotted along the horizontal axis and t_2 along the vertical axis. The top left panel shows a contour plot of the correlation function $\texttt{corr}_{\texttt{Hip}}(t_1,t_2)$ for the hip angles alone, and the bottom right panel shows the corresponding plot for the knee angles. The cross-correlation functions $\texttt{corr}_{\texttt{Hip},\texttt{Knee}}$ and $\texttt{corr}_{\texttt{Knee},\texttt{Hip}}$ are plotted in the top right and bottom left panels respectively; since, in general, $\texttt{corr}_{X,Y}(t_1,t_2) = \texttt{corr}_{Y,X}(t_2,t_1)$, these are transposes of one another, in that each is the reflection of the other about the main diagonal $t_1 = t_2$. Note that each axis is labelled by the generic name of relevant data function, `Hip` or `Knee`, rather than by the argument value t_1 or t_2.

In this figure, different patterns of variability are demonstrated by the individual correlation functions $\texttt{corr}_{\texttt{Hip}}$ and $\texttt{corr}_{\texttt{Knee}}$ for the hip and knee angles considered separately. The hips show positive correlation throughout, so that if the hip angle is larger than average at one point in the cycle it will have a tendency to be larger than average everywhere. The contours on this plot are more or less parallel to the main diagonal, implying that the correlation is approximately a function of $t_1 - t_2$ and that the variation of the hip angles can be considered as an approximately stationary process.

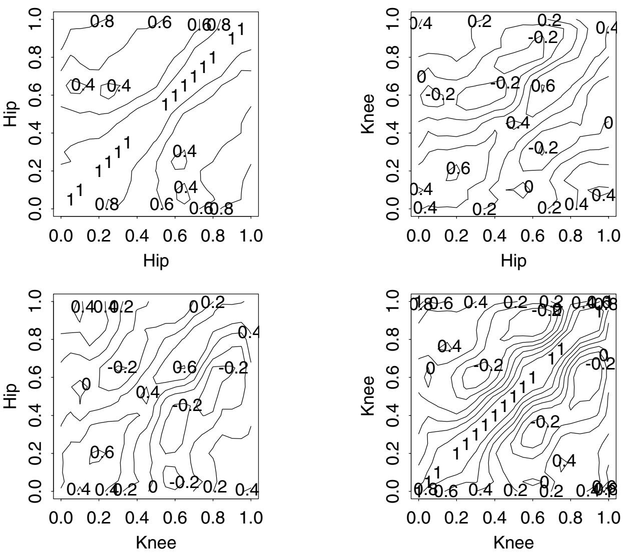

Figure 2.3. Contour plots of the correlation and cross-correlation functions for the gait data. In each panel t_1 is plotted on one axis and t_2 on the other; the legends indicate which observations are being correlated against each other.

On the other hand, the knee angles show behavior that is clearly nonstationary; the correlation between the angle at time 0.0 and time 0.3 is about 0.4, while that between times 0.3 and 0.6 is actually negative. In the middle of the cycle the correlation falls away rapidly as one moves away from the main diagonal, while at the ends of the cycle there is much longer range correlation. The hip angles show a slight, but much less marked, departure from stationarity of the same kind. These features may be related to the greater effect on the knee of external factors such as the heel strike and the associated weight placed on the joint, whereas the hip acts under much more even muscular control throughout the cycle.

The ridge along the main diagonal of the cross-correlation plots indicates that $\texttt{Hip}(t_1)$ and $\texttt{Knee}(t_2)$ are most strongly correlated when t_1 and t_2 are approximately equal, though the main ridge shows a slight reverse S shape (in the orientation of the top right panel). The analysis developed in Chapter 11 will elucidate the delays in the dependence of one joint on the other. Apart from this, there are differences in the way that the cross-correlations behave at different points of the cycle, but the cross-correlation function does not make it clear what these mean in terms of dependence between the functions.

