13

Modelling functional responses with multivariate covariates

13.1 Introduction

We now consider how to use data on a set of scalar predictor variables or covariates $z_j, j = 1, \ldots, p$ to fit the features of a functional response or outcome variable x. Both of the examples in this chapter can be described as *functional analyses of variance* because the values of the covariates are 0's and 1's coding the categories of factor variables, but the techniques that we develop here apply equally well to measured covariates.

13.2 Predicting temperature curves from climate zones

Let's have a look at the Canadian weather data introduced in Chapter 1. Monthly means for temperature and precipitation are available for each of 35 weather stations distributed across the country, and we can use the smoothing techniques of Chapters 4 and 5 to represent each record as a smooth function. Thus, two periodic functions, `Temp` and `Prec`, denoting temperature and precipitation, respectively, are available for each station.

How much of the pattern of annual variation of temperature in a weather station is explainable by its geographical area? Dividing Canada into Atlantic, Continental, Pacific and Arctic meteorological zones, we want to study the characteristic types of temperature patterns in each zone.

This is an *analysis of variance* problem with four treatment groups. *Multivariate analysis of variance* (MANOVA) is the extension of the ideas of analysis of variance to deal with problems where the dependent variable is multivariate. Because our dependent variable is the functional observation `Temp`, the methodology we need is a *functional analysis of variance*, abbreviated FANOVA.

In formal terms, we have a number of stations in each group g, and the model for the mth temperature function in the gth group, indicated by $\texttt{Temp}_{mg}$, is

$$\texttt{Temp}_{mg}(t) = \mu(t) + \alpha_g(t) + \epsilon_{mg}(t). \tag{13.1}$$

The function μ is the grand mean function, and therefore indicates the average temperature profile across all of Canada. The terms α_g are the specific effects on temperature of being in climate zone g. To be able to identify them uniquely, we require that they satisfy the constraint

$$\sum_g \alpha_g(t) = 0 \text{ for all } t. \tag{13.2}$$

The residual function ϵ_{mg} is the unexplained variation specific to the mth weather station within climate group g.

We note in passing that the smoothing problem discussed in Chapters 3, 4, and 5 is contained within this model by using a single covariate whose values are all ones.

We can define a 35×5 design matrix $\mathbf{Z}$ for this model, with one row for each individual weather station, as follows. Use the label (mg) for the row corresponding to station m in group g; this row has a one in the first column, a one in column $g+1$, and zeroes in the rest. Write $z_{(mg)j}$ for the value in this row and in the jth column of $\mathbf{Z}$.

We can then define a corresponding set of five regression functions β_j by setting $\beta_1 = \mu$, $\beta_2 = \alpha_1$, and so on to $\beta_5 = \alpha_4$, so that the functional vector $\boldsymbol{\beta} = (\mu, \alpha_1, \alpha_2, \alpha_3, \alpha_4)'$. In these terms, the model (13.1) has the equivalent formulation

$$\texttt{Temp}_{mg}(t) = \sum_{j=1}^{5} z_{(mg)j}\beta_j(t) + \epsilon_{mg}(t) \tag{13.3}$$

or, more compactly in matrix notation,

$$\texttt{Temp} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{13.4}$$

where `Temp` is the functional vector containing the 35 temperature functions, $\boldsymbol{\epsilon}$ is a vector of 35 residual functions, and $\boldsymbol{\beta}$ is the 5-vector of parameter functions. The design matrix $\mathbf{Z}$ has exactly the same structure as for the corresponding univariate or multivariate one-way analysis of variance. The only way in which (13.4) differs from the corresponding equations in standard elementary textbooks on the general linear model is

that the parameter $\boldsymbol{\beta}$, and hence the predicted observations $\mathbf{Z}\boldsymbol{\beta}$, are vectors of functions rather than vectors of numbers.

13.2.1 Fitting the model

If (13.4) were a standard general linear model, the standard least squares criterion would say that $\boldsymbol{\beta}$ should be chosen to minimize the residual sum of squares. To extend the least squares principle to the functional case, we need only reinterpret the residual sum of squares in an appropriate way. The quantity $\texttt{Temp}_i(t) - \mathbf{Z}_i\boldsymbol{\beta}(t)$ is now a function, and so the unweighted least squares fitting criterion becomes

$$\texttt{LMSSE}(\boldsymbol{\beta}) = \sum_{g}^{4} \sum_{m}^{N_g} \int [\texttt{Temp}_{mg}(t) - \sum_{j}^{q} z_{(mg),j}\beta_j(t)]^2 \, dt. \tag{13.5}$$

Minimizing $\texttt{LMSSE}(\boldsymbol{\beta})$ subject to the constraint $\sum_2^5 \beta_j = 0$ (equivalent to $\sum_1^4 \alpha_g = 0$) gives the least squares estimates $\hat{\boldsymbol{\beta}}$ of the functional parameters μ and α_g. Section 13.4 contains some remarks about the way LMSSE is minimized in practice.

Figure 13.1 displays the resulting estimated region effects α_g for the four climatic zones, and Figure 13.2 displays the composite effects $\mu + \alpha_g$. We see that the region effects are more complex than the constant or even sinusoidal effects that one might expect:

- The Atlantic stations appear to have a temperature around 5 degrees C warmer than the Canadian average.
- The Pacific weather stations have a summer temperature close to the Canadian average, but are much warmer in the winter.
- The Continental stations are slightly warmer than average in the summer, but are colder in the winter by about 5 degrees C.
- The Arctic stations are certainly colder than average, but even more so in March than in January.

The cross-hatched areas indicate 95% confidence regions for the location of the curves at fixed points. These will be discussed in Section 13.4.

13.2.2 Assessing the fit

In estimating and plotting the individual regional temperature effects, we have taken our first step towards achieving the goal of characterizing the typical temperature pattern for weather stations in each climate zone. We may wish to move on and not only confirm that the total zone-specific effect α_g is nonzero, but also investigate whether this effect is substantial at

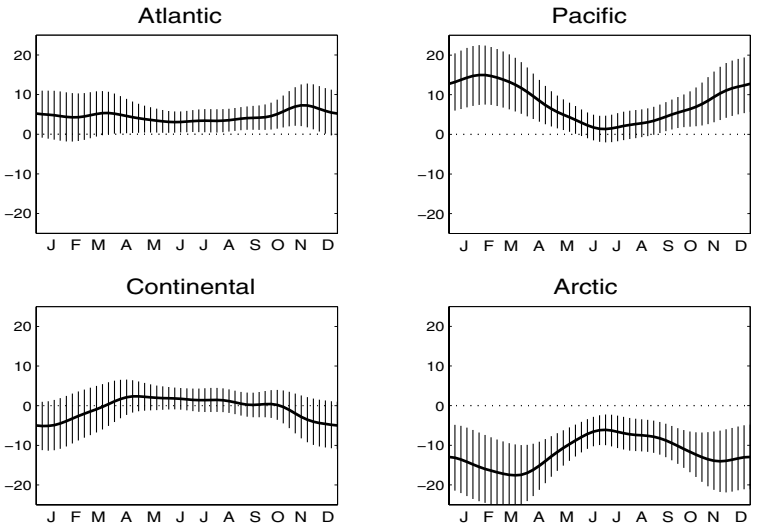

Figure 13.1. The region effects α_g for the temperature functions in the functional analysis of variance model $\texttt{Temp}_{mg}(t) = \mu(t) + \alpha_g(t) + \epsilon_{mg}(t)$. The effects $\alpha_g(t)$ are required to sum to 0 for all t. The cross-hatched areas indicate 95% point-wise confidence intervals for the true effects.

