THE UNIVERSITY OF CALGARY

Functional Data Analysis of Foot Orthotics

by

Carina Wang

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF MATHEMATICS AND STATISTICS

CALGARY, ALBERTA

January, 2004

THE UNIVERSITY OF CALGARY

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled "Functional Data Analysis of Foot Orthotics" submitted by Carina Wang in partial fulfillment of the requirements for the degree of MASTER OF SCIENCE.

Supervisor, Dr. P.F. Ehlers,
Department of Mathematics and Statistics

Dr. E. Enns,
Department of Mathematics and Statistics

Dr. W.W. Zwirner,
Department of Applied Psychology

Jan 29, 2004

Date

Abstract

It is the aim of this paper to investigate the suitability of applying functional data analysis (FDA) methods to the study of foot orthotics data, provided by an experiment designed to assess the effect of shoe inserts on some relevant kinematic and kinetic variables.

After a brief introduction, the following chapters provide data analysis that consists of several procedures commonly adapted to FDA. First, we prepare the data for functional representation and proper data display, involving smoothing and registration techniques. Second, this data preparation is followed by some conventional statistical methods. These are basic t-tests, variance analysis and correlation coefficient analysis. Third, again, a functional counterpart of a method in multivariate data analysis – principal components analysis – is employed to explore data variation.

With sufficiency of the data set, we are able to draw some conclusive results using the proposed FDA methods above; the definite outcomes in turn assure the suitability of FDA application on such data.

Acknowledgements

I would like to thank my supervisor Dr. Peter F. Ehlers for providing a good environment, StatCar, for me to work on this project, and for his vast enthusiasm matched only by his patience for my endless questions. I also wish to acknowledge Dr. Ernest G. Enns and Dr. Walter W. Zwirner for their comments and suggestions.

After all is said and done, this one is for Jackie. I just can't imagine having done this without his confidence in me and endless support. I can't possibly thank you enough Jackie.

to Jackie

Table of Contents

List of Figures

Chapter 1

Introduction

Functional data analysis (FDA) is a recently active area. There are growing numbers of scientific and research fields in which data are collected through a process naturally described as functional; for example, each observation y_i, $i = 1, ..., N$, is a function. Historically, FDA goes back at least to the attempts of Gauss and Legendre to estimate comet trajectories [RS2]. Today, with sophisticated monitoring and imaging equipment used in medicine, seismology, meteorology, and so on, massive data sets can easily be scanned in the form of curves, forming sets of functional data. More precisely, the observations of individuals at different time points are recorded. For example, a weather station generates monthly data (temperature, precipitation) that are real functions of one variable – time. In another example, the temperature profile across the whole country at a particular time of the day is a real function of two variables – longitude and latitude. The defining quality of functional data is that they consist of functions. Functions are usually estimated from discrete data by smoothing techniques.

What is the difference between longitudinal data analysis and FDA? The main difference between these is in the dimensionality of the data vector, though such a distinction is not always clear. The dimensionality in functional data analysis is usually much higher, and hence smoothing techniques are needed. In addition, FDA tends to give a critical role to one or more derivatives.

FDA has the same fundamental aims as those of more conventional statistics:

to address problems that are responsive to statistical consideration and analysis; to work out ways of displaying data that highlight interesting and significant features; to explore variability and average characteristics, and so on. What sort of data may be considered as functional data? In some cases, functional data can be obtained by interpolating their original observations from longitudinal data, quantities observed as they evolve through time. Non-stationary time series data are appropriate for FDA most commonly [RD]. However, functional data come in many other forms. Here is the beauty of FDA: each new functional dataset offers new challenges and opens a door to explore new ideas and techniques.

The field of functional data analysis is still in its early development, and the boundaries between functional data analysis and other aspects of statistics are vague. A fundamental assumption in FDA is that observed data functions are single entities. In practice, however, very often – by the nature of digital technology – functional data are recorded in a discrete manner.

The data set, foot orthotics data, in this thesis is provided by the Human Performance Laboratory at University of Calgary. It came from the study of the effect of several types of shoe inserts on relevant biomechanical variables during the human gait cycle. Even if the raw data exhibit a certain amount of roughness, it seems reasonable to assume that the human gait is a continuous process. Orthotics data are functional in the sense of being representable by smooth functions. Functional data analysis is a set of methods for the analysis of samples of curves, and our data are natural candidates for such analysis. The main aim of this thesis is to investigate the suitability of functional data analysis approaches to analyzing the data set.

The first task of FDA is to transform the raw discrete values into true functional

form since the foot orthotics data were obtained discretely. Chapter 2 reviews some techniques for smoothing, including basis function expansion and roughness penalty methods. Due to both timing and magnitude differences seen in the sampled functions, transforming functions by transforming their arguments – which is called registration – becomes the other subject of this chapter. After preliminary smoothing and registration steps, we perform data analysis in Chapter 3, which applies classical summary statistics used in functional form in FDA, such as pointwise t-tests and correlation functions. Chapter 4 introduces the principal components analysis of functional data to continue the data analysis. This is a key exploratory method, which gives us the tools to observe the main features characterizing functions. It was the first method considered by early literature on FDA.

The data set used in this dissertation is adequate, and even with subjective variability, some outcomes are still conclusive. Hence, it shows that the methods proposed here have considerable promise for data of this type.

Chapter 2

First Steps in Functional Data Analysis

In this chapter, we first discuss the ideas of data smoothing which are the techniques for converting raw functional data into true functional form. This is followed by an essential preliminary to functional data analysis, the registration or alignment of salient curve features. Finally, we take a close look at the data with which this dissertation is concerned and process the data to prepare them for further analysis.

2.1 Representing functional data as smooth functions

Smoothing methods can aid in data analysis in two important ways. First, they are able to highlight the underlying structure in the data. Second, due to being free from rigid parametric distribution assumptions, they can provide both flexible and robust analysis [Si2]. Moreover, in any exploratory analysis, we should not underemphasize the importance of looking at the data. Smoothing methods provide a way of meeting this end desirably – often even the simplest graphical smoothing methods will highlight important structure clearly.

2.1.1 Usage of smoothing methods

Functional data analysis concerns an observed data function as a whole as being a datum instead of a sequence of single observations. In practice, functional data are usually observed and recorded discretely. Assuming that a set of discretely

observed values, $y_{j1}, ..., y_{jn}$, is a functional datum for replication j, we should first convert the y values to a real function, $x_j(t)$, where t is an argument ranging over an arbitrary interval τ. The process may involve interpolation or smoothing, depending on whether the discrete values are observed errorless.

Smoothness implies a certain degree of derivatives, and smoothing converts raw data into true smooth and continuous functional form. The discrete raw data are supposed to have functional form, but they are usually imposed with observational error or noise. We write

$$y_j = x(t_j) + \epsilon_j, \tag{2.1}$$

where ϵ_j denotes the error term and results as a roughness to the raw data. Therefore, as an efficient approximation technique, smoothing is required to filter out the noise of raw data in order to obtain a better functional representation.

Now we turn to a discussion of various smoothing methods designed for direct observational error. One particular class of methods, based on roughness penalties, plays a particular role in our development of functional data analysis methods. Before introducing roughness penalties in Section 2.1.3, we briefly discuss linear regression first.