Another example is provided by the Canadian weather data. Contour plotting in Figure 2.4 shows the correlation functions between tempera-

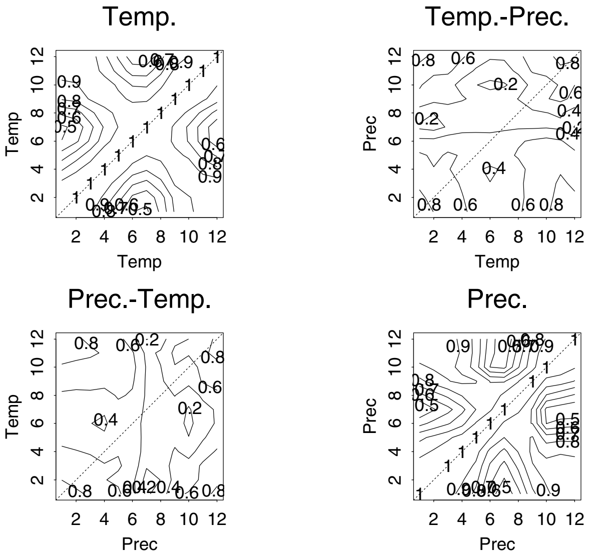

Figure 2.4. Contour plots of the correlation and cross-correlation functions for 35 Canadian weather stations for temperature and log precipitation. The cross-correlation functions are those in the upper right and lower left panels.

ture and log precipitation based on monthly data. The correlation is high for both temperature and precipitation on either side of the midsummer period, so that autumn weather tends to be highly correlated with spring weather. By contrast, winter and summer weather have a weaker correlation of around 0.5. The cross-correlations show that midsummer precipitation has a near zero correlation with temperature at any point in the year, but that midwinter temperature and midwinter precipitation are highly correlated. This is due to the fact that, in continental weather stations, both measures tend to be especially low in midwinter, whereas in marine stations, the tendency is for both temperature and precipitation to be higher.

2.4 The anatomy of a function

2.4.1 Functional features

What interests us when we consider functions such as the height acceleration curves in Figure 1.2? Certainly the *peak* and *valley* defining the pubertal growth spurt, as well as the smaller peaks at age 6 for most girls. *Crossings* of specified levels can also be important markers, such as the age at which acceleration is zero in the middle of the pubertal growth spurt, marking out the point of peak growth velocity. *Levels* are function values that we consider significant, such as the zero level that a growth acceleration reaches after growth has stopped.

We can consider each of these functional features as events that are associated with a specific value of the argument t. That is, most features are characterized by a *location*. Many are also defined by *amplitude*, a measure of their size. For example, the height or depth of a peak or valley, respectively, is a matter of amplitude, as is the steepness with which a line crosses a specified level. Finally, events like peaks and valleys are also characterized by *widths*; the first peak in the knee angle curves in Figure 1.8 is narrower than the second peak.

In this sense, levels are one-dimensional events, crossings are two-dimensional, and peaks and valleys are three-dimensional. That is, in ideal errorless circumstances, we would need three pieces of information to fully define a peak, namely location, amplitude, and width. This corresponds to the fact that peaks look somewhat like parabolas, which are defined by three parameters; crossings look like lines, requiring two parameters; and levels are like points.

The *dimensionality* of a functional feature tells us a great deal about how much information we will need to estimate it. For example, even a tiny bit of observational error in the data will force us to provide five rather than three function values at locations within a peak, and for data with error levels common in functional data analyses, seven to eleven values per peak would be wise.

2.4.2 Data resolution and functional dimensionality

This suggests the notation of the *resolving power* or *resolution* of a set of data. This is inversely related to the width of the narrowest event that can be estimated to our satisfaction. We mean by the phrase "high resolution data" that they can pin down small events. The resolution of a set of data can be a rather more useful concept than simply the number of observations taken.

Resolution leads in turn to the notion of the *dimensionality* of a function. Expertise in the mathematical area of functional analysis is necessary to understand this concept in depth, but it is easy to say some common sensical things about the dimension of a curve. Roughly speaking, it is the sum across functional "features" of the numbers of pieces of information that are required to define each feature or event.

We can say that the *practical dimensionality* of a function is the total amount of information required to define it to some required level. This notion inevitably depends on the goals of the functional data analysis, since it supposes that we ignore error and other sources of high frequency variation that would increase the actual dimensionality of the function greatly.

