Use R!

Series Editors:
Robert Gentleman   Kurt Hornik   Giovanni Parmigiani

For other titles published in this series, go to
http://www.springer.com/series/6991

J.O. Ramsay • Giles Hooker • Spencer Graves

Functional Data Analysis with R and MATLAB

Springer

J.O. Ramsay
2748, Howe Street
Ottawa, ON K2B 6W9
Canada
ramsay@psych.mcgill.ca

Giles Hooker
Department of Biological Statistics
& Computational Biology
Cornell University
1186, Comstock Hall
Ithaca, NY 14853
USA
gjh27@cornell.edu

Spencer Graves
Productive Systems Engineering
751, Emerson Ct.
San Jose, CA 95126
USA
spencer.graves@prodsyse.com

ISBN 978-0-387-98184-0 e-ISBN 978-0-387-98185-7
DOI 10.1007/978-0-387-98185-7
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2009928040

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This contribution to the useR! series by Springer is designed to show newcomers how to do functional data analysis in the two popular languages, Matlab and R. We hope that this book will substantially reduce the time and effort required to use these techniques to gain valuable insights in a wide variety of applications.

We also hope that the practical examples in this book will make this learning process fun, interesting and memorable. We have tried to choose rich, real-world problems where the optimal analysis has yet to be performed. We have found that applying a spectrum of methods provides more insight than any single approach by itself. Experimenting with graphics and other displays of results is essential.

To support the acquisition of expertise, the "scripts" subdirectory of the companion fda package for R includes files with names like "fdarm-ch01.R", which contain commands in R to reproduce virtually all of the examples (and figures) in the book. This can be found on any computer with R and fda installed using `system.file('scripts', package='fda')`. The Matlab code is provides as part of the fda package for R. From within R, it can be found using `system.file( 'Matlab', package='fda')`. It also can obtained by downloading the `.tar.gz` version of the fda package for R from the Comprehensive R Archive Network (CRAN, `www.r-project.org`), unzipping it and looking for the `inst/Matlab` subdirectory.

The contents of a book are fixed by schedules for editing and printing. These script files are not similarly constrained. Thus, in some cases, the script files may perform a particular analysis differently from how it is described in the book. Such differences will reflect improvements in our understanding of preferred ways of performing the analysis described in the book. The web site www.functionaldata.org is a resource for ongoing developments of software, new tools and current events.

The support for two languages is perhaps a bit unusual in this series, but there are good reasons for this. Matlab is expensive for most users, but its for capacity modeling dynamical systems and other engineering applications has been critical in the development of today's fda package, especially in areas such chemical engineering where functional data are the rule rather than the exception and where Matlab is widely used. On the other hand, the extendibility of R, the easy interface with lowerlevel languages, and above all its cost explain its popularity in many fields served by statisticians, students and new researchers. We hope that we can help many of our readers to appreciate the strengths of each language, so as to invest wisely later on. Secondarily, we hope that any user of either language wanting to learn the other can benefit from seeing the same analyses done in both languages.

As with most books in this useR! series, this is not the place to gain enough technical knowledge to claim expertise in functional data analysis nor to develop new tools. But we do hope that some readers will find enough of value here to want to turn to monographs on functional data analysis already published, such as Ramsay and Silverman (2005), and to even newer works.

We wish to end this preface by thanking our families, friends, students, employers, clients and others who have helped make us what we are today and thereby contributed to this book and to our earlier efforts. In particular, we wish to thank John Kimmel of Springer for organizing this series and inviting us to create this book.

James Ramsay, McGill University
Giles Hooker, Cornell University
Spencer Graves, San Jose, CA

Contents

Chapter 1
Introduction to Functional Data Analysis

The main characteristics of functional data and of functional models are introduced. Data on the growth of girls illustrate samples of functional observations, and data on the US nondurable goods manufacturing index are an example of a single long multilayered functional observation. Data on the gait of children and handwriting are multivariate functional observations. Functional data analysis also involves estimating functional parameters describing data that are not themselves functional, and estimating a probability density function for rainfall data is an example. A theme in functional data analysis is the use of information in derivatives, and examples are drawn from growth and weather data. The chapter also introduces the important problem of registration: aligning functional features.