Linear regression is one of the oldest and most widely used statistical techniques [GS]. The natural way to view linear regression is as a method fitting a model of the form

$$y = a + bt + error \tag{2.2}$$

to the observed data. It is natural to draw a straight line to emphasize linear trend when there exists the linear relationship between the design variable t and the re-

sponse variable y. However, there are many data sets where it is clearly inappropriate to fit a straight line to the model of the form (2.2) and where a model of the form

$$y = g(t) + error \tag{2.3}$$

is called for, where g is a curve of some sort. If g is a polynomial, this approach is called polynomial regression. The disadvantages for this approach are that individual observations can exert an influence on remote parts of the curve, and that the model elaboration implicit in increasing the polynomial degree happens in discrete steps and cannot be controlled continuously.

Due to the drawbacks of the two previous approaches, a more flexible approach is required.

2.1.2 Basis expansion approach

The basis expansion method is one of the most familiar smoothing procedures, and its basic philosophy is to represent a function by a linear combination of K known basis functions ϕ_k,

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t), \tag{2.4}$$

where c_k are the coefficients of the expansion. The number K of basis functions controls the degree to which the data y_j are smoothed. When $K = n$, the coefficients c_k are chosen to yield $x(t_j) = y_j$ for each j, and thus an exact interpolation is generated.

The linear smoother, $x(t)$, can be obtained once we determine c_k by minimizing

the least squares criterion

$$S(y|c) = \sum_{j=1}^{n} \left\{ y_j - \sum_{k=1}^{K} c_k \phi_k (t_j) \right\}^2 . \tag{2.5}$$

The choice of basis functions is important, and there is no good universal basis. The type of basis functions should be chosen so that their characteristics match those of the functions being estimated. Therefore, we can choose the basis by observing the raw data. One of the basis choices, Fourier series, is useful for periodic or extremely stable functions. For example, for data such as the Canadian annual temperature data in the book by Ramsay and Silverman [RS2], Fourier expansion is an appropriate choice. The other types of basis are such as polynomial bases, wavelet bases, and the most common one, B-splines.

B-splines

The inability for Fourier and polynomial bases to accommodate local features led to the development of polynomial splines. B-splines is one of the most popular kind. A B-spline basis is defined by a set of knots, and our strategy is to place a knot at each time point corresponding to an observation. B-splines are constructed from polynomial pieces, joined at certain values of x, the knots [EM]. Once the knots are given, it is easy to compute the B-splines recursively for any desired degree of the polynomial.

In practice, B-splines are a set of special spline functions that can be used to construct piece-wise polynomials by computing the appropriate linear combination. Computational convenience is derived from the fact that any B-spline basis function is nonzero over at most m (order of B-spline) adjacent intervals. Consider as an illustration the very common case where the order m is 4 for all polynomials, so that

the degree of each polynomial is $m - 1 = 3$. That is, the polynomials, and therefore the B-splines are cubic.

It is recommended to choose a B-spline basis for any function that is not periodic and that has no other restriction on its shape [RS3].

2.1.3 Roughness penalties

The task of smoothing is not merely fitting the data, but also capturing a slowly changing trend hidden by the local variation in the curve. The roughness penalty or regularization is a method to quantify the rapid local variation of x and to make an optimal trade-off between regularity and goodness of fit of the curve. In its simplest form, the roughness penalty approach relaxes the model assumptions in classical linear regression along lines a little different from polynomial regression.

Given a curve x defined on an interval $[a, b]$, there are many different ways of measuring how "rough" or "wiggly" the curve x is. An intuitively appealing way of measuring the roughness of a twice-differentiable curve x is to calculate its integrated squared second derivative. This estimates the total curvature in x. Consequently, high values of

$$\int_a^b \{x''(s)\}^2 \, ds \tag{2.6}$$

can be expected to result from highly variable functions due to their large second derivatives. Then, we use

$$S(x, \lambda) = \sum_{j=1}^{n} \{y_j - x(t_j)\}^2 + \lambda \int_a^b \{x''(s)\}^2 \, ds \tag{2.7}$$

to define the *penalized* residual sum of squares. Our estimate of the function x is obtained by minimizing $S(x, \lambda)$ over the space of functions x.

As a smoothing parameter, λ measures the rate of exchange between fit to the data and and variability of the function x, as quantified by the first term of Equation (2.7) and by Expression (2.6), respectively. The penalty, measured by the second term of Equation (2.7), controls the size of the second derivative of $x(t)$, i.e. the curvature of the derivative $x'(s)$. That is, as $\lambda \to \infty$, the fitted curve x approaches the standard linear regression to the observed data. On the other hand, as $\lambda \to 0$, the penalty matters less and less, and the smoothing function will pass as closely as any curve x can to the actual data. That is, the curve x approaches an interpolant to the data. Ramsay and Li [RL] recommend that one should penalize with a derivative two orders higher than the highest derivative required in the model.

2.1.4 Choosing the key parameters

Once the type of basis function is chosen, the involved parameter values are yet left to be decided. These include the number of knots and their positions, number and order of the basis functions, and value of the smoothing parameter λ. Each of these parameters plays a role in governing the degree of smoothness and fitness of the estimated curve. The choice of knots has been a subject of much research: too many knots lead to over-fitting of the data; too few knots lead to under-fitting. Friedman and Silverman [FS] suggested to begin with a dense set of knots, and then eliminate unneeded knots by an algorithmic procedure similar to variable selection techniques used in multiple regression. Where do we position the knots? Knots could be either equally or unequally spaced. When we need to account for the varying amount of curvature or local roughness, the latter is used. For the foot orthotics data set, knot positions are evenly spaced, approximately between every five time points. We use

25 order 6 B-splines. We choose 25 because this is judged sufficient to capture the complexity of the function curves, and we choose order 6 because it allows modeling of higher derivatives of the data.

A major problem of any smoothing technique is the choice of the optimal amount of smoothing; therefore, the choice of the smoothness parameter λ is a delicate matter. The many methods discussed in the nonparametric regression literature tend to be based on the assumption that it is only the characteristics of $x(t)$ that matter, but it is known that stable estimation of derivatives requires more smoothing. Moreover, most of these methods assume that observational error is independently and identically distributed, but the indications of substantial serial correlation have been noted in many studies of orthotics data. Smoothing parameter λ can either be chosen by the inspection of smoothness, or by an automatic procedure such as generalized cross-validation (GCV). Unfortunately, according to Ramsay and Bock [RB], methods like cross-validation are known to be highly sensitive to correlational structure in these errors. They also experimented with calibrating the smoothing process on simulated data, using some parametric models and a realistic observational error structure. However, these models are probably too smooth, and this calibration process tends to over-smooth the data. Finally, we tend to agree with Chaudhury and Marron [CM] that the best strategy is to view the data over a range of λ to see what features emerge at various smoothness levels. By varying the smoothing parameter, features of the data that arise on different "scales" can be obtained by a subjective choice. Moreover, in exploratory analysis designed to hint at what might be seen in future data sets with more information, it may be better to be less conservative than otherwise. In the present context, we have found it satisfactory to choose the

smaller values of λ that still provide a smooth and interpretable estimate of $x(t)$.

2.1.5 Examples

Figure 2.1 plots the knee external rotation moment (Variable 6) for one subject with medial posting (Condition 2). The raw data points are shown as the circles. The solid line is constructed by connecting all the observed values, the circles, and thus is an interpolation function with zero residual sum of squares. The dashed line represents a smoothed curve with nonzero residual sum of squares, but its smoothness shows clearer data pattern and less local variation.