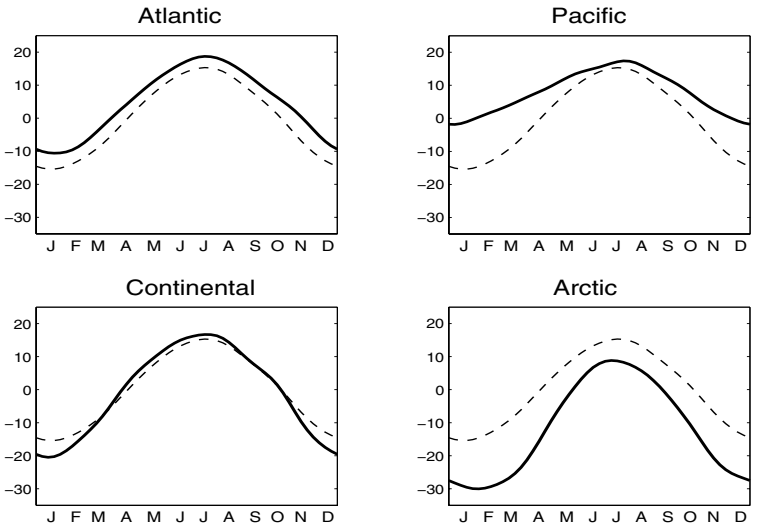

Figure 13.2. The estimated climate zone temperature profiles $\mu + \alpha_g$ for the temperature functions in the functional analysis of variance model (solid curves). The dashed curve is the Canadian mean function μ.

a specific time t. As in ordinary analysis of variance, we look to summarize these issues in terms of error sum of squares functions LMSSE, squared correlation functions RSQ, and F-ratio functions FRATIO. It is the dependence of these quantities on t that makes the procedure different from the standard multivariate case.

As in the multivariate linear model, the primary source of information in investigating the importance of the zone effects α_g is the sum of squares function

$$\mathtt{SSE}(t) = \sum_{mg} [\mathtt{Temp}_{mg}(t) - \mathbf{Z}_{mg}\hat{\boldsymbol{\beta}}(t)]^2. \tag{13.6}$$

This function can be compared to the error sum of squares function based on using only the Canadian average $\hat{\mu}$ as a model,

$$\mathtt{SSY}(t) = \sum_{mg} [\mathtt{Temp}_{mg}(t) - \hat{\mu}(t)]^2$$

and one way to make this comparison is by using the squared multiple correlation function RSQ with values

$$\mathtt{RSQ}(t) = [\mathtt{SSY}(t) - \mathtt{SSE}(t)]/\mathtt{SSY}(t). \tag{13.7}$$

Essentially, this function considers the drop in error sum of squares produced by taking climate zone into effect relative to error sum of squares without using climate zone information.

We can also compute the functional analogues of the quantities entered into the ANOVA table for a univariate analysis. For example, the mean squared for error function MSE has values

$$\mathtt{MSE} = \mathtt{SSE}/\mathtt{df}(\text{error}),$$

where df(error) is the degrees of freedom for error, or the sample size N less the number of mathematically independent functions β_q in the model. In this problem, the zero sum restriction on the climate zone effects α_g implies that there are four degrees of freedom lost to the model, or df(error) = 31.

Similarly, the mean square for regression is the difference between SSY (or, more generally, whatever reference model we employ that is a specialization of the model being assessed) and SSE, divided by the difference between the degrees of freedom for error for the two models. Let the difference in degrees of freedom be denoted by df(regression), which in this case is 3. Thus

$$\mathtt{MSR}(t) = \frac{\mathtt{SSY}(t) - \mathtt{SSE}(t)}{\mathtt{df}(\text{regression})} .$$

Finally, we can compute the F-ratio function,

$$\mathtt{FRATIO} = \frac{\mathtt{MSR}}{\mathtt{MSE}}. \tag{13.8}$$

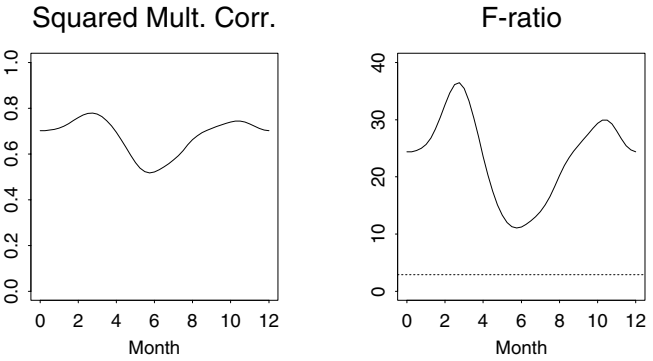

Figure 13.3. The left panel contains the squared multiple correlation function `RSQ` and the right panel the corresponding F-ratio function `FRATIO`. The horizontal dotted line indicates the 5% significance level for the F-distribution with 3 and 31 degrees of freedom.

Figure 13.3 shows the two functions `RSQ` and `FRATIO`. We can see that the squared correlation is relatively high and that the F-ratio is everywhere substantially above the 5% significance level of 2.92. It is interesting to note that the differences between the climate zones are substantially stronger in the spring and autumn, rather than in the summer and winter as we might expect.

Basically, then, most of the statistical machinery available for univariate analysis of variance is readily applicable to this functional problem. We can consider, for example, contrast functions, post-hoc multiple comparison functions, F-ratios associated with constrained estimates of region effects, and so on, essentially because the functional analysis of variance problem is really a univariate ANOVA problem for each specific value of t.

One question not addressed in the discussion of this example is an overall assessment of significance for the difference between the climate zones, rather than an assessment for each individual time t. We remind ourselves that the classical significance level was designed to be used for a single hypothesis test, rather than a continuum of them as in here. Although there is no reasonable doubt here that climate zone has an important effect somewhere in the year, in other applications we will want to protect ourselves more effectively against falsely declaring significance somewhere in the interval. Section 13.3.3 provides an approach to this question using simulation in the context of a different example.

A second question is, "How can we compute confidence intervals for the estimated regression functions?" Because this topic involves substantial mathematical detail, we put this off until Section 13.4.

13.3 Force plate data for walking horses

This section describes some interesting data on equine gait. The data were collected by Dr. Alan Wilson of the Equine Sports Medicine Center, Bristol University, and his collaborators. Their kindness in allowing use of the data is gratefully acknowledged. The data provide an opportunity to discuss various extensions of our functional linear modelling and analysis of variance methodology. For further details of this example, see Wilson et al. (1996).

13.3.1 Structure of the data

The basic structure of the data is as follows. It is of interest to study the effects of various types of shoes, and various walking surfaces, on the gait of a horse. One reason for this is simply biomechanical: the horse is an animal particularly well adapted to walking and running, and the study of its gait is of intrinsic scientific interest. Secondly, it is dangerous to allow horses to race if they are lame or likely to go lame. Careful study of their gait may produce diagnostic tests of incipient lameness which do not involve any invasive investigations and may detect injuries at a very early stage, before they become serious or permanent. Thirdly, it is important to shoe horses to balance their gait, and understanding the effects of different kinds of shoe is necessary to do this. Indeed, once the normal gait of a horse is known, the measurements we describe can be used to test whether a blacksmith has shod a horse correctly, and can therefore be used as an aid in the training of farriers.