The use of code is not taken up in this chapter, but R code to reproduce virtually all of the examples (and figures) appears in files ”fdarm-ch01.R” in the ”scripts” subdirectory of the companion ”fda” package for R, but without extensive explanation in this chapter of why we used a specific command sequence.

1.1 What Are Functional Data?

1.1.1 Data on the Growth of Girls

Figure 1.1 provides a prototype for the type of data that we shall consider. It shows the heights of 10 girls measured at a set of 31 ages in the Berkeley Growth Study (Tuddenham and Snyder, 1954). The ages are not equally spaced; there are four measurements while the child is one year old, annual measurements from two to eight years, followed by heights measured biannually. Although great care was taken in the measurement process, there is an average uncertainty in height values of at least three millimeters. Even though each record is a finite set of numbers, their values reflect a smooth variation in height that could be assessed, in principle, as

J.O. Ramsay et al., *Functional Data Analysis with R and MATLAB*, Use R,
DOI: 10.1007/978-0-387-98185-7_1,

often as desired, and is therefore a height *function*. Thus, the data consist of a sample of 10 *functional* observations $\texttt{Height}_i(t)$.

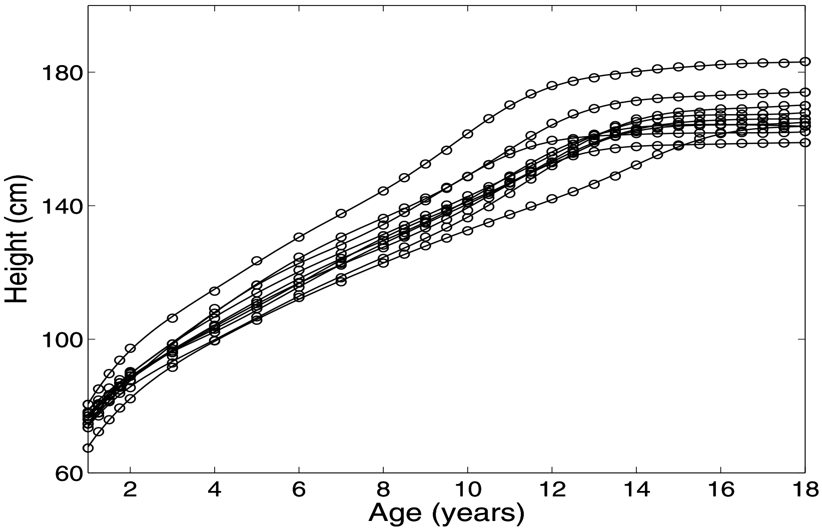

Fig. 1.1 The heights of 10 girls measured at 31 ages. The circles indicate the unequally spaced ages of measurement.

There are features in these data too subtle to see in this type of plot. Figure 1.2 displays the acceleration curves $D^2\texttt{Height}_i$ estimated from these data by Ramsay et al. (1995a) using a technique discussed in Chapter 5. We use the notation D for differentiation, as in

$$D^2\texttt{Height} = \frac{d^2\texttt{Height}}{dt^2}.$$

The pubertal growth spurt shows up as a pulse of strong positive acceleration followed by sharp negative deceleration. But most records also show a bump at around six years that is termed the midspurt. We therefore conclude that some of the variation from curve to curve can be explained at the level of certain derivatives. The fact that derivatives are of interest is further reason to think of the records as functions rather than vectors of observations in discrete time.

The ages are not equally spaced, and this affects many of the analyses that might come to mind if they were. For example, although it might be mildly interesting to correlate heights at ages 9, 10 and 10.5, this would not take account of the fact that we expect the correlation for two ages separated by only half a year to be higher than that for a separation of one year. Indeed, although in this particular example the ages at which the observations are taken are nominally the same for each girl, there is no real need for this to be so. In general, the points at which the functions are observed may well vary from one record to another.