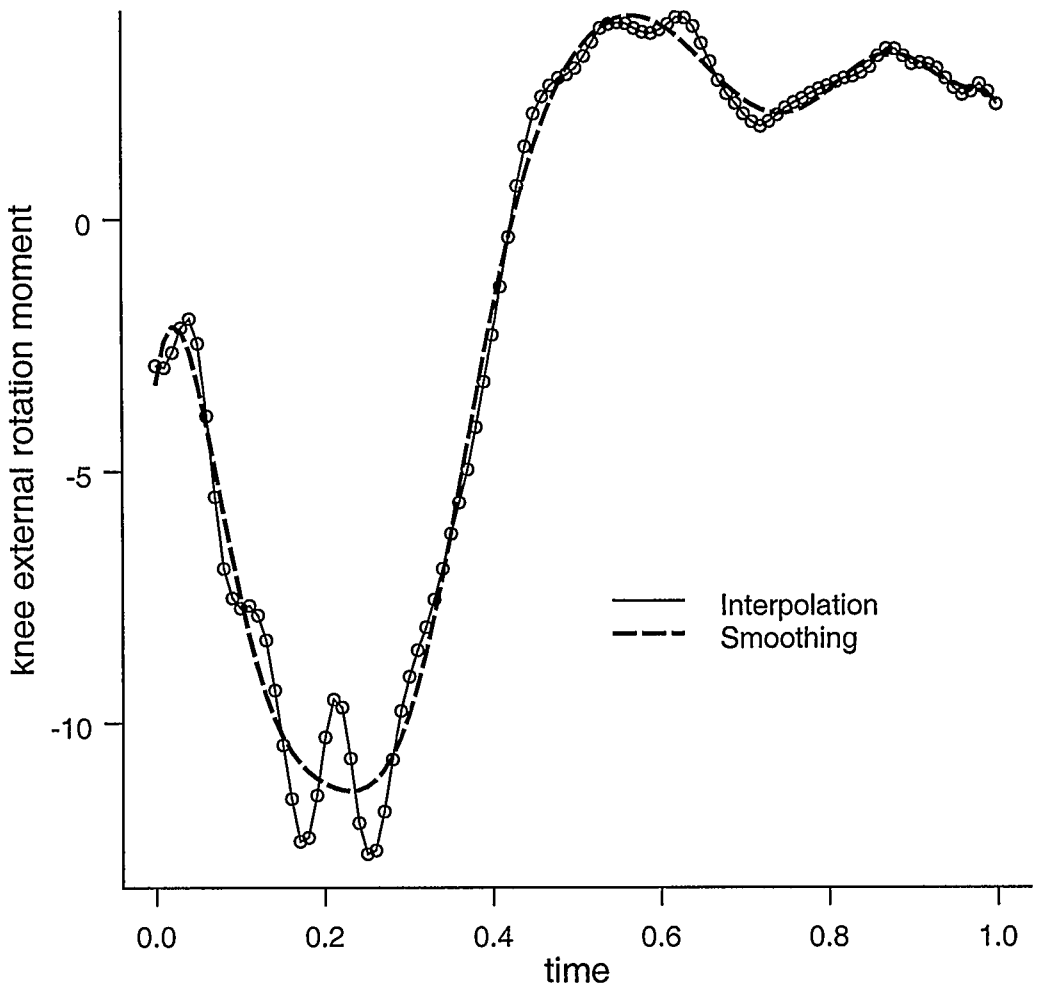

Figure 2.1: Interpolated and smoothed curves for the raw data.

The effect of choosing different values for the smoothing parameter λ is seen in Figure 2.2. The solid curve, with λ as zero, shows the largest amount of goodness-of-fit. This amount decreases with increasing value of λ, resulting in more and more smoothness. For smoothing our data, λ is set to be in the range of $(10^{-8}, 10^{-6})$ for the rough curve, depending on different variables.

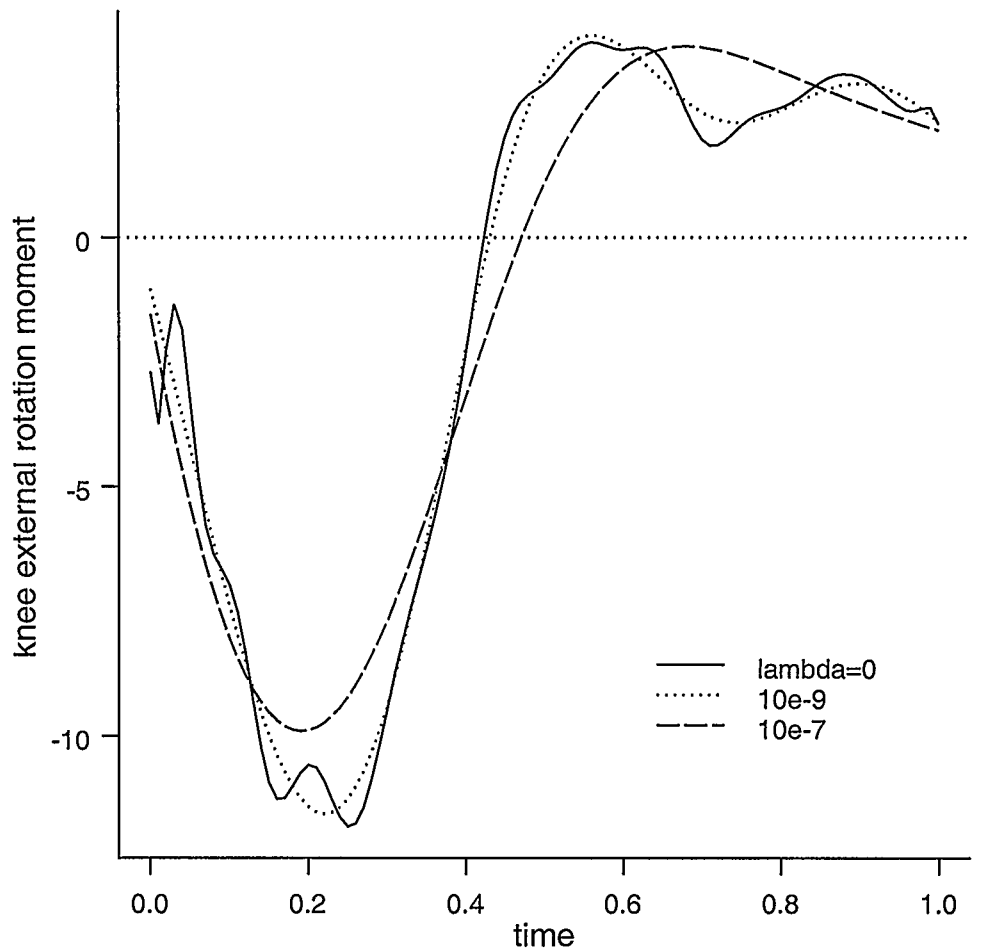

Figure 2.2: Effect of different values of the smoothing parameter.

The effect of different roughness penalty integrands at fixed λ is shown in Figure 2.3. Roughness of the curve in the top panel is measured by using the integral of the square of its second derivative. The curves in the middle and bottom panels result from penalizing the third and fifth derivative function, respectively. It is obvious that the latter two show more and more extensive smoothness.