In this experiment, horses walk on to a plate about 1 meter square set into the ground and equipped with meters at each corner measuring the force in the vertical and the two horizontal directions. We consider only the vertical force. During the period that the horse's hoof is on the ground (the stance phase) the four measured vertical forces allow the instrument to measure the point of resultant vertical force. The hoof itself does not move during the stance phase, and the position of the hoof is measured by dusting the plate with sawdust or is inferred from the point of force at the end of the stride, when only the front tip of the hoof is in contact with the ground.

The vertical force increases very rapidly at the beginning of the stance phase but reduces more slowly at the end. Operationally, the stance phase is defined as starting at the moment where the total vertical force first reaches 30% of its maximum value and ending where it falls to 8% of its maximum value. For each replication, the point of force is computed for 100 time points equally spaced in this time interval.

A typical functional observation is therefore a two-dimensional function of time $\texttt{Force} = (\texttt{ForceX}, \texttt{ForceY})$ where t varies from 0 to 1 during the stance phase, and $\texttt{ForceX}(t)$ and $\texttt{ForceY}(t)$ are the coordinates of the point

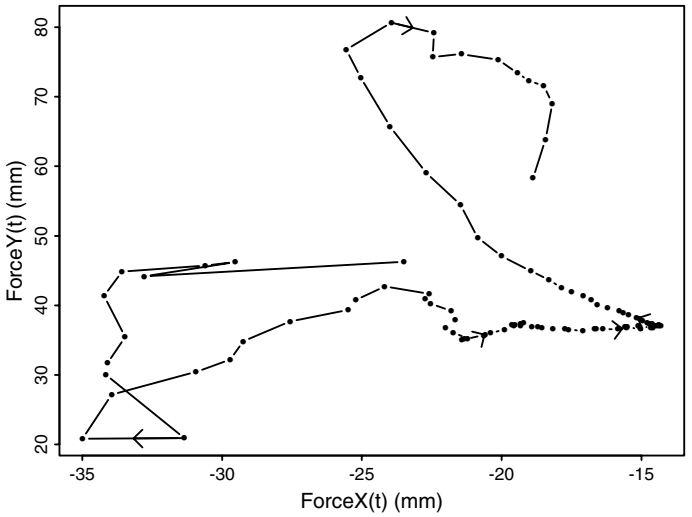

Figure 13.4. A typical trace of the resultant point of force during the stance phase of a horse walking onto a force plate. One hundred points equally spaced in time are indicated on the curve. The arrows indicate the direction of time.

of force at time t. Here Y is the direction of motion of the horse, and X measures distance in a perpendicular direction towards the body of the horse. Thus the coordinates are defined as if looking at the plate from above if a left foot is being measured, but with the X direction reflected if a right foot is being measured.

The data set consists of 592 separate runs and involves 8 horses, each of which has a number of measurements on both its right and left forelimbs. The nine shoeing conditions are as follows: first, the horse is observed unshod; it is then shod and observed again; then its shoe is modified by the addition of various wedges, either building up its toe or heel or building up one side or the other of its hoof. Not every horse has every wedge applied. In the case of the toe and heel wedges, the horse is observed immediately after the wedge is fitted and one day later, after it has become accustomed to the shoe. Finally the wedges are removed and the horse is observed with a normal shoe.

Figure 13.4 shows a typical (`ForceX`, `ForceY`) plot. This realization is among the smoother curves obtained. The 100 points that are equally spaced in time are marked on the curve, and the direction of time is indicated by arrows (also evenly spaced in time). We can see, not surprisingly, that the point of force moves most rapidly near the beginning and end of

the stance phase. The accuracy of a point measurement was about 1 mm in each direction.

13.3.2 A functional linear model for the horse data

The aim of this experiment is to investigate the effects of various shoeing conditions, and particularly to study the effects of the toe and heel wedges, which change as the horse becomes accustomed to the wedge. We fit a model of the form

$$\texttt{Force}_{ijkl} = \mu + \alpha_{ij} + \theta_k + \epsilon_{ijkl}, \tag{13.9}$$

where all the terms are two-dimensional functions of t, $0 \leq t \leq 1$. The suffix $ijkl$ refers to the data collected for the lth observed curve for side j of horse i under condition k.

For any particular curve, use labels x and y where necessary to denote the x and y coordinates of the vector function. The following identifiability constraints are placed on the various effects, each valid for all t:

$$\sum_{i,j} \alpha_{ij}(t) = \sum_{k=1}^{9} \theta_k(t) = 0. \tag{13.10}$$

We estimate the various effects by carrying out a separate general linear model fit for each t and for each of the x and y coordinates. Since the data are observed at 100 discrete times in practice, each $\texttt{Force}_{ijkl}$ corresponds to two vectors, each of length 100, one for the x coordinates and one for the y coordinates. The design matrix relating the expected value of $\texttt{Force}_{ijkl}$ to the various effects is the same for all 200 observed values, so although the procedure involves the fitting of 200 separate models, considerable economy of effort is possible. The model (13.9) can be written as

$$\texttt{Force} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{13.11}$$

where Force and $\boldsymbol{\epsilon}$ are both vectors of length 592, each of whose elements is a two-dimensional function on $[0, 1]$. The vector $\boldsymbol{\beta}$ is a vector of the 26 two-dimensional functions μ, α_{ij} and θ_k, and $\mathbf{Z}$ is a 592×26 design matrix relating the observations Force to the effects $\boldsymbol{\beta}$. The identifiability constraints (13.10) are incorporated by augmenting the matrix $\mathbf{Z}$ by additional rows corresponding to the constraints, and by augmenting the data vector Force by zeroes. Standard theory of the general linear model of course then gives as the estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\,\texttt{Force}. \tag{13.12}$$

Figure 13.5 plots the estimated overall mean curve $\mu = (\mu_x, \mu_y)$ in the same way as Figure 13.4. Although the individual observations are somewhat irregular, the overall mean is smooth, even though no smoothing is incorporated into the procedure.

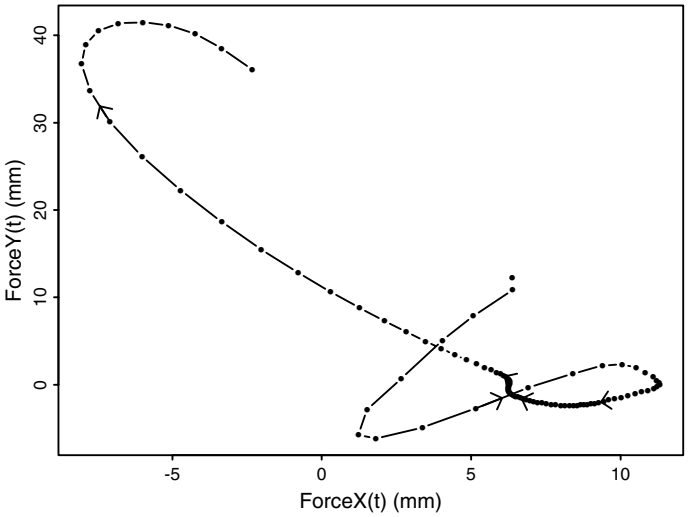

Figure 13.5. Estimate of the overall mean curve (μ_x, μ_y) obtained from the 592 observed point-of-force curves using model (13.9).