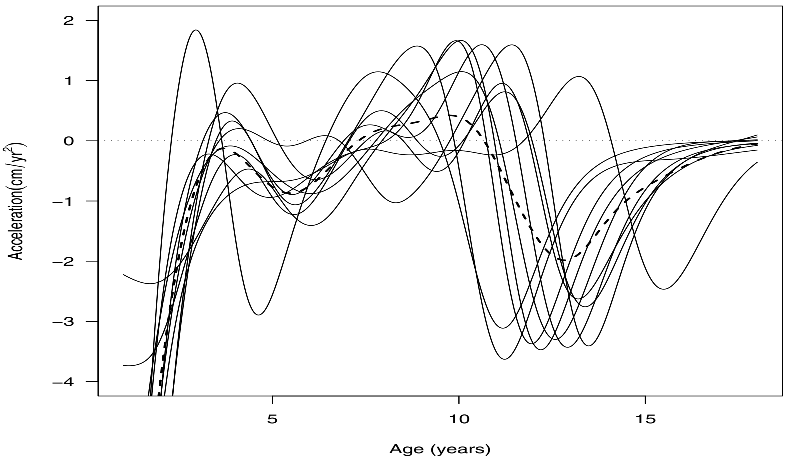

Fig. 1.2 The estimated accelerations of height for 10 girls, measured in centimeters per year per year. The heavy dashed line is the cross-sectional mean and is a rather poor summary of the curves.

The replication of these height curves invites an exploration of the ways in which the curves vary. This is potentially complex. For example, the rapid growth during puberty is visible in all curves, but both the timing and the intensity of pubertal growth differ from girl to girl. Some type of principal components analysis would undoubtedly be helpful, but we must adapt the procedure to take account of the unequal age spacing and the smoothness of the underlying height functions.

It can be important to separate variation in *timing* of significant growth events, such as the pubertal growth spurt, from variation in the *intensity* of growth. We will look at this in detail in Chapter 8 where we consider *curve registration*.

1.1.2 Data on US Manufacturing

Not all functional data involve independent replications; we often have to work with a single long record. Figure 1.3 shows an important economic indicator: the nondurable goods manufacturing index for the United States. Data like these often show variation as multiple levels.

There is a tendency for the index to show geometric or exponential increase over the whole century, and plotting the logarithm of the data in Figure 1.4 makes this trend appear linear while giving us a better picture of other types of variation. At a finer scale, we see departures from this trend due to the depression, World War II, the end of the Vietnam War and other more localized events. Moreover, at an

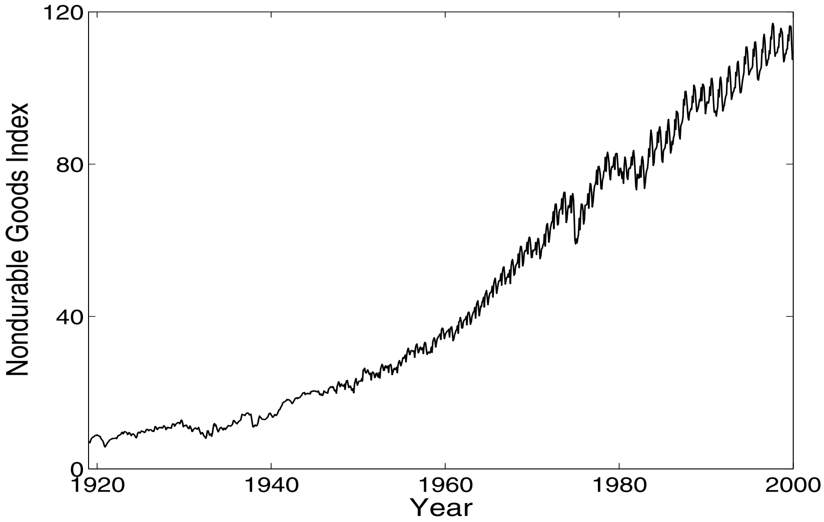

Fig. 1.3 The monthly nondurable goods manufacturing index for the United States.

even finer scale, there is a marked annual variation, and we can wonder whether this *seasonal trend* itself shows some longer-term changes. Although there are no independent replications here, there is still a lot of repetition of information that we can exploit to obtain stable estimates of interesting curve features.