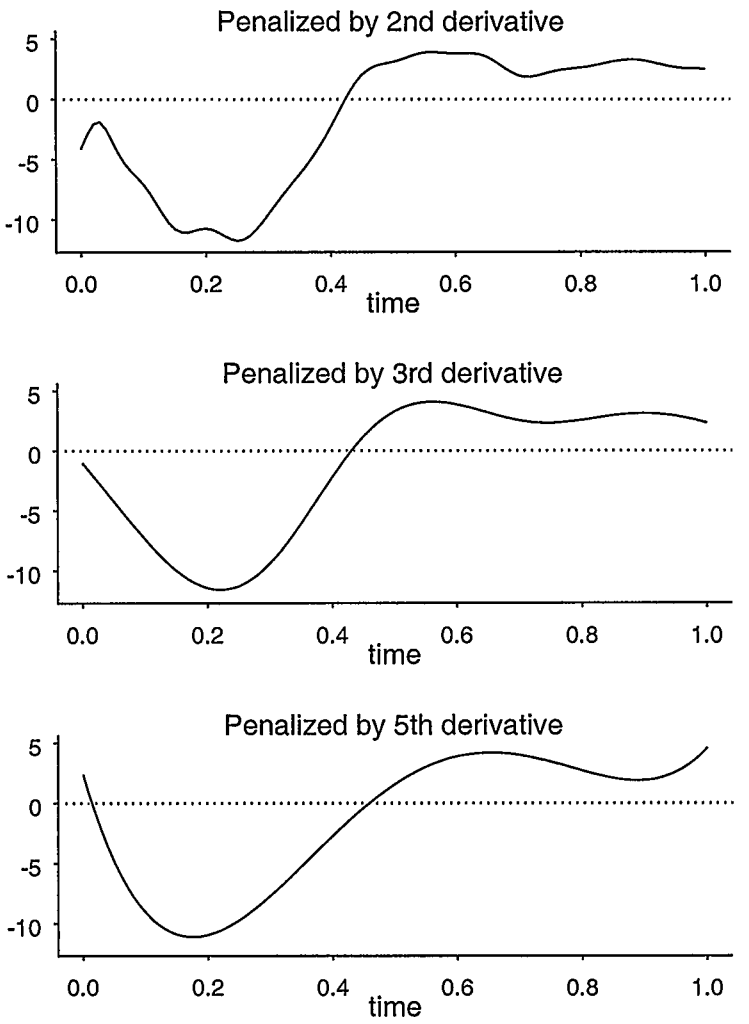

Figure 2.3: Roughness penalty: effect of three levels of order of derivative.

2.2 Foot orthotics data

The shoe orthotics data were provided by A. Muendermann of the Human Performance Laboratory at the University of Calgary. Several kinematic and kinetic factors have been suggested to increase a runner's risk for injuries. It has been speculated that foot orthotics can be used to reduce injury-related complaints or even to prevent running injuries by affecting these factors. Therefore, the objective of the experiment is to quantify the effects of posting and custom-molding of foot orthotics on lower extremity kinematics and kinetics during running.

Twenty-one volunteers participated in this study. Data for Subject 1 were excluded due to technical errors in the analog data. All subjects had neither history of lower extremity injuries nor had previously worn foot orthotics. Kinematic and kinetic data were accessed during nine sessions over a three-week period. In each of the nine sessions, each subject performed twelve running trials (12 trials was designed for the experiment; however, some sessions have fewer trials due to unavailability, and several sessions have 13 trials.) for each of the four insert conditions as follows:

Condition 1: control

Condition 2: medial posting

Condition 3: custom-molding

Condition 4: combination of medial posting and custom-molding

The four insert conditions were tested in randomized order. All subjects ran the same distance for each condition, and thus uncontrolled effects of mileage in foot orthotics on the outcome of this investigation were eliminated. The angle, force

and moment curves of the three-dimensional lower extremity kinematics and kinetics were collected for a single step normalized to touch-down and toe-off resulting in 101 data points per curve per trial. The variables of interest are:

Variable 1: foot inversion

Variable 2: foot inversion velocity

Variable 3: tibia rotation (ankle)

Variable 4: ankle inversion moment

Variable 5: knee adduction moment

Variable 6: knee external rotation moment

Variable 7: knee adduction angle

Variable 8: tibia rotation (knee)

Variable 9: vertical loading rate

Variable 10: vertical ground reaction force

Variable 11: flexion moment ankle

Variable 12: flexion moment knee

Variable 13: ankle flexion angle

Variable 14: knee flexion angle

Variable 15: tibia rotation velocity (ankle)

Suggested by Muendermann [Mu], the variables in bold are studied more profoundly.

Now, before processing our data, we should look at the raw data to get an initial impression. Figure 2.4 plots the raw data for foot inversion for two subjects

(Subjects 6 and 12) in the same session with all four conditions. All the treatment conditions differ from the control condition in some ways. Most obviously, medial posting (Condition 2) and custom-molding (Condition 3) vary considerably, but custom molding (Condition 3) and the combination of medial posting and custom-molding (Condition 4) are similar. Figure 2.5 shows foot inversion with posting condition in all nine sessions for one subject. Apparently, triàls within sessions look more alike than trials between sessions. These impressions are consistent with what Muendermann [Mu] stated in her thesis.