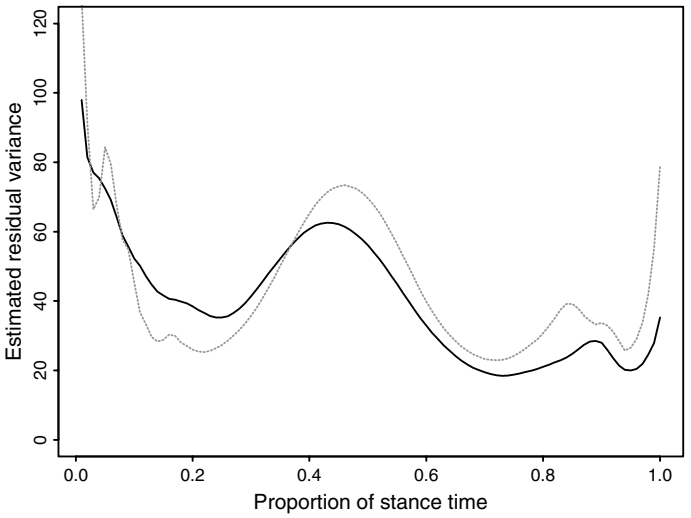

Figure 13.6. The estimated residual variance in the x coordinate (solid curve) and the y coordinate (dotted curve) as the stance phase progresses.

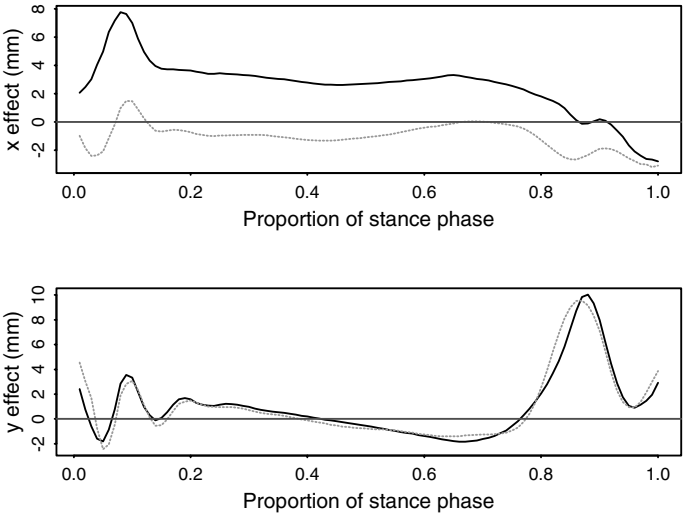

Figure 13.7. The effects of the application of a toe wedge, x coordinate in the upper panel and y coordinate in the lower. Solid curves are the immediate effect, and dashed curves are the effect on the following day.

The general linear model fitted for each coordinate at each time point allows the calculation of a residual sum of squares, and hence an estimated residual variance, at each point. The residual variance curves $\texttt{MSE}_x$ and $\texttt{MSE}_y$ for the x and y coordinates are plotted in Figure 13.6. It is very interesting to note that the residual variances in the two coordinates are approximately the same size, and vary in roughly the same way, as the stance phase progresses.

13.3.3 Effects and contrasts

We can now explain how the linear model can be used to investigate various effects of interest. We concentrate on two specific effects, corresponding to the application of the toe wedges, and illustrate how various inferences can be drawn. In Figure 13.7, we plot the effects of the toe wedge immediately after it has been applied and the following day. The x and y coordinates of the relevant functions θ_k are plotted separately. It is interesting to note that the y effects are virtually the same in both cases: The application of the wedge has an immediate effect on the way in which the point of force moves in the forward-backward direction, and this pattern does not change appreciably as the horse becomes accustomed to the wedge. The effect in the side-to-side direction is rather different. Immediately after the wedge

is applied, the horse tends to put its weight to one side, but the following day the effect becomes much smaller, and the weight is again placed in the same lateral position as in the average stride.

To investigate the significance of this change, we now consider the contrast between the two effects, which shows the expected difference between the point of force function for a horse 24 hours after a toe wedge has been applied and that immediately after applying the wedge. Figure 13.8 shows the x and y coordinates of the difference of these two effects. The standard error of this contrast is easily calculated. Let u be the vector such that the estimated contrast is the vector function $\texttt{Contrast} = u'\hat{\boldsymbol{\beta}}$, so that the component of u corresponding to toe wedge 24 hours after application is $+1$, that corresponding to toe wedge immediately after application is -1, and all the other components are zero. Define a by $a^2 = u'(\mathbf{Z}'\mathbf{Z})^{-1}u$. The squared point-wise standard errors of the x and y coordinates of the estimated contrasts are then $a^2\texttt{MSE}_x$ and $a^2\texttt{MSE}_y$, respectively. Plots of ± 2 times the relevant standard error are included in Figure 13.8. Because the degrees of freedom $(592 - 26 + 2)$ for residual variance are so large, these plots indicate that point-wise t tests at the 5% level would demonstrate that the difference in the y coordinate of the two toe wedge effects is not significant, except possibly just above time 0.8, but that the x coordinate is significantly different from zero for almost the whole stance phase.

How should we account for the correlation in the tests at different times in assessing the significance of any difference between the two conditions? We can consider the summary statistics

$$M_x = \sup_t |\texttt{Contrast}_x(t)/a\sqrt{\texttt{MSE}_x(t)}|$$

and

$$M_y = \sup_t |\texttt{Contrast}_y(t)/a\sqrt{\texttt{MSE}_y(t)}|.$$

The values of these statistics for the data are $M_x = 5.03$ and $M_y = 2.01$. A permutation-based significance value for each of these statistics was obtained by randomly permuting the observed toe wedge data for each leg of each horse between the conditions immediately after fitting of wedge and 24 hours after fitting of wedge, keeping the totals the same within each condition for each leg of each horse. The statistics M_x and M_y were calculated for each random permutation of the data. In 1000 realizations, the smallest value of M_x observed was 3.57, so the observed difference in the x direction of the two conditions is highly significant. A total of 177 of the 1000 simulated M_y values exceeded the observed value of 2.01, and so the estimated p-value of this observation was 0.177, showing no evidence that the y coordinate of point of force alters its time behavior as the horse becomes accustomed to the wedge.

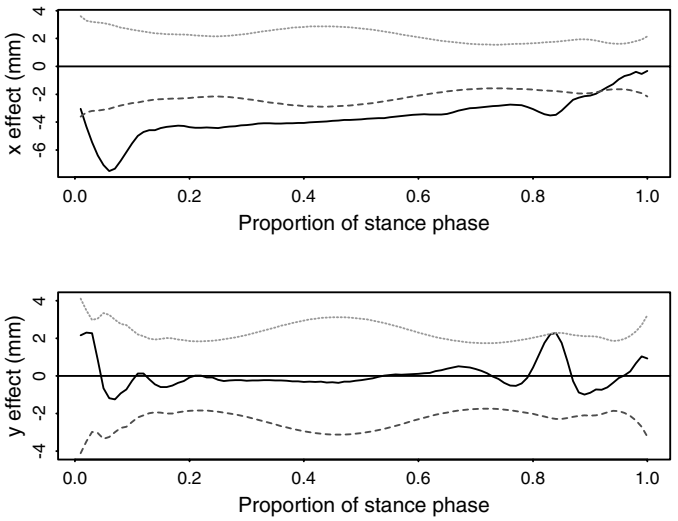

Figure 13.8. Solid curves are the differences between the effect of a toe wedge after 24 hours and its immediate effect. Dotted curves indicate plus and minus two estimated standard errors for the point-wise difference between the effects. The upper panel contains x coordinates and the lower the y coordinates.