1.1.3 Input/Output Data for an Oil Refinery

Functional data also arise as input/output pairs, such as in the data in Figure 1.5 collected at an oil refinery in Texas. The amount of a petroleum product at a certain level in a distillation column or cracking tower, shown in the top panel, reacts to the change in the flow of a vapor into the tray, shown in the bottom panel, at that level. How can we characterize this dependency? More generally, what tools can we devise that will show how a system responds to changes in critical input functions as well as other covariates?

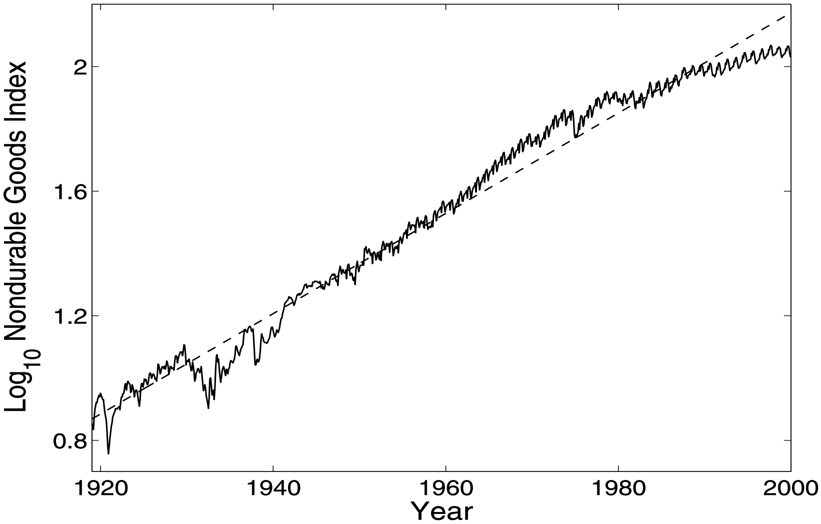

Fig. 1.4 The logarithm of the monthly nondurable goods manufacturing index for the United States. The dashed line indicates the linear trend over the whole time period.

1.2 Multivariate Functional Data

1.2.1 Data on How Children Walk

Functional data are often multivariate. Our third example is in Figure 1.6. The Motion Analysis Laboratory at Children's Hospital, San Diego, CA, collected these data, which consist of the angles formed by the hip and knee of each of 39 children over each child's gait cycle. See Olshen et al. (1989) for full details. Time is measured in terms of the individual gait cycle, which we have translated into values of t in $[0,1]$. The cycle begins and ends at the point where the heel of the limb under observation strikes the ground. Both sets of functions are periodic and are plotted as dotted curves somewhat beyond the interval for clarity. We see that the knee shows a two-phase process, while the hip motion is single-phase. It is harder to see how the two joints interact: The figure does not indicate which hip curve is paired with which knee curve. This example demonstrates the need for graphical ingenuity in functional data analysis.

Figure 1.7 shows the gait cycle for a single child by plotting knee angle against hip angle as time progresses round the cycle. The periodic nature of the process implies that this forms a closed curve. Also shown for reference purposes is the same relationship for the average across the 39 children. An interesting feature in this plot is the cusp occurring at the heel strike as the knee momentarily reverses its extension to absorb the shock. The angular velocity is clearly visible in terms of the spacing between numbers, and it varies considerably as the cycle proceeds.

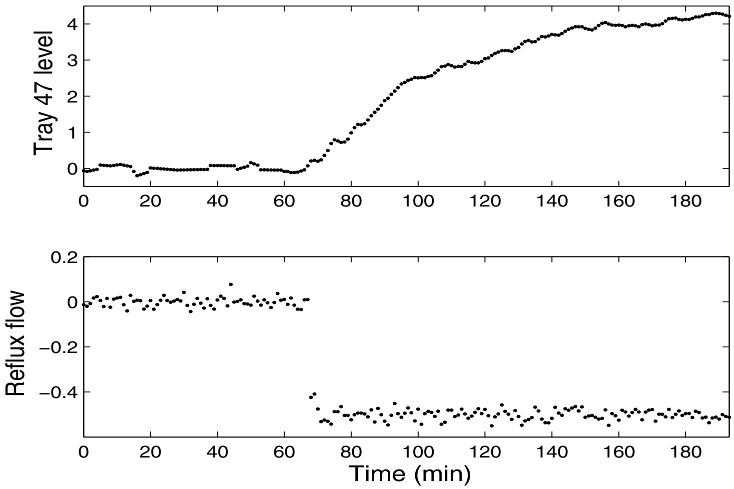

Fig. 1.5 The top panel shows 193 measurements of the amount of petroleum product at tray level 47 in a distillation column in an oil refinery. The bottom panel shows the flow of a vapor into that tray during an experiment.