Functions are potentially infinite dimensional. That is, a complete specification of a function x could conceivably require us to know its value $x(t)$ at each possible argument value t, and since there are an infinity of these, the dimensionality of a function can be arbitrarily large. Or, put another way, if a function can pack an infinite number of peaks and valleys within any interval, no matter how small, we will need infinite resolving power in any set of data concerning this curve. For example, the terms like "Brownian motion" and "white noise" are used to describe functions so erratic that no information is contained in $x(t)$ about the value $x(t+\delta)$, no matter how small δ is. This is somewhat depressing, because it implies that we can never collect enough data to estimate functions like these exactly.

However, in practice we work with functions that do not display so much complexity. It is more or less accepted, for example, that from 12 to 16 pieces of information, in a sense to be made precise in the next chapter, are required to describe growth curves like those in Figure 1.1. Almost always there are several ways in which we can use this much information to get about the same result, and in the growth curve literature there are several competing parametric models. But what matters is that all of the successful growth curve models seem to need at least this much information.

2.4.3 The size of a function

Something like *energy* tends govern the behavior of many functional variables, just as it does in physics. By this we mean that change requires effort or work, and typically the systems that we study can only muster a limited amount of whatever brings change per unit time. For example, even a process as seemingly chaotic as the stock market reflects, on a time scale small enough, the effort required to move money and information from one place to another. Biological systems like growing children likewise cannot make very rapid changes to their status due to the need to burn calories to bring this change about. Because on a short time scale the energy available in a system is essentially conserved, we can expect to see smooth changes, just as we will not see extremely large accelerations in mechanical systems with substantial mass.

Consequently, the dimensionality of a function is actually a measure of its *size* in the same way that its amplitude is. That is, both amplitude and dimensionality require energy to produce. For example, white noise is an infinitely large function, even if its values are always within specified limits such as $[-1, 1]$, because it would take an infinite amount of energy to produce this much variability. Similarly, what mathematicians refer to as Brownian motion is an abstraction inspired by the seemingly chaotic but actually limited movements of small particles due to collision with molecules in the medium in which they are suspended. One learns in functional analysis, for example, that an infinite dimensional hyper-sphere of radius one is infinitely large. Statisticians are referring to something like this by the colorful phrase “the curse of dimensionality.”

Dimensionality matters a great deal as a size indicator in functional data analysis. We will return to this important theme in the next chapter when we consider what the terms “noise” and “observational error” might mean in a functional sense, and when we take up the notion *multi-resolution analysis.*

2.5 Phase-plane plots of periodic effects

The two concepts of energy and of functional data having variation on more than one time scale lead to the graphical technique of plotting one derivative against another, something that we will call *phase-plane plotting.* We saw an example in Figure 1.13, and we now return to the U.S. nondurable goods manufacturing index to illustrate these ideas.

Like most economic indicators, the nondurable goods index tends to exhibit exponential increase, corresponding to percentage increases over fixed time periods. Moreover, the index tends to increase in size and volatility at the same time, so that the large relative effects surrounding the Second World War seem to be small relative to the large changes in the 1970s and 1980s, and seasonal variation in recent years dwarfs that in early years.

2.5.1 The log nondurable goods index

We prefer, therefore, to study the logarithm of this index, displayed in Figure 2.5. The log index has a linear trend with a slope of 0.016, corresponding to an annual rate of increase of 1.6%, and the sizes of the seasonal cycles are also more comparable across time. We now see that the changes in the depression and war periods are now much more substantial and abrupt than those in recent times. The growth rate is especially high from 1960 to 1975, when the baby boom was in the years of peak consumption; but in subsequent years seems to be substantially lower, perhaps because middle-aged “boomers” consume less, or possibly because the nature of the index itself has changed.

The goods index exhibits variation on four time scales:

- The longest scale is the century-long nearly linear increase in the log index, or exponential trend in the index itself.

- There are events that last a decade or more, such as the depression, the unusually rapid growth in the 1960s, and the slower growth in the last two decades.