13.4 Computational issues

13.4.1 The general model

We explain how to compute the least squares estimates in the functional linear model. Let Y be an N-vector of functional observations, and let the q-vector $\boldsymbol{\beta}$ contain the regression functions. In the force-plate data example, the individual elements of both Y and $\boldsymbol{\beta}$ were themselves two-dimensional functions. We assume that $\mathbf{Z}$ is an $N \times q$ design matrix, and that the expected value of $\mathbf{y}(t)$ for each t is modelled as $\mathbf{Z}\boldsymbol{\beta}(t)$. The functional linear model is then

$$\mathbf{y}(t) = \mathbf{Z}\boldsymbol{\beta}(t) + \boldsymbol{\epsilon}(t) \ . \tag{13.13}$$

Any linear constraints on the parameters $\boldsymbol{\beta}$, such as the requirement in the temperature data example that the individual climate zone effects sum to zero, are expressed as $\mathbf{L}\boldsymbol{\beta} = 0$ for some suitable matrix $\mathbf{L}$ with q columns. By using a technique such as the QR-decomposition, described in Section A.3.3 of the Appendix, we may then say that

$$\boldsymbol{\beta} = \mathbf{C}\alpha \tag{13.14}$$

for some matrix $\mathbf{C}$. In this case we find ourselves back at the basic model (13.13), except that $\mathbf{Z}$ is now replaced by $\mathbf{ZC}$ and $\boldsymbol{\beta}$ by $\boldsymbol{\alpha}$.

Our aim is to minimize the least squares fitting criterion

$$\texttt{LMSSE}(\boldsymbol{\beta}) = \int [\mathbf{y}(t) - \mathbf{Z}\boldsymbol{\beta}(t)]'[\mathbf{y}(t) - \mathbf{Z}\boldsymbol{\beta}(t)]\, dt \ , \tag{13.15}$$

It is possible that an order N weighting matrix $\mathbf{W}$ should be included in this expression between the two factors on the right in order to allow for possible variation of importance across replications, for dependencies. However, most situations assume independence and constant error variation, and so we will not bother with this feature in our discussion.

13.4.2 Pointwise minimization

If there are no particular restrictions on the way in which $\boldsymbol{\beta}(t)$ varies as a function of t, we can minimize $\texttt{LMSSE}(\boldsymbol{\beta})$ by minimizing $\|\mathbf{y}(t) - \mathbf{Z}\boldsymbol{\beta}(t)\|^2$ individually for each t. That is, we calculate $\hat{\boldsymbol{\beta}}(t)$ for a suitable grid of values of t using ordinary regression analysis, and then interpolate between these values. This was the technique used in the force plate example, where the grid of values was chosen to correspond with the discretization of the original data. The fact that the same design matrix is involved for each t makes for considerable economy of numerical effort.

13.4.3 Functional linear modelling with regularized basis expansions

We have already noted, however, that the use of regularized basis function expansions gives us continuous control over smoothness while still permitting as much high frequency detail in the model as the data require. The use of roughness penalties or regularization can play an important role in a functional linear model. In particular, one may adopt the philosophy that the representation of the response functions should be allowed to be of high resolution, and that smoothness is imposed only on the functional parameters to be estimated in $\boldsymbol{\beta}$. In this way, we do not risk smoothing away important information that may impact the estimate of $\boldsymbol{\beta}$ when smoothing the data giving rise to y.

Let us now assume that the observed functions y_i and regression functions β_j are expressed in basis expansion form, as the coefficients of a Fourier series or B-spline or some other basis system. This means that

$$\mathbf{y}(t) = \mathbf{C}\boldsymbol{\phi}(t),$$

where

- the N-vector $\mathbf{y}$ contains the N observed response functions,

- the K_y-vector $\boldsymbol{\phi}$ contains the linearly independent basis functions, and
- the N by K_y matrix matrix $\mathbf{C}$ contains the coefficients of expansion of function y_i in its ith row.

Let us now expand the estimated parameter vector $\hat{\boldsymbol{\beta}}$ in terms of a basis vector $\boldsymbol{\theta}$ of length K_β, expressing $\hat{\boldsymbol{\beta}} = \mathbf{B}\boldsymbol{\theta}$ for a $q \times K_\beta$ matrix $\mathbf{B}$. In some cases, we may choose to use the same basis that was used to expand the response functions, in which case $\boldsymbol{\theta} = \boldsymbol{\phi}$, and consequently some of what follows becomes simpler. However, there are plenty of situations where we need to keep the two basis systems separate.

Note, though, that we have made things somewhat easier on ourselves by assuming that the same basis system $\boldsymbol{\theta}$ is used for all q regression functions β_j. In the next chapter, we will relax this constraint, but for the time being this assumption has the advantage of keeping the notation reasonably simple.

Now suppose that we use a linear differential operator L to define a roughness penalty for $\boldsymbol{\beta}$ as

$$\texttt{PEN}_L(\beta) = \int [L\boldsymbol{\beta}(s)]'[L\boldsymbol{\beta}(s)]\, ds \ . \tag{13.16}$$

In addition, we need to define these four matrices:

$$\mathbf{J}_{\phi\phi} = \int \boldsymbol{\phi}\boldsymbol{\phi}' \ , \quad \mathbf{J}_{\theta\theta} = \int \boldsymbol{\theta}\boldsymbol{\theta}' \ , \quad \mathbf{J}_{\phi\theta} = \int \boldsymbol{\phi}\boldsymbol{\theta}'$$

$$\mathbf{R} = \int (L\boldsymbol{\theta})(L\boldsymbol{\theta})' \ . \tag{13.17}$$

Note that we dropped "(s)" and "ds" from the expressions in (13.17); this makes the expressions more readable, and the context makes it clear that what we really mean is an expression like (13.16) where they were left in. Note, too, that since $\boldsymbol{\phi}$ is a column vector of length K_y of basis functions, $\boldsymbol{\phi}\boldsymbol{\phi}'$ is a square matrix of order K_y containing all possible pairs of these functions, and consequently $\mathbf{J}_{\phi\phi}$ is constant symmetric order K_y matrix of integrated products, and similarly for the other three matrices.

We can then obtain the following expressions for the penalized least squares criterion:

$$\begin{aligned} \texttt{PENSSE}(y|\boldsymbol{\beta}) \quad &= \quad \int (\mathbf{C}\boldsymbol{\phi} - \mathbf{Z}\mathbf{B}\boldsymbol{\theta})'(\mathbf{C}\boldsymbol{\phi} - \mathbf{Z}\mathbf{B}\boldsymbol{\theta}) + \\ &\qquad \lambda \int (L\mathbf{B}\boldsymbol{\theta})'(L\mathbf{B}\boldsymbol{\theta}) \ . \end{aligned} \tag{13.18}$$

The operation of integration and the summations implied by the matrix products in these expressions can be interchanged, and consequently we

can re-expressing this as

$$\begin{aligned}\texttt{PENSSE}(y|\boldsymbol{\beta}) &= \operatorname{trace}(\mathbf{C}'\mathbf{C}\mathbf{J}_{\phi\phi}) + \operatorname{trace}(\mathbf{Z}'\mathbf{Z}\mathbf{B}\mathbf{J}_{\theta\theta}\mathbf{B}') - \\ &\quad 2\operatorname{trace}(\mathbf{B}\mathbf{J}_{\theta\phi}\mathbf{C}'\mathbf{Z}) + \lambda\operatorname{trace}(\mathbf{B}\mathbf{R}\mathbf{B}') \ , \end{aligned} \tag{13.19}$$

where the operation trace is defined as

$$\operatorname{trace}\mathbf{A} = \sum_i a_{ii}$$

for a square matrix $\mathbf{A}$. One of the properties of the trace is that its value remains the same under any cyclic permutation of the matrix factors, so that, for example, $\operatorname{trace}(\mathbf{B}\mathbf{R}\mathbf{B}') = \operatorname{trace}(\mathbf{B}'\mathbf{B}\mathbf{R})$.