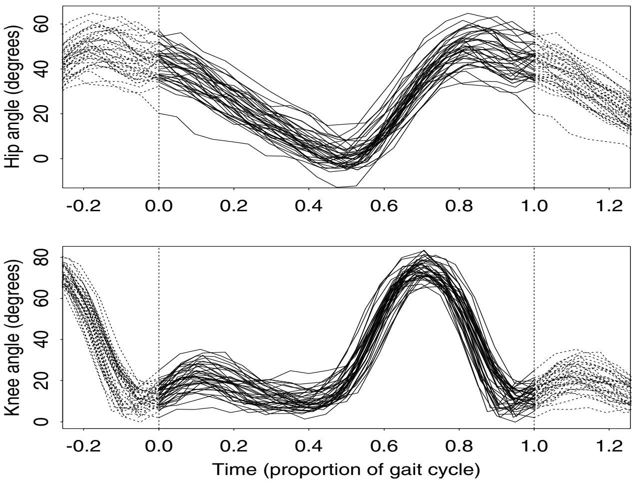

Fig. 1.6 The angles in the sagittal plane formed by the hip and knee as 39 children go through a gait cycle. The interval $[0,1]$ is a single cycle, and the dotted curves show the periodic extension of the data beyond either end of the cycle.

The child whose gait is represented by the solid curve differs from the average in two principal ways. First, the portion of the gait pattern in the C–D part of the cycle shows an exaggeration of movement relative to the average. Second, in the part of the cycle where the hip is most bent, this bend is markedly less than average; interestingly, this is not accompanied by any strong effect on the knee angle. The overall shape of the cycle for this particular child is rather different from the average. The exploration of variability in these functional data must focus on features such as these.

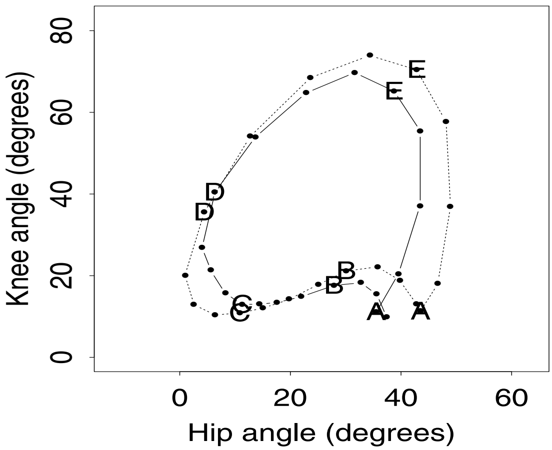

Fig. 1.7 Solid line: The angles in the sagittal plane formed by the hip and knee for a single child plotted against each other. Dotted line: The corresponding plot for the average across children. The points indicate 20 equally spaced time points in the gait cycle. The letters are plotted at intervals of one fifth of the cycle with A marking the heel strike.

1.2.2 Data on Handwriting

Multivariate functional data often arise from tracking the movements of points through space, as illustrated in Figure 1.8, where the X-Y coordinates of 20 samples of handwriting are superimposed. The role of time is lost in plots such as these, but can be recovered to some extent by plotting points at regular time intervals.

Figure 1.9 shows the first sample of the writing of “statistical science” in simplified Chinese with gaps corresponding to the pen being lifted off the paper. Also plotted are points at 120-millisecond intervals; many of these points seem to coincide with points of sharp curvature and the ends of strokes.