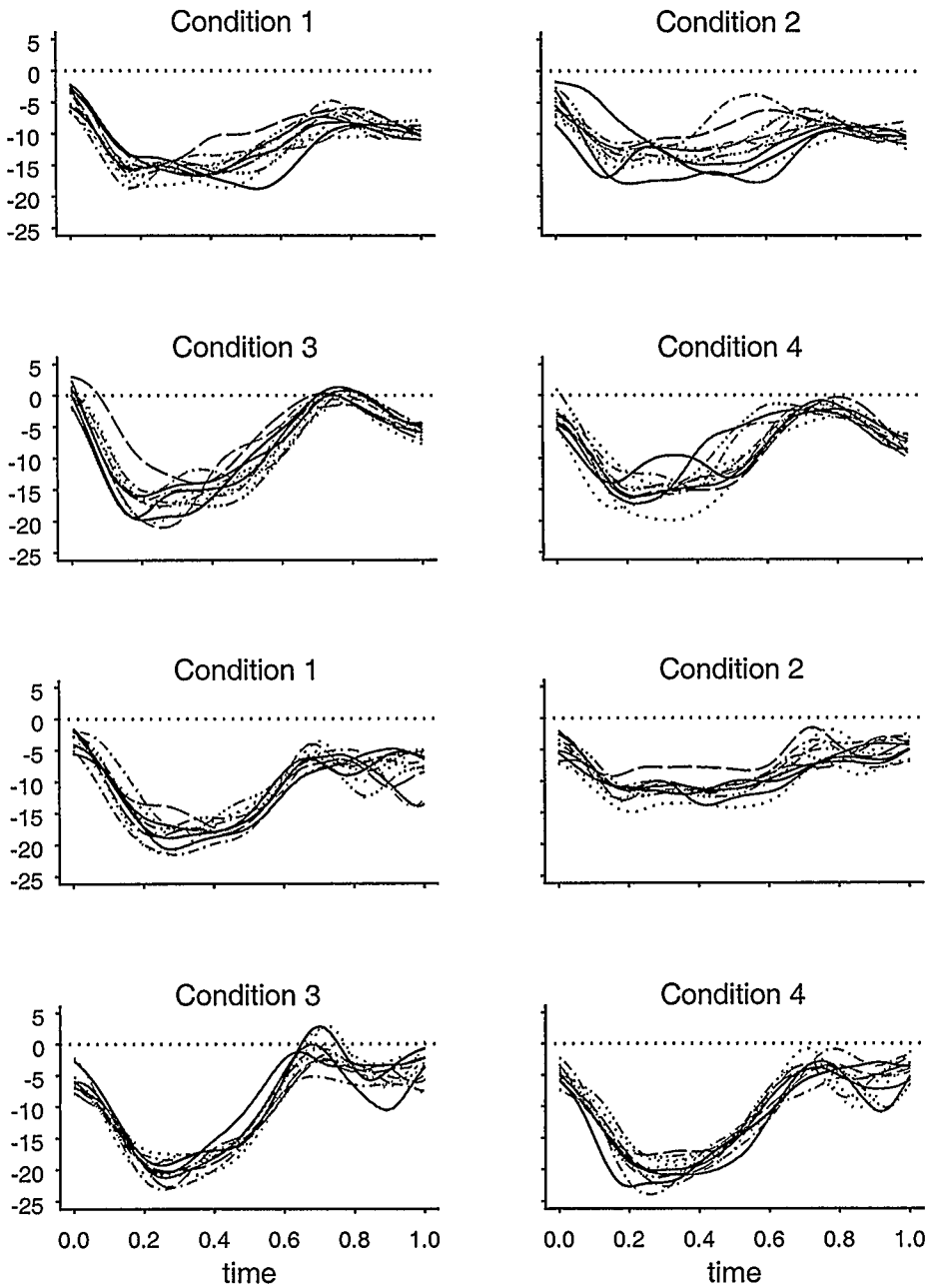

Figure 2.4: Raw data for foot inversion (Variable 1) with all conditions in one session for Subject 6 (top 4) and Subject 12 (bottom 4).

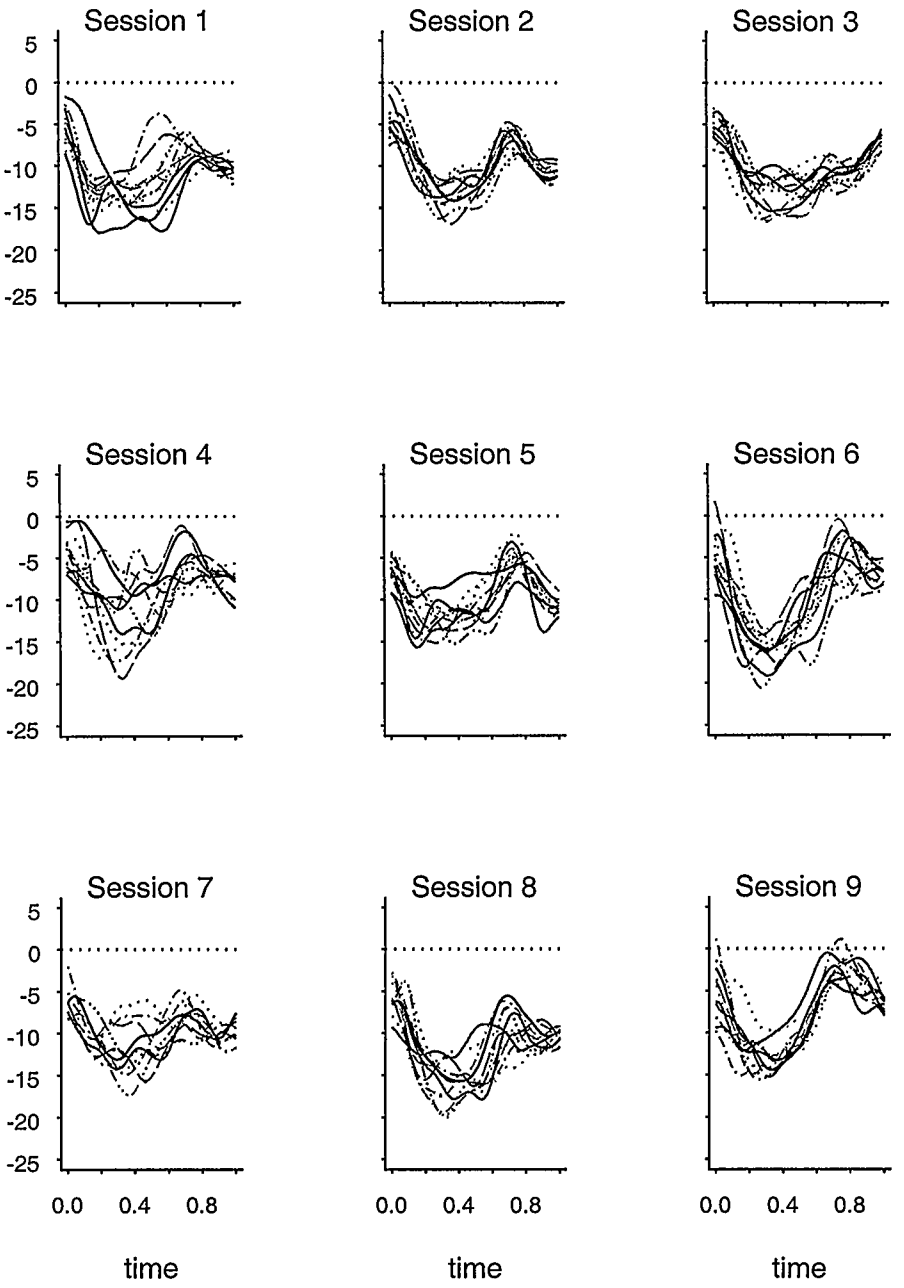

Figure 2.5: Raw data for foot inversion (Variable 1) with posting condition for Subject 6.

2.3 Data processing

Before analyzing our data, we need to clean up the data in the sense of transforming the discretely obtained data to functional form, performing some smoothing, and removing the phase variation in order to align the amplitude variation. To prepare the data for analysis, we follow the preparation steps introduced by Araki [Ar]:

Step 1: Transform the raw data to functional form

Step 2: Register the functional data among trials

Step 3: Reduce the dimension of functional data

Step 4: Register the summary functions

Transform the raw data

According to Muendermann [Mu], within-day repeatability is greater than between-day repeatability for kinematic and kinetic data. Figure 2.5 supports this statement. To assess this statement, one can compare within-session and between-session variability. Figure 2.6 illustrates such a comparison for one subject with one condition in Variable 1. Nine plain solid lines are cross-sectional standard deviation curves of each session. The bold solid line presents the cross-sectional average of those nine plain solid lines, while the dashed line presents the cross-sectional standard deviation curve over all trials from nine sessions. The dashed line's being higher than the bold solid line exhibits that between-session variability is greater than within-session variability. This is also the case for the other subjects, conditions and variables. Due to this fact, one should analyze the data within session in order to retain as

much power as possible, but data analysis becomes a more complicated and time consuming matter. For such a trade-off situation, we might effect a compromise, a suboptimal choice of combining trials of all sessions for further analysis, ignoring less repeatability between sessions. Before making such a choice, it is still necessary to detect whether there is session effect or not.