13.4.4 Using the Kronecker product to express $\hat{\mathbf{B}}$

We now need to compute the derivative of (13.19) with respect to matrix $\mathbf{B}$ and set the result to zero. Using the fact that the derivative of $\operatorname{trace}(\mathbf{B}'\mathbf{A})$ with respect to matrix $\mathbf{B}$ is $\mathbf{A}$, we find that $\mathbf{B}$ satisfies the matrix system of linear equations

$$(\mathbf{Z}'\mathbf{Z}\mathbf{B}\mathbf{J}_{\theta\theta} + \lambda\mathbf{B}\mathbf{R}) = \mathbf{Z}'\mathbf{C}\mathbf{J}_{\phi\theta} \ . \tag{13.20}$$

The solution $\mathbf{B}$ to this equation can be expressed explicitly in conventional matrix algebra if we use Kronecker products. The Kronecker product $\mathbf{A}\otimes\mathbf{C}$ is the super or composite matrix consisting of sub-matrices $a_{ij}\mathbf{C}$. Its usefulness in this situation derives from the fact that the matrix expression $\mathbf{A}\mathbf{B}\mathbf{C}'$ can be re-expressed as

$$\operatorname{vec}(\mathbf{A}\mathbf{B}\mathbf{C}') = (\mathbf{C}\otimes\mathbf{A})\operatorname{vec}(\mathbf{B}) \ ,$$

where $\operatorname{vec}(\mathbf{B})$ indicates the vector of length qK_θ obtained by writing matrix $\mathbf{B}$ as a vector column-wise. Moreover, the Kronecker product is also *bilinear* in the sense that

$$\operatorname{vec}(\mathbf{A}_1\mathbf{B}\mathbf{C}_1' + \mathbf{A}_2\mathbf{B}\mathbf{C}_2') = (\mathbf{C}_1\otimes\mathbf{A}_1 + \mathbf{C}_2\otimes\mathbf{A}_2)\operatorname{vec}(\mathbf{B}).$$

The Appendix contains a discussion of properties of the Kronecker product that have been used to obtain these expression, and other properties that will be used subsequently.

Consequently, applying these relations to the two terms involving $\mathbf{B}$ on the left side of (13.20), we obtain

$$[\mathbf{J}_{\theta\theta}\otimes(\mathbf{Z}'\mathbf{Z}) + \mathbf{R}\otimes\lambda\mathbf{I}]\operatorname{vec}(\mathbf{B}) = \operatorname{vec}(\mathbf{Z}'\mathbf{C}\mathbf{J}_{\phi\theta}) \ . \tag{13.21}$$

Now we have a system of qK_θ linear equations expressed in the conventional way that must be solved to obtain the elements of $\mathbf{B}$. The solution is

$$\operatorname{vec}(\mathbf{B}) = [\mathbf{J}_{\theta\theta}\otimes(\mathbf{Z}'\mathbf{Z}) + \mathbf{R}\otimes\lambda\mathbf{I}]^{-1}\operatorname{vec}(\mathbf{Z}'\mathbf{C}\mathbf{J}_{\phi\theta}) \ . \tag{13.22}$$

In (13.21) and (13.22) we have assumed a single smoothing parameter λ to impose the same level of smoothness on each component $\beta_j, j = 1, \ldots, q$

of the vector $\boldsymbol{\beta}$, but we may need the additional flexibility of controlling the smoothness of each component independently by using a smoothing parameter λ_j using a separate roughness penalty for each component. This involves the following minor modification of (13.21): Replace $\lambda\mathbf{I}$ in this expression by the diagonal matrix Λ containing the λ_j's in its diagonal.

Constraining $\boldsymbol{\beta}$ to be smooth is not the same thing as constraining the fit $\hat{y}$ to be smooth, and this latter strategy may be more important in some situations. Again a modification of (13.21) will serve: Replace $\mathbf{I}$ by $\mathbf{Z}'\mathbf{Z}$, in which case (13.22) simplifies to

$$\text{vec}\,(\mathbf{B}) = [(\mathbf{J}_{\theta\theta} + \lambda\mathbf{R}) \otimes (\mathbf{Z}'\mathbf{Z})]^{-1}\text{vec}\,(\mathbf{Z}'\mathbf{C}\mathbf{J}_{\phi\theta}) \ . \tag{13.23}$$

As in Chapter 5, smoothing parameters may be chosen by cross-validation, generalized cross-validation and other methods.

13.4.5 Fitting the raw data

We have been assuming that the response variable $\mathbf{y}(t)$ is a result of previously smoothing the discrete data, but in some applications we may prefer to go straight from the raw response data matrix $\mathbf{Y}$ to estimates of the regression coefficient functions. The penalized least squares criterion in this case is

$$\|\mathbf{Y} - \mathbf{Z}\mathbf{B}\boldsymbol{\Theta}'\|^2 + \lambda\|\mathbf{L}\boldsymbol{\beta}(t)\|^2,$$

where $\boldsymbol{\Theta}$ is the N by K_β matrix of values of the basis functions for $\boldsymbol{\beta}$ evaluated at the sampling points for the response functions. The normal equations to be solved are in this case

$$(\mathbf{Z}'\mathbf{Z})\mathbf{B}(\boldsymbol{\Theta}'\boldsymbol{\Theta}) + \lambda\mathbf{B}\mathbf{R} = \mathbf{Z}'\mathbf{Y}\boldsymbol{\Theta} \ , \tag{13.24}$$

or, using Kronecker products,

$$[(\boldsymbol{\Theta}'\boldsymbol{\Theta}) \otimes (\mathbf{Z}'\mathbf{Z}) + \mathbf{R} \otimes \lambda\mathbf{I}]\text{vec}\,(\mathbf{B}) = (\boldsymbol{\Theta}' \otimes \mathbf{Z}')\text{vec}\,(\mathbf{Y}) \ . \tag{13.25}$$

13.5 Confidence intervals for regression functions

13.5.1 How to compute confidence intervals

The technique for computing point-wise confidence limits for regression functions is essentially the same as we used in Section 5.5. Recall that we required there two mappings. The first was the mapping `y2cMap` from the raw data vector y to the coefficient vector c, corresponding to the n by K matrix $\mathbf{S}_{\phi,\lambda}$ in the equation

$$\mathbf{c} = \mathbf{S}_{\phi,\lambda}\mathbf{y}.$$

We still need this mapping here, but we now assume that there are N replications, and consequently that the raw data reside in an N by n

matrix $\mathbf{Y}$ to be mapped to the N by K_y matrix $\mathbf{C}$ of coefficients of the basis function expansions of the functions N functions x_i. Consequently, as in Section 5.5, the `y2cMap` is represented by the matrix equation

$$\mathbf{C} = \mathbf{Y}\mathbf{S}_{\phi,\lambda} \ .$$

This may be re-expressed in Kronecker product notation as

$$\text{vec}\,\mathbf{C} = (\mathbf{S}_{\phi,\lambda} \otimes \mathbf{I})\text{vec}\,\mathbf{Y},$$

and this permits us to express what we can now denote as `Y2CMap` as

$$\texttt{Y2CMap} = \mathbf{S}_{\phi,\lambda} \otimes \mathbf{I}. \tag{13.26}$$

However, we now have a new linear mapping, namely that of the linear model itself, which maps a coefficient matrix $\mathbf{B}$ to an N by n fit to the data $\hat{\mathbf{Y}}$ by

$$\hat{\mathbf{Y}} = \mathbf{Z}\mathbf{B}\mathbf{\Theta}'.$$

We therefore need an expression for the mapping `C2BMap` that maps the coefficient matrix $\mathbf{C}$ for the response functions to the q by K_β coefficient matrix $\mathbf{B}$ for the regression function vector $\boldsymbol{\beta}$.