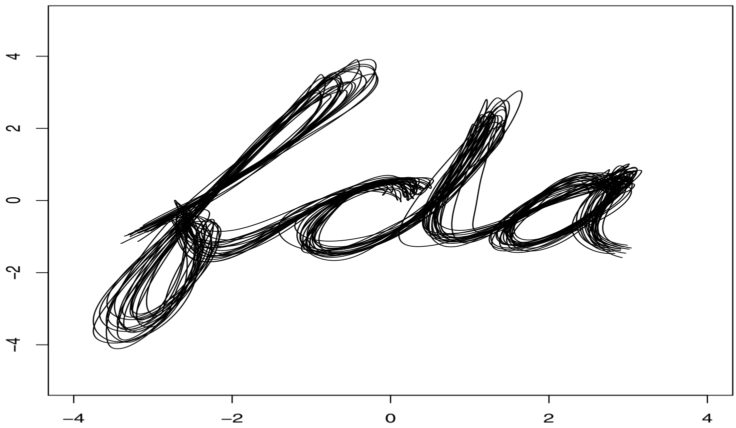

Fig. 1.8 Twenty samples of handwriting. The axis units are centimeters.

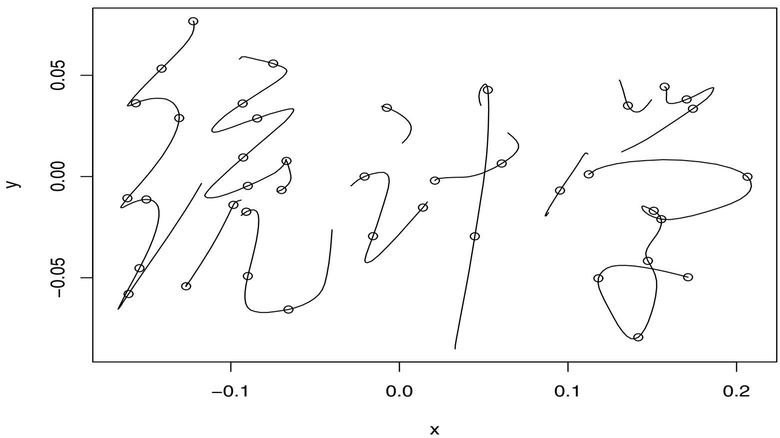

Fig. 1.9 The first sample of writing "statistical science" in simplified Chinese. The plotted points correspond to 120-millisecond time steps.

Finally, in this introduction to types of functional data, we must not forget that they may come to our attention as full-blown functions, so that each record may consist of functions observed, for all practical purposes, everywhere. Sophisticated online sensing and monitoring equipment now routinely used in research in fields such as medicine, seismology, meteorology and physiology can record truly functional data.

1.3 Functional Models for Nonfunctional Data

The data examples above seem to deserve the label "functional" since they so clearly reflect the smooth curves that we assume generated them. Beyond this, functional data analysis tools can be used for many data sets that are not so obviously functional.

Consider the problem of estimating a probability density function p to describe the distribution of a sample of observations $x_1,\ldots,x_n$. The classic approach to this problem is to propose, after considering basic principles and closely studying the data, a *parametric model* with values $p(x|\theta)$ defined by a fixed and usually small number of parameters in the vector θ. For example, we might consider the normal distribution as appropriate for the data, so that $\theta = (\mu,\sigma^2)'$. The parameters themselves are usually chosen to be descriptors of the shape of the density, as in location and spread for the normal density, and are therefore the focus of the analysis.

But suppose that we do not want to assume in advance one of the many textbook density functions. We may feel, for example, that the application cannot justify the assumptions required for using any of the standard distributions. Or we may see features in histograms and other graphical displays that seem not to be captured by any of the most popular distributions. *Nonparametric density* estimation methods assume only smoothness, and permit as much flexibility in the estimated $p(x)$ as the data require or the data analyst desires. To be sure, parameters are often involved, as in the density estimation method of Chapter 5, but the number of parameters is not fixed in advance of the data analysis, and our attention is focused on the density function p itself, not on parameter estimates. Much of the technology for estimation of smooth *functional parameters* was originally developed and honed in the density estimation context, and Silverman (1986) can be consulted for further details.