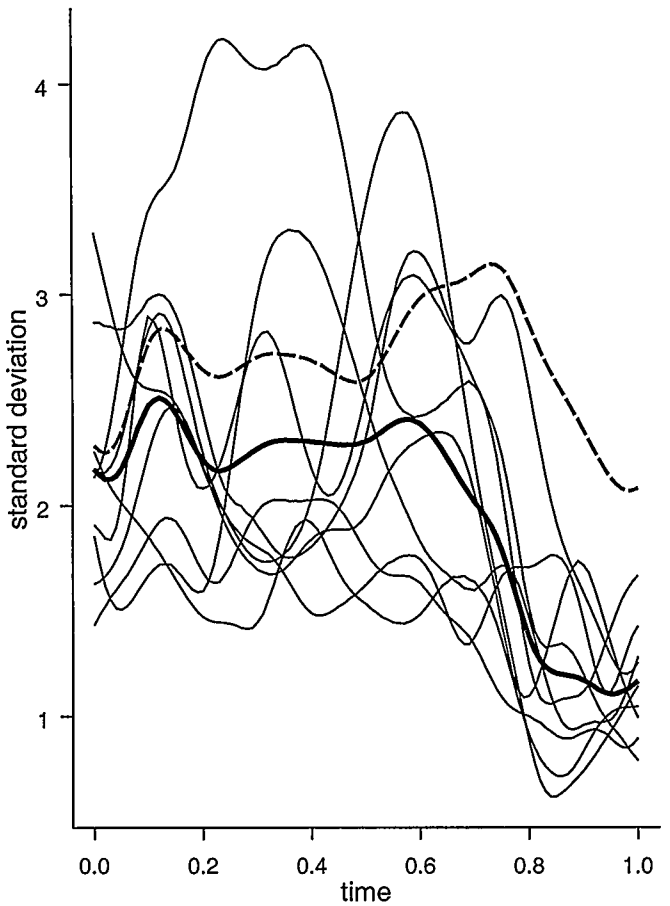

Figure 2.6: Comparison of variability between sessions and within sessions for Subject 6 with Condition 2 in foot inversion (Variable 1).

Figure 2.7 plots cross-sectional mean curves of each of nine sessions to show the ranks of them for one subject with Condition 2 in Variable 1. It does not seem that there is obvious session trend within these curves, neither does any of other unshown plots for other subjects, conditions or variables. Further, as a formal measure of the visual impression of a lack of session effect, we performed Binomial tests on the ranks of these nine curves at time points 0.1, 0.3, 0.5, 0.7 and 0.9, separately. The corresponding p-values of each test at a different time point are around 0.73 or higher; also, they are insignificant for other subjects, conditions and variables. Hence, we can conclude that no session effect exists, meaning that we can combine all sessions in processing the data.

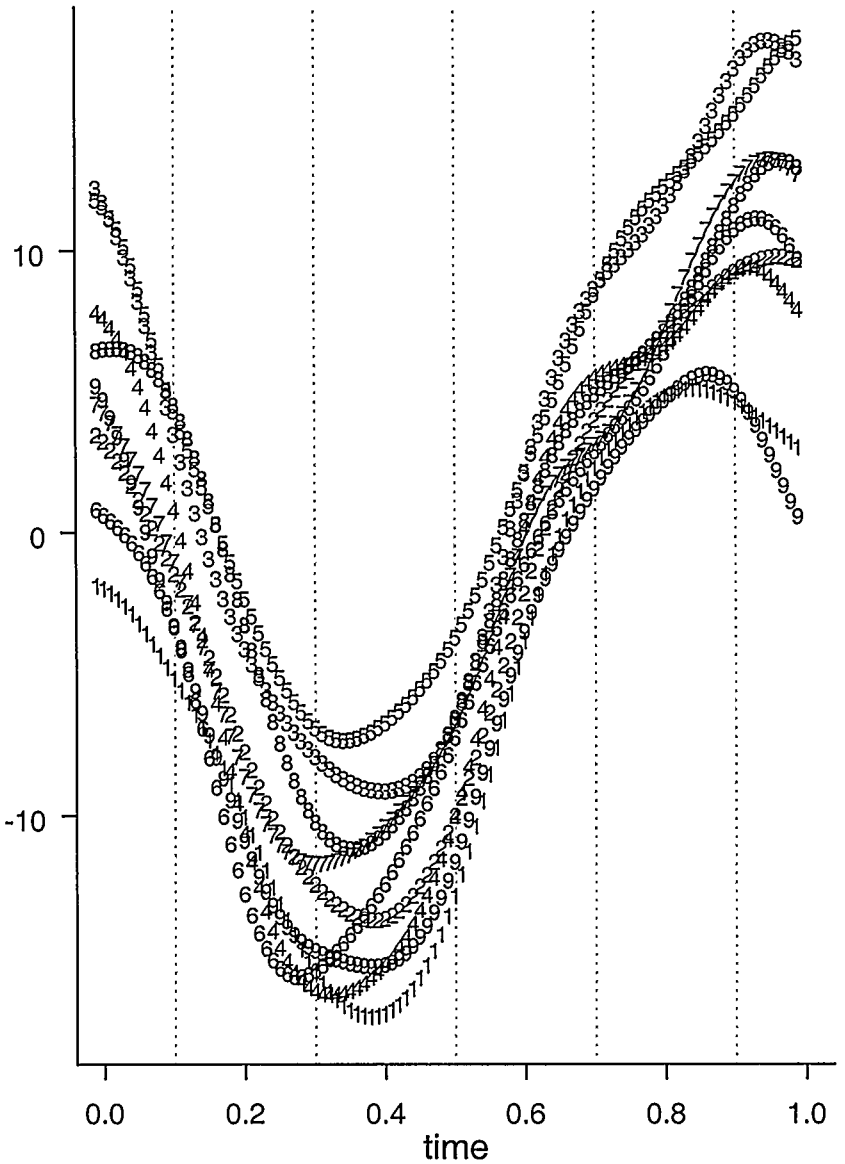

Figure 2.7: Mean curves of nine sessions (identified by numbers) for Subject 20 with posting condition in foot inversion.

Now, we can start the normal first task, which is to smooth the raw data so as to obtain the functional form. As previously mentioned, choosing the values of smoothing parameters is quite important, and it depends mainly on the features of the raw data and how much detailed information we want to keep. Ideally, one should do smoothing on the raw data for each variable, each condition, each subject and each trial. The functional data need to be registered later, and registration requires iteration, which is extremely time consuming in S-PLUS. Although the language Matlab can do a faster job at registration, we chose S-PLUS because of our familiarity with it. Due to the concern of time issue, we decided to process a randomly selected sample of 25 out of 108 trials (when there is no missing session or trial) per subject, per condition and per variable. The size of the random sample, 25, is deemed to be adequate to keep a reasonable amount of information, through our visualizing the rough plots of some summary statistics such as mean curves of the sample and all the trials.

By looking at the raw data of different variables, one should use appropriate parameter values according to the features shown from different variables while performing smoothing on them. For variables having too "wiggly" curves, more smoothness is required to present a clearer trend. We applied the penalized B-spline basis expansion method to the sample data. According to Araki [Ar], it is proper to use the 25 B-spline basis functions of order 6 on all the variables. Figure 2.8 compares the raw data to the smoothed data for Variable 1 of one subject, and it is evident that these basis functions closely resemble in shape the raw data trials.