Finally, we will want to compute confidence intervals for some functional contrast or linear probe $\rho(\boldsymbol{\beta})$, and this will require a mapping that we can indicate by `B2RMap` that maps the coefficient matrix $\mathbf{B}$ to $\rho(\boldsymbol{\beta})$. For example, we may want to estimate the standard error of a regression function at a value t, and this is the value of the evaluation function $\rho_t(\boldsymbol{\beta})$. But we may also be interested in *functional contrasts* that probe for special effects that interest us as well.

The last stage in actually computing confidence limits is the computing of the composite mapping $\texttt{Y2RMap} = \texttt{B2RMap} \circ \texttt{C2BMap} \circ \texttt{Y2CMap}$ and applying it to each side of $\mathbf{\Sigma}_e$ to get an estimate of the sampling variance of the quantity of interest.

Now let us derive each of these matrix mappings, and put them together as required. That is, we compute the matrix mapping the raw data to the coefficients of the basis function expansions for the β_j's, and then we multiply this by the matrix mapping the regression coefficients to whatever quantities or functionals that interest us.

The first step, then, is to compute the matrix mapping $\mathbf{S}_{\phi,\lambda_y}$ from the data to these coefficients. This is, using the results in Section 5.5 for the response functions,

$$\mathbf{S}_{\phi,\lambda_y} = \mathbf{\Phi}(\mathbf{\Phi}'\mathbf{\Phi} + \lambda_y\mathbf{R}_y)^{-1}\mathbf{\Phi}' \ ,$$

where λ_y is the smoothing parameter used to smooth the data and $\mathbf{R}_y$ is the corresponding roughness penalty matrix. Matrix $\mathbf{S}_{\phi,\lambda_y}$ is then substituted in (13.26) to obtain `Y2CMap`.

From (13.25), we have the mapping from the data coefficients to the regression function coefficients expressed as

$$\text{vec}\,(\mathbf{B}) = [\mathbf{J}_{\theta\theta} \otimes \mathbf{Z}'\mathbf{Z} + \mathbf{R}_\beta \otimes \lambda_\beta \mathbf{I}]^{-1}(\mathbf{J}'_{\phi\theta} \otimes \mathbf{Z}')\text{vec}\,(\mathbf{C}'),$$

where λ_β and $\mathbf{R}_\beta$ are the smoothing parameter and roughness matrix associated with the regularization of the functions β_j. The qK_β by NK_y matrix

$$\texttt{C2BMap} = \mathbf{S}_\beta = [\mathbf{J}_{\theta\theta} \otimes \mathbf{Z}'\mathbf{Z} + \mathbf{R}_\beta \otimes \lambda_\beta \mathbf{I}]^{-1}(\mathbf{J}'_{\phi\theta} \otimes \mathbf{Z}') \tag{13.27}$$

is the matrix mapping that we need.

These last two expressions can then be combined into

$$\texttt{Y2BMap} = \text{vec}\,(\mathbf{B}) = \mathbf{S}_\beta(\mathbf{S}_{\phi,\lambda_y} \otimes \mathbf{I})\text{vec}\,(\mathbf{Y}') \ . \tag{13.28}$$

The variance of the raw data arranged as a vector is given by

$$\texttt{Var}[\text{vec}\,(\mathbf{Y}')] = \mathbf{\Sigma}_e \otimes \mathbf{I} \ ,$$

where $\mathbf{\Sigma}_e$ is the variance-covariance matrix of the residual vectors e_i and $\mathbf{I}$ is of order N. Note that these residuals are for the linear model (13.13) and not the residuals involved in smoothing the raw data for the response variable.

We can now put this all together to get what we need in terms of the coefficients of the expansions of the β_j;

$$\texttt{Var}[\text{vec}\,(\mathbf{B})] = \mathbf{S}_\beta(\mathbf{S}_{\phi,\lambda_y} \otimes \mathbf{I})(\mathbf{\Sigma}_e \otimes \mathbf{I})(\mathbf{S}_{\phi,\lambda_y} \otimes \mathbf{I})\mathbf{S}'_\beta \ . \tag{13.29}$$

If our objective is an estimate of $\texttt{Var}[\text{vec}\,(\hat{\boldsymbol{\beta}})]$, then this is

$$\texttt{Var}[\text{vec}\,(\hat{\boldsymbol{\beta}})] = (\mathbf{\Theta}\otimes\mathbf{I})\mathbf{S}_\beta(\mathbf{S}_{\phi,\lambda_y}\otimes\mathbf{I})(\mathbf{\Sigma}_e\otimes\mathbf{I})(\mathbf{S}_{\phi,\lambda_y}\otimes\mathbf{I})\mathbf{S}'_\beta(\mathbf{\Theta}\otimes\mathbf{I})' \ . \tag{13.30}$$

If both $\mathbf{J}_{\theta\theta}$ and $\mathbf{Z}'\mathbf{Z}$ are invertible, then this expression can be simplified to

$$\texttt{Var}[\text{vec}\,(\hat{\boldsymbol{\beta}})] = [\mathbf{J}^{-1}_{\theta\theta}\mathbf{J}'_{\phi\theta}\mathbf{S}_{\phi,\lambda_y}\mathbf{\Sigma}_e\mathbf{S}_{\phi,\lambda_y}\mathbf{J}_{\phi\theta}\mathbf{J}^{-1}_{\theta\theta}] \otimes (\mathbf{Z}'\mathbf{Z})^{-1} \ . \tag{13.31}$$

If the raw data are fit directly, the corresponding expression is

$$\texttt{Var}[\text{vec}\,(\hat{\boldsymbol{\beta}})] = [\mathbf{\Theta}(\mathbf{\Theta}'\mathbf{\Theta})^{-1}\mathbf{\Theta}'\mathbf{\Sigma}_e\mathbf{\Theta}(\mathbf{\Theta}'\mathbf{\Theta})^{-1}\mathbf{\Theta}'] \otimes (\mathbf{Z}'\mathbf{Z})^{-1} \ . \tag{13.32}$$

13.5.2 Confidence intervals for climate zone effects

We now illustrate this method for computing confidence intervals by estimating climate zone effects for the daily mean temperature data for 35 weather stations. We smoothed these data with a Fourier series basis with 65 basis functions without regularization in order to economize on computer time and work with temperature profiles that were somewhat smoother than those that we obtained in Chapter 5. The figures in Section 13.2 were obtained using these functional responses.