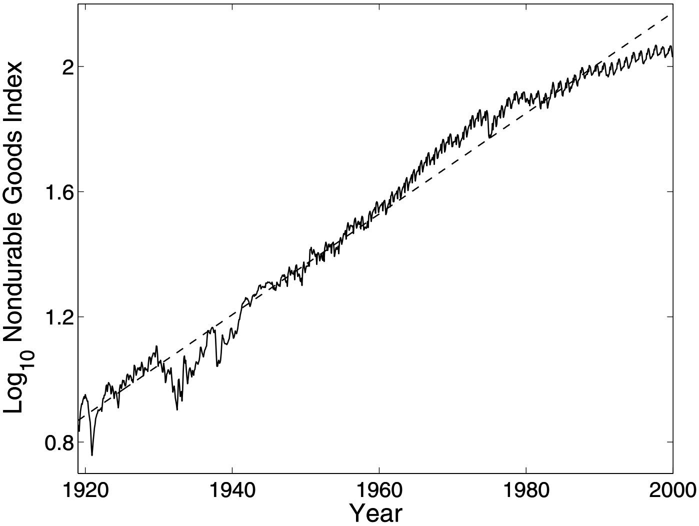

Figure 2.5. The monthly nondurable goods production of the United States shown in Figure 1.3 plotted on a logarithmic scale. The dotted straight line is estimated by least squares regression, and has a slope of 0.016, corresponding to a 1.6% increase in the index per year.

- Shorter term perturbations are also visible, such as World War II and the end of the Vietnam War in 1974.
- On the shortest scale there is seasonal variation over an annual cycle that tends to repeat itself.

A closer look at a comparatively stable period, 1964 to 1967 shown in Figure 2.6, suggests that the index varies fairly smoothly and regularly within each year. The solid line is a smooth of these data using the roughness penalty method described in Chapter 5. We now see that the variation within this year is more complex than Figure 2.5 can possibly reveal. This curve oscillates three times during the year, with the size of the oscillation being smallest in spring, larger in the summer, and largest in the autumn. In fact each year shows smooth variation with a similar amount of detail, and we now consider how we can explore these within-year patterns.

2.5.2 Phase–plane plots show energy transfer

Now that we have derivatives at our disposal, we can learn new things by studying how derivatives relate to each other. Our tool is a plot of acceleration against velocity. To see how this might be useful, consider the phase-plane plot of the function $\sin(2\pi t)$, shown in Figure 2.7. This simple function describes a basic *harmonic process*, such as the vertical position

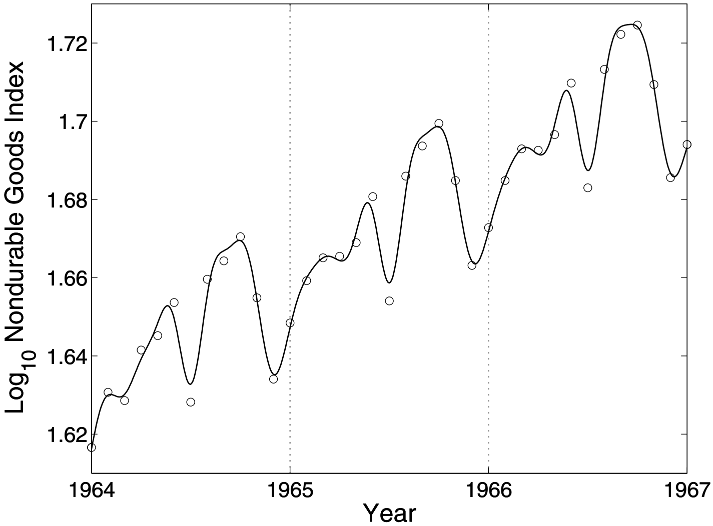

Figure 2.6. The log nondurable goods index for 1964 to 1967, a period of comparative stability. The solid line is a fit to the data using a polynomial smoothing spline. The circles indicate the value of the log index at the first of the month.

of the end of a suspended spring bouncing with a period of one time unit and starting at position zero at time $t = 0$.