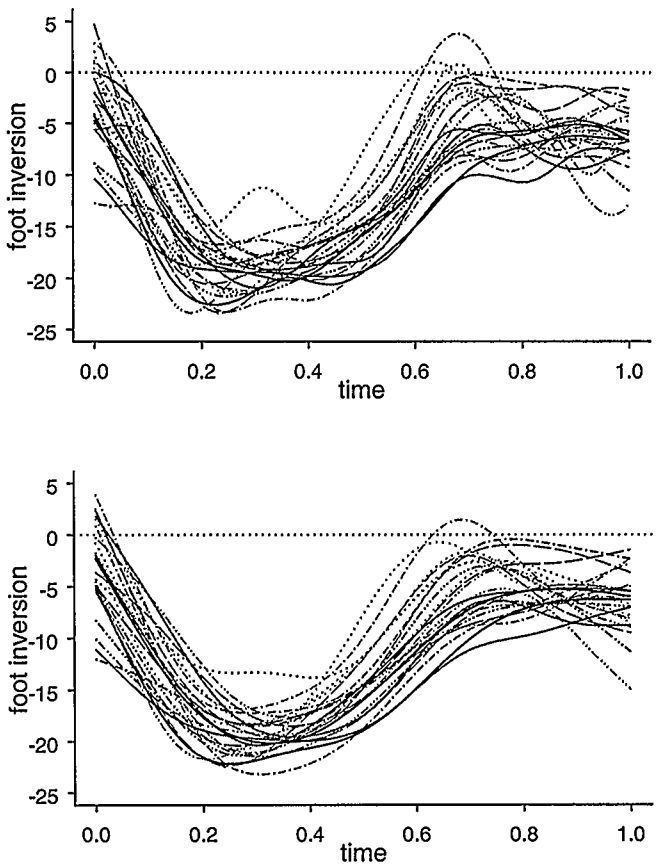

Figure 2.8: Effect of B-spline smoothing for foot inversion (Variable 1) with control condition. Top: raw data. Bottom: smoothed data with $\lambda = 10^{-6}$ and order = 6.

For different variables, the amount of smoothness required varies according to the features exhibited in the variable. Figure 2.9 shows the raw and smoothed data for Variable 2. Obviously, there is more difference between raw and smoothed data shown in this figure than in the previous figure. Because the raw data of Variable 2 are more "wiggly" than those of Variable 1, the amount of smoothness imposed on Variable 2 is greater than that for Variable 1 to make adequately representative functions.

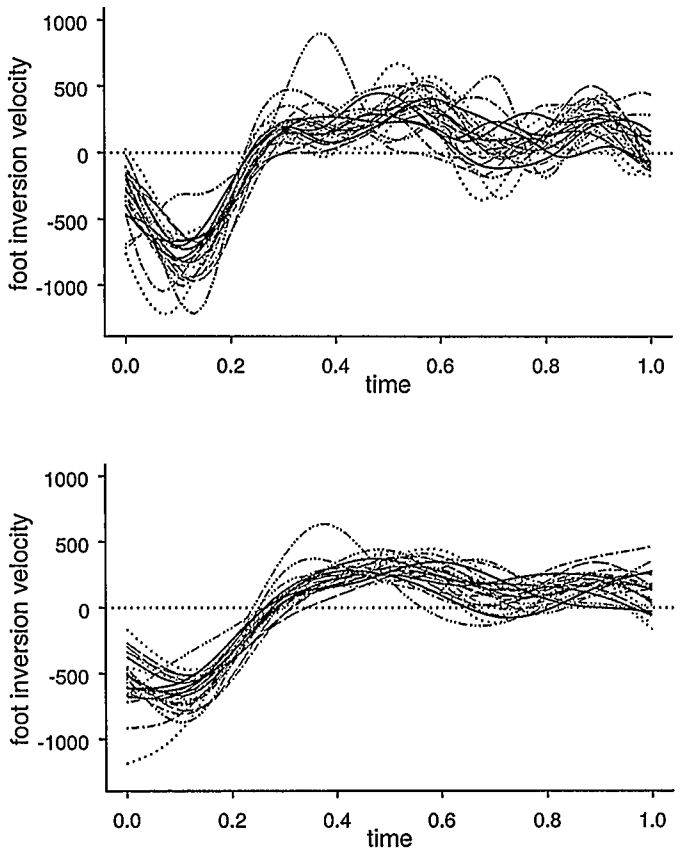

Figure 2.9: Effect of B-spline smoothing for foot inversion velocity (Variable 2) with control condition. Top: raw data. Bottom: smoothed data.

Register sample curves

At this stage, all the data are represented in functional form. As in any data analysis, the important aims for the orthotics data are to estimate the average features of each foot orthotics, and to get an impression of their variability across variables. These tasks are straightforward for univariate and multivariate data. However, Figure 2.10 (top) illustrates a challenge that commonly occurs with functional data. The problem is that the human gait curves exhibit two types of variability. The first type, called

amplitude variability, pertains to merely the intensity of particular features such as the peak in knee external rotation moment (Variable 6), ignoring their timings. The other type, called phase variability, as opposed to the first type, is the variation in the timings of salient curve features without considering their sizes. Before undertaking nearly any analysis, it is essential for us to separate these two types of variation so that features such as "bump" occur at roughly the same "time" for all curves. The technique involves a curve registration process which removes phase variation from the data.

The phase variation does not disappear, though; it is captured by a transformation of time t, which we call a *time warping function.* The following is one way to express the curve registration problem formally. Let $h_i(t)$ be a transformation of time t for curve i, x_i. We assume there exists a standard interval $[0, T_0]$ over which the argument t ranges, while the values of $h_i(t)$ range over x_i's interval $[0, T_i]$. There are constraints $h_i(0) = 0$ and $h_i(T_0) = T_i$ needed to be satisfied. Thus, $h_i(t)$ maps the standard interval $[0, T_0]$ to x_i's interval $[0, T_i]$. The general registration task is to estimate such an h_i for each curve x_i so that the de-warped components x_i can be studied separately [RL].

This general registration method involves using the entire curve, rather than just the location of certain features. Although more technical, the method is completely automatic, and is especially handy when features are hard to identify in certain curves and/or a large number of curves has to be processed. The effect of this method is shown at the bottom of Figure 2.10. As we can see, the curves, which have several extreme features, are expanded or shrunk to make the horizontal difference as small as possible.