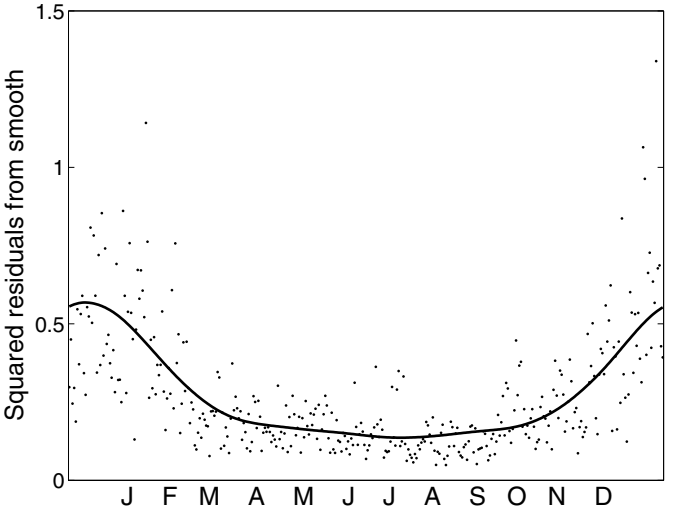

Figure 13.9. The points indicate daily estimates of the standard error of measurement for the mean temperature data computed across 35 weather stations, and the solid line is a positive smooth of these values.

Figure 13.9 shows the raw daily standard error estimates taken across the 35 stations as well as a positive smooth of these estimates of the kind that we discussed in Chapter 6. These vary between 0.4 degrees Celsius in the summer to 0.7 in the winter. This is a fair amount of variation, and so we put the reciprocal of the smoothed variances of measurement in the diagonal of weight matrix $\mathbf{W}$ in (5.3), and then re-smoothed the data.

In the functional analysis of variance step, we defined $\mathbf{Z}$ to be the 35 by 5 matrix containing the value 1 in column 1, and coding membership in the Atlantic, Pacific, Continental and Arctic regions in columns two to five as above. We used 21 Fourier basis functions to represent the 5 regression functions β_j.

The order 365 variance-covariance matrix $\mathbf{\Sigma}_e$ for the residuals from the linear model was estimated in the same way that we described in Chapter 2.

Figure 13.1 displays the 95% point-wise confidence limits on the estimated curve climate zone effect function. We see, for example, that it is only in the winter months that the temperature in Pacific zones can be considered as significantly warmer than those represented by the intercept function. For the record, a direct approximation of the raw average temperatures produced function and confidence limit estimates that were effectively indistinguishable from these results.

13.5.3 Some cautions on interpreting confidence intervals

Many things can go wrong in interpreting confidence interval estimates such as those in Figure 13.1, and it is important here to stress their limitations so as to avoid misleading ourselves and others in applications leading to serious decisions.

First of all, point-wise limits are not the same thing as confidence regions for the entire estimated curve. As mentioned in Section 5.5, although can envisage of pairs of curves between which there is a specified probability that the entire true curve is to be found, this requires the use of computationally intensive resampling methods. Point-wise regions are useful, of course, but they are based on the assumption that conclusions are only to be drawn *at that point.* So when we indicated above that we could conclude that summer temperatures in the Pacific zone are not much different from the national average, this was, strictly speaking, an abuse of the concept of a point-wise region.

Secondly, the confidence region estimates that we have developed are based on strong assumptions that may not be true. We have, as is usual in the analysis of variance, implicitly assumed that the distribution of the residuals is the same within each group. In fact, it is hard to imagine that, given enough weather stations, we would not also see systematic differences in covariances and other aspects of dispersion from one climate zone to another. In any case, the very idea of basing a confidence region on an estimated covariance involves the strong assumption that the joint distribution of two residuals is well summarized by a covariance, as would be the case if they were normally distributed. In fact, if the actual residual distribution is strongly skewed, if it has long tails, if it is multi-modal, or any one of many other violations of what normality implies is operative, these regions will not work as advertised. That is, in this case, we cannot claim them to contain the true curve with the specified probability.

Thirdly, we are estimating something whose potential dimension can outstrip any quantity of data that we can afford to collect. A curve can be made arbitrarily complex given enough information. It seems likely, surely, that someone working with ten times the number of weather stations, fifty years worth of data, and taking spatial dependencies into account will discover features of temperature curves that we could not capture with the number of basis functions that we used. Consequently, any claims that we may make about the precision with which we have estimated these curves must be understood to be conditional on the amount and quality of information that we have at our disposal. There is no asymptotic sample size when it comes to estimating a curve. Period. Although the results we have in this section are what some texts would call "small sample" results, in fact, we had to do here what is almost always done in practice, that is, substitute a sample estimate for a population quantity, namely $\mathbf{\Sigma}_e$. We really don't know what the full implication of this might be.

We can put the problem this way. In classical statistics, we usually work with a fixed number of parameters. But in nonparametric curve estimation, the number of basis functions K has the characteristics of a random variable. Two investigators working with different objectives, different number of sampling points, different residual variance levels, different ranges, and so forth are quite likely to work with different values of K. We really need, therefore, to take into account variation in K in our interval estimates, something that this chapter has not done. The interval estimation problem is, as far as K is concerned, more suitable for a Bayesian approach than for the classical methodology used here.

Elsewhere, moreover, we will have to substitute approximations for so-called "exact" results. Because, for example, the monotone smoother is not a linear function of the data, it was necessary to replace an exact calculation of the mapping `y2cMap` by a first-order approximation. This is inevitably a crude approximation in many situations, and always has the potential to be misleading.

So what to do? In the end, there is probably no safe substitute for computationally intensive methods such as simulation, bootstrapping of various kinds, and cross-validation methods. If these methods give results in essential agreement with these cheaper exact or asymptotically correct estimates, perhaps we can breathe a sigh of relief and carry on. But we should always assume that our decisions will only be reasonable until better data become available.

13.6 Further reading and notes

Brumback and Rice (1998) reported a functional analysis of variance involving daily progesterone metabolite concentrations over the menstrual cycles of 91 women enrolled in an artificial insemination clinic. The main experimental factor was whether conception occurred (21) or not (70). Within and between woman variation was also assessed. This work proceeded independently of Ramsay and Silverman (1997), but ended up using rather similar methods, and identified some serious computational difficulties involved with working with random functional factors. The discussion that followed the paper highlighted a number of issues. We strongly recommend reading this paper as a supplement to this chapter.

Faraway (1997) used functional ANOVA to study three–dimensional movement trajectories in a complex industrial design setting. Muñoz Maldonado, Staniswalis, Irwin and Byers (2002) suggest three ways of testing the equality of curves collected from samples of young and old rats. Another application of functional ANOVA can be found in Ramsay, Munhall, Gracco and Ostry (1996), where variation in lip movement during the production of four syllables is analyzed at the level of both position and acceleration.

Spitzner, Marron and Essick (2003) combine functional ANOVA with a mixed-model approach to study human tactile perception. Fan and Lin (1998) proposed a method for testing for significance when the response variable is functional. Yu and Lambert (1999) fit tree models to functional responses.

Li, Aragon, Shedden and Thomas Agnan (2003) offer an approach that combines elements of the concurrent functional linear model discussed in Chapter 14, principal components analysis and the varying-coefficient model. This paper applies the *sliced inverse regression* or *SIR* method developed by Li (1991) to a functional response predicted by one or scalar independent variables.

Chiou, Müller and Wang (2003) describe an interesting variant of the functional linear model. They propose that principal components scores $f_{im}, m = 1, \ldots, M$, associated with the response functions x_i are related to the covariate values z_{ij} through

$$f_{im} = \alpha_m(\sum_j \beta_{mj} z_{ij}) + \epsilon_{im}. \tag{13.33}$$

This model combines a linear model for the arguments of the regression coefficient functions α_m with a principal components model for the response. Models in which argument values are themselves linear combinations of covariates are often referred to as *single index models*. The med-fly life history data to which the authors apply the model have been analyzed in many fascinating and original ways, and a collection of the papers on these data makes fascinating reading for anyone wishing to see functional data analysis in action.