The spring oscillates because energy is exchanged between two states: *potential* and *kinetic*. At times $\pi, 3\pi, \ldots$ the spring is at one or the other end of its trajectory, and the restorative force due to its stretching has brought it to a standstill. At that point, its potential energy is maximized, and so is the force, which is acting either upward (positively) or downward. Since force is proportional to acceleration, the second derivative of the spring position, $-(2\pi)^2 \sin(2\pi t)$, is also at its highest absolute value, in this case about ± 40. On the other hand, when the spring is passing through the position 0, its velocity, $2\pi \cos(2\pi t)$, is at its greatest, about ± 8, but its acceleration is zero. Since kinetic energy is proportional to the square of velocity, this is the point of highest kinetic energy. The phase-plane plot shows this energy exchange nicely, with potential energy being maximized at the extremes of Y and kinetic energy at the extremes of X.

Now harmonic processes and energy exchange are found in many situations besides mechanics. In economics, potential energy corresponds to available capital, human resources, raw material, and other resources that are at hand to bring about some economic activity, in this case the manufacture of nondurable goods. Kinetic energy corresponds to the manufacturing process in full swing, when these resources are moving along the assembly line, and the goods are being shipped out the factory door.

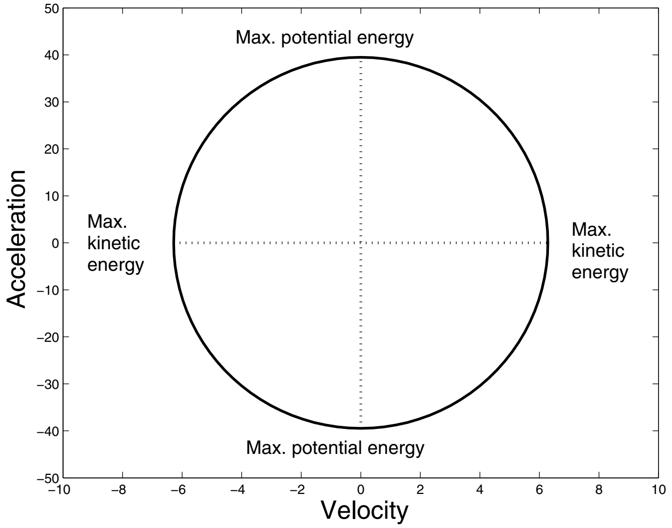

Figure 2.7. A phase-plane plot of the simple harmonic function $\sin(2\pi t)$. Kinetic energy is maximized when acceleration is 0, and potential energy is maximized when velocity is 0.

The process moves from strong kinetic to strong potential energy when the rate of change in production goes to zero. We see this, for example, after a period of rapid increase in production when labor supply and raw material stocks become depleted, and consequently potential energy is actually in a negative state. Or it happens when management winds down production because targets have been achieved, so that personnel and material resources are piling up and waiting to be used anew.

After a period of intense production, or at certain periods of crisis that we examine shortly, we may see that both potential and kinetic energy are low. This corresponds to a period when the phase-plane curve is closer to zero than is otherwise the case.

To summarize, here's what we are looking for:

- a substantial cycle;
- the size of the radius: the larger it is, the more energy transfer there is in the event;
- the horizontal location of the center: if it is to the right, there is net positive velocity, and if to the left, there is net negative velocity;
- the vertical location of the center: if it is above zero, there is a net velocity increase; if below zero, there is velocity decrease; and
- changes in the shapes of the cycles from year to year.

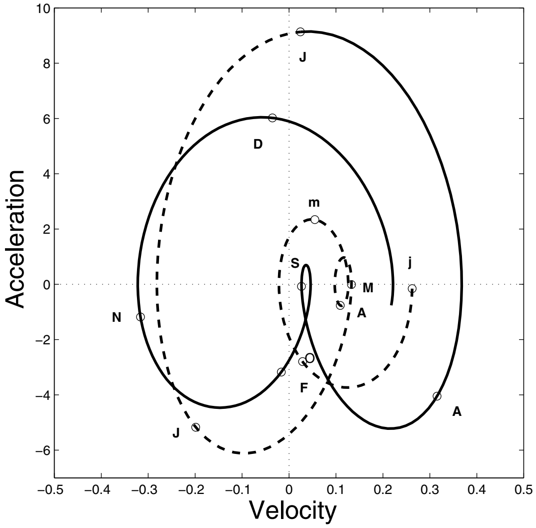

Figure 2.8. A phase-plane plot of the first derivative or velocity and the second derivative or acceleration of the smoothed log nondurable goods index for 1964. Letters indicate mid–months, with lowercase letters used for January and March. For clarity, the first half of the year is plotted as a dashed line, and the second half as a solid line.