Psychometrics or mental test theory also relies heavily on functional models for seemingly nonfunctional data. The data are usually zeros and ones indicating unsuccessful and correct answers to test items, but the model consists of a set of *item response functions*, one per test item, displaying the smooth relationship between the probability of success on an item and a presumed latent ability continuum. Figure 1.10 shows three such functional parameters for a test of mathematics estimated by the functional data analytic methods reported in Rossi et al. (2002).

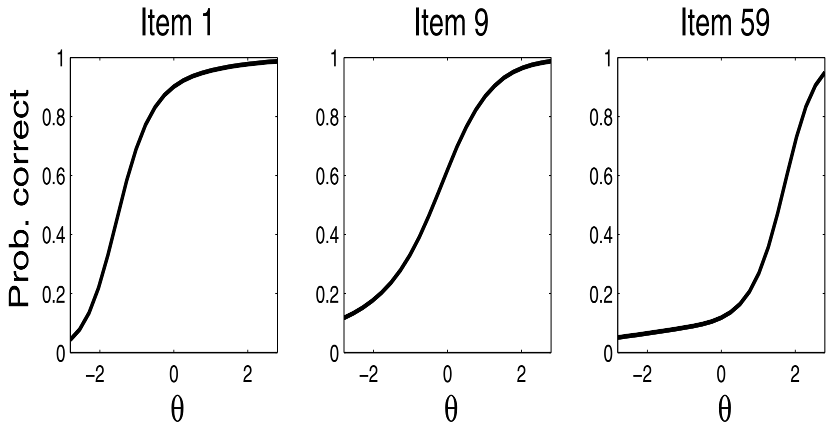

Fig. 1.10 Each panel shows an item response function relating an examinee's position θ on a latent ability continuum to the probability of a correct response to an item in a mathematics test.

1.4 Some Functional Data Analyses

Data in many fields come to us through a process naturally described as functional. Consider Figure 1.11, where the mean temperatures for four Canadian weather stations are plotted as smooth curves. Montreal, with the warmest summer temperature, has a temperature pattern that appears to be nicely sinusoidal. Edmonton, with the next warmest summer temperature, seems to have some distinctive departures from sinusoidal variation that might call for explanation. The marine climate of Prince Rupert is evident in the small amount of annual variation in temperature. Resolute has bitterly cold but strongly sinusoidal temperatures.

One expects temperature to be periodic and primarily sinusoidal in character and over the annual cycle. There is some variation in the timing of the seasons or phase, because the coldest day of the year seems to be later in Montreal and Resolute than in Edmonton and Prince Rupert. Consequently, a model of the form

$$\texttt{Temp}_i(t) \approx c_{i1} + c_{i2}\sin(\pi t/6) + c_{i3}\cos(\pi t/6) \tag{1.1}$$

should do rather nicely for these data, where $\texttt{Temp}_i$ is the temperature function for the ith weather station, and (c_{i1}, c_{i2}, c_{i3}) is a vector of three parameters associated with that station.

In fact, there are clear departures from sinusoidal or simple harmonic behavior. One way to see this is to compute the function

$$L\texttt{Temp} = (\pi/6)^2 D\texttt{Temp} + D^3\texttt{Temp}. \tag{1.2}$$

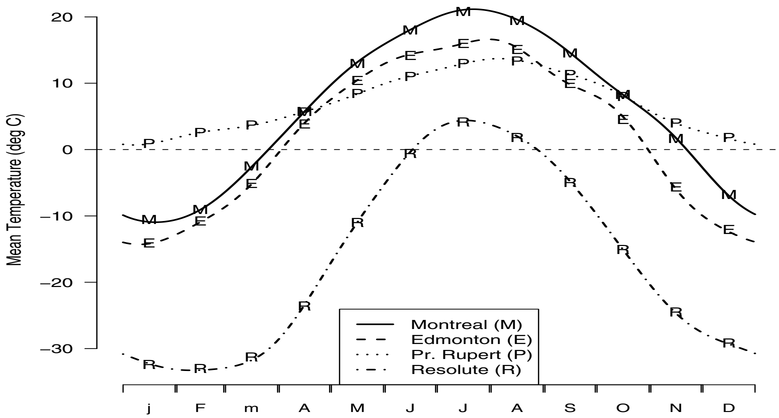

Fig. 1.11 Mean temperatures at four Canadian weather stations.