2.5.3 The nondurable goods cycles

We use the phase-plane plot, therefore, to study the energy transfer within the economic system. We can examine the cycle within individual years, and also see more clearly how the structure of the transfer has changed throughout the twentieth century. Figure 2.8, a reproduction here of Figure 1.13, phase-plane plots the year 1964, a year in a relatively stable period for the index. To read the plot, find the lower-case “j” in the middle right of the plot, and move around the diagram clockwise, noting the letters indicating the months as you go. You will see that there are two large cycles surrounding zero, plus some small cycles that are much closer to the origin.

The largest cycle begins in mid-May (M), with positive velocity but near zero acceleration. Production is increasing linearly or steadily at this point. The cycle moves clockwise through June (first J) and passes the horizontal zero acceleration line at the end of the month, when production is now decreasing linearly. By mid-July (second J) kinetic energy or velocity is near zero because vacation season is in full swing. But potential energy or acceleration is high, and production returns to the positive kinetic/zero potential phase in early August (A), and finally concludes with a cusp at summer’s end (S). At this point the process looks like it has run out of both potential and kinetic energy.

The cusp, near where both derivatives are zero, corresponds to the start of school in September, and to the beginning of the next big production cycle passing through the autumn months of October through November. Again this large cycle terminates in a small cycle with little potential and kinetic energy. This takes up the months of February and March (F and m). The tiny subcycle during April and May seems to be due to the spring holidays, since the summer and fall cycles, as well as the cusp, don't change much over the next two years, but the spring cycle cusp moves around, reflecting the variability in the timings of Easter and Passover.

To summarize, the production year in the 1960s has two large cycles swinging widely around zero, each terminating in a small cusp–like cycle. This suggests that each large cycle is like a balloon that runs out of air, the first at the beginning of school, and the second at the end of winter. At the end of each cycle, it may be that new resources must be marshalled before the next production cycle can begin.

2.6 Further reading and notes

These notes on other sources of information are intended only if you have some need to go beyond what is in this book. Otherwise, please push on to the following chapters, where we have tried to provide introductions to any concepts that you need to deal with at least the core topics for functional data analysis.

We find that inner product notation is appearing more and more often in statistics, and that it is already routinely used in engineering in fields such as signal analysis. Moore (1985) is an example of a reference oriented to applications of functional analysis that can be consulted for further information on many topics in this and subsequent chapters.

There have been many books that have used the notation of functional analysis to describe multivariate statistics, with a view to generalizing that methodology and synthesizing results within a common notational framework, but unfortunately not many that would be readable by anyone except mathematics specialists. Two references, however, have landmark qualities. Cailliez and Pagès (1976) attempted to write a text that combined high mathematics with an applied data analysis orientation, and the result was a unique and exciting approach that still merits attention for those able to read French. Our treatment of summary statistics in Section 2.3 is extended in many ways in their work. Grenander (1980) is a much more advanced book that we think of as dealing with many of the topics covered in this volume.

To see more of phase–plane plotting in action, consult Ramsay and Silverman (2002), where the method is used to show changes in the seasonal trend over longer time scales. The idea is taken directly from elementary physics, where conservation of energy is used in so many ways. This graphical tool links naturally to differential equation models that are considered Chapter 17 and subsequently.

Since observed curves are often complex objects requiring large numbers of parameters to describe adequately, as we shall see in the next three chapters, finding ways to summarize their distribution can be a challenge. In fact, it is relatively routine to have the number of curves N rather less than the number of parameters n that must be estimated per curve. We will use principal components analysis in Chapters 8 to 10 to capture at least a few dimensions of the variation across curves. Hall and Heckman (2002) propose an ingenious technique using what they call *density ascent lines* to provide interesting summaries of the probability density function for curve data.