The notation D^m`Temp` means "take the mth derivative of function `Temp`," and the notation L`Temp` stands for the function which results from applying the *linear differential operator* $L = (\pi/6)^2 D + D^3$ to the function `Temp`. The resulting function, L`Temp`, is often called a *forcing function.* If a temperature function is truly sinusoidal, then L`Temp` should be exactly zero, as it would be for any function of the form (1.1). That is, it would conform to the *differential equation*

$$L\texttt{Temp} = 0 \text{ or } D^3\texttt{Temp} = -(\pi/6)^2 D\texttt{Temp}.$$

But Figure 1.12 indicates that the functions $L\texttt{Temp}_i$ display systematic features that are especially strong in the summer and autumn months. Put another way, temperature at a particular weather station can be described as the solution of the *nonhomogeneous* differential equation corresponding to $L\texttt{Temp} = u$, where the forcing function u can be viewed as input from outside of the system, or as an exogenous influence. Meteorologists suggest, for example, that these spring and autumn effects are partly due to the change in the reflectance of land when snow or ice melts, and this would be consistent with the fact that the least sinusoidal records are associated with continental stations well separated from large bodies of water.

Here, the point is that we may often find it interesting to remove effects of a simple character by applying a differential operator, rather than by simply subtracting them. This exploits the intrinsic smoothness in the process. Long experience in the natural and engineering sciences suggests that this may get closer to the underlying driving forces at work than just adding and subtracting effects, as is routinely done in multivariate data analysis. We will consider this idea in depth in Chapter 11.

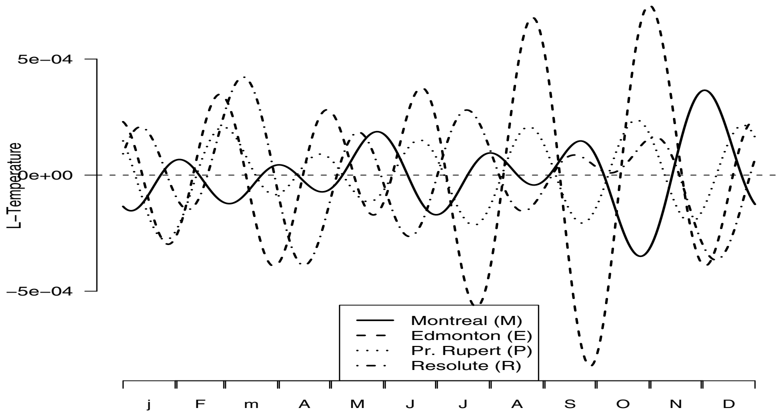

Fig. 1.12 The result of applying the differential operator $L = (\pi/6)^2 D + D^3$ to the estimated temperature functions in Figure 1.11. If the variation in temperature were purely sinusoidal, these curves would be exactly zero.

1.5 First Steps in a Functional Data Analysis

1.5.1 Data Representation: Smoothing and Interpolation

Assuming that a functional datum for replication i arrives as a finite set of measured values, $y_{i1}, \ldots, y_{in}$, the first task is to convert these values to a function x_i with values $x_i(t)$ computable for any desired argument value t. If these observations are assumed to be errorless, then the process is *interpolation*, but if they have some observational error that needs removing, then the conversion from (finite) data to functions (which can theoretically be evaluated at an infinite number of points) may involve *smoothing.*

Chapter 5 offers a survey of these procedures. The *roughness penalty* smoothing method discussed there will be used much more broadly in many contexts throughout the book, and not merely for the purpose of estimating a function from a set of observed values. The daily precipitation data for Prince Rupert, one of the wettest places on the continent, is shown in Figure 1.13. The curve in the figure, which seems to capture the smooth variation in precipitation, was estimated by penalizing the squared deviations in *harmonic acceleration* as measured by the differential operator (1.2).

The gait data in Figure 1.6 were converted to functions by the simplest of interpolation schemes: joining each pair of adjacent observations by a straight line segment. This approach would be inadequate if we required derivative information. However,