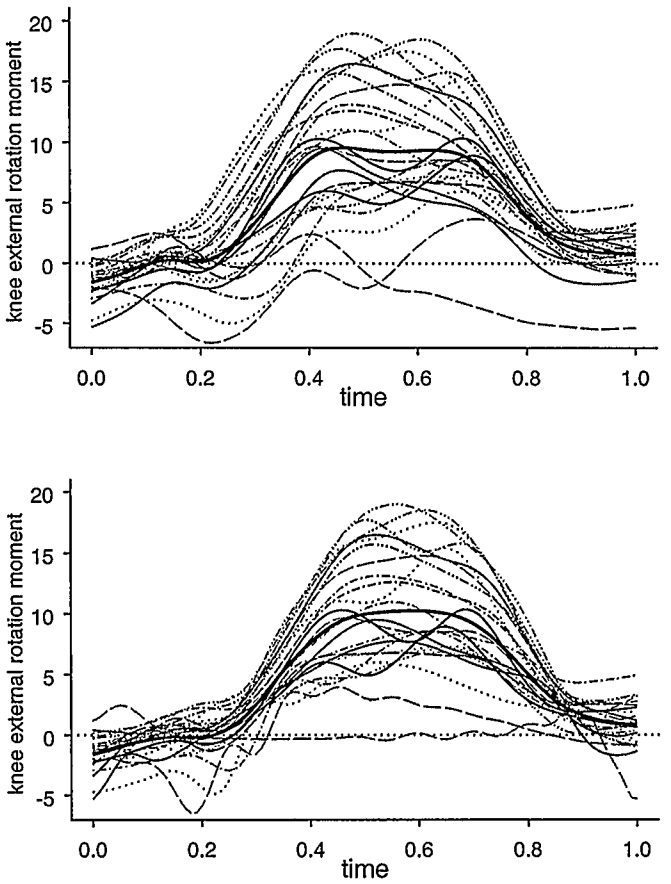

Figure 2.10: Effect of registration on knee external rotation moment (Variable 6), Condition 3, Subject 3. Top: unregistered data. Bottom: registered data. The heavy solid lines are the cross-sectional means, and the bottom one is a better summary of the curves than the top one.

There are other methods for registering curves. One called marker or landmark registration, requires the identification of the location of a number of visible features, such as peaks or valleys, in each curve to be registered. This method is easy to use and understand. However, it can present some problems. Marker events may be missing from certain curves; landmark identification will be a time-consuming and

tedious exercise when large number of curves are to be registered.

The simple time shifting method is one of the others. It simply aligns all curves at one single target time point, ignoring other interested features. This method can be applied when dealing with cyclic data, but in practice, it is inappropriate to use it for most data.

Reduce the dimension of functional data

The original raw data set is presented as a six-dimensional array of dimensions (20, 9, 12, 4, 15, 101) representing (subject, session, trial, condition, variable, time). Combining all sessions results in a reduction of session dimension, and the dimension of the time variable is reduced by forming smoothed and registered time functions instead of keeping discrete time values. Furthermore, the resulting 25 sample functions can be summarized as a single function by taking their mean function, and thus the trial dimension is eliminated.

It is particularly convenient to take the mean instead of the median for calculations while using a basis expansion method to estimate functions. The same basis functions with the mean coefficient matrix can be used to construct the mean function in a basis expansion.

The calculation for the median function in terms of basis expansion functions is not as straightforward as for the mean function. Although in general the median is more robust than the mean, the sample size we use is large enough to make the mean quite representative.

Register the summary functions

There are 80 summary functions (20 subjects $\times$ 4 conditions) for each variable resulted from the above steps. Before any statistical method is applied for data analysis, a second registration may be needed to synchronize the summary functions. This second registration could be over conditions or subjects, depending on which one is to be compared. For example, in order to compare different conditions, the registration may be performed over subjects. Figure 2.11 plots mean functions of four conditions separately for each subject, and it illustrates that subjective variability is greater than condition variability for both Variables 1 and 4.

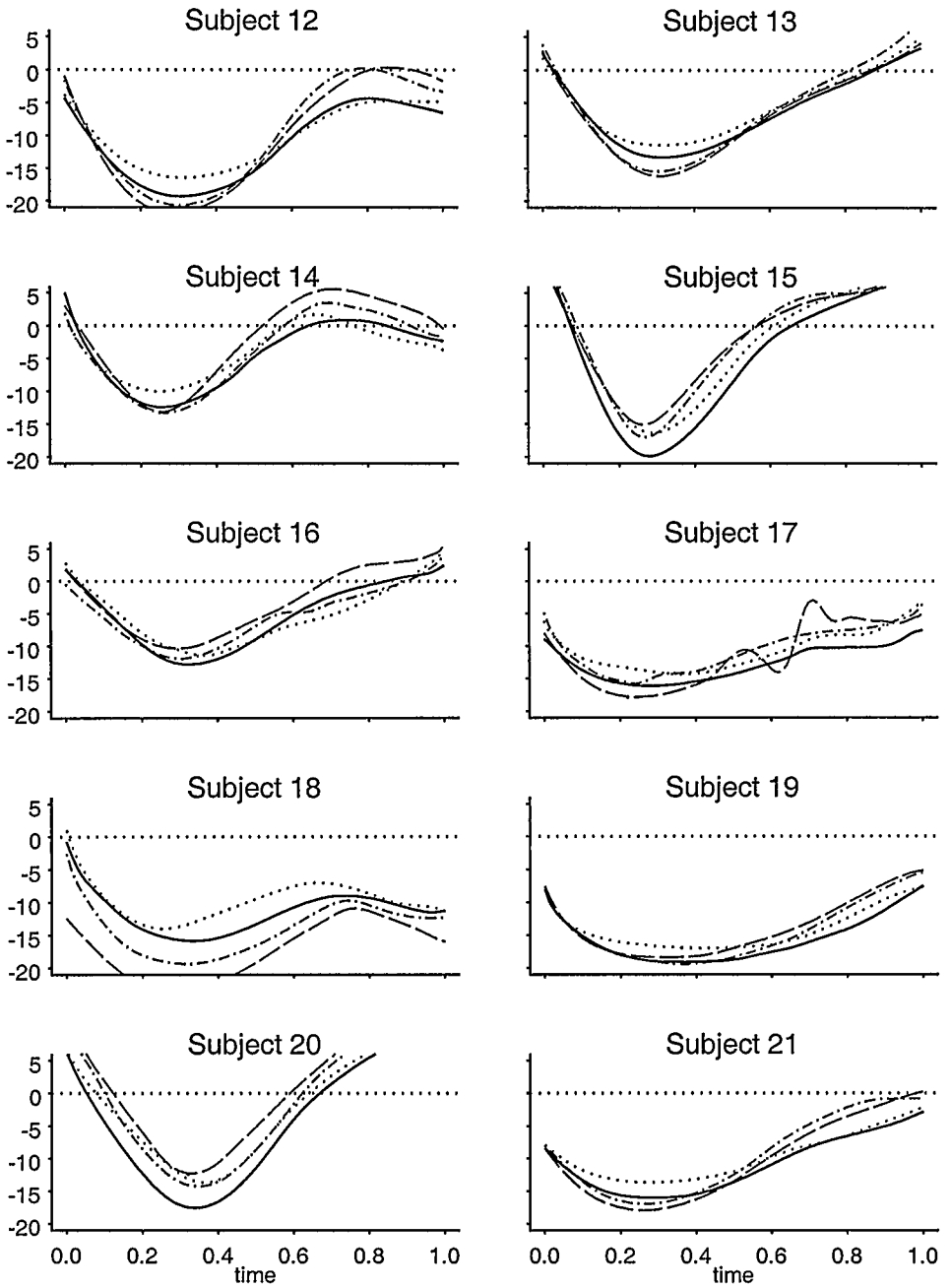

Figure 2.11: (a) Subject and condition variation for foot inversion (Variable 1): Subjects 12 to 21. Solid line: C1(Condition 1); dotted: C2; dot-dashed: C3; dashed: C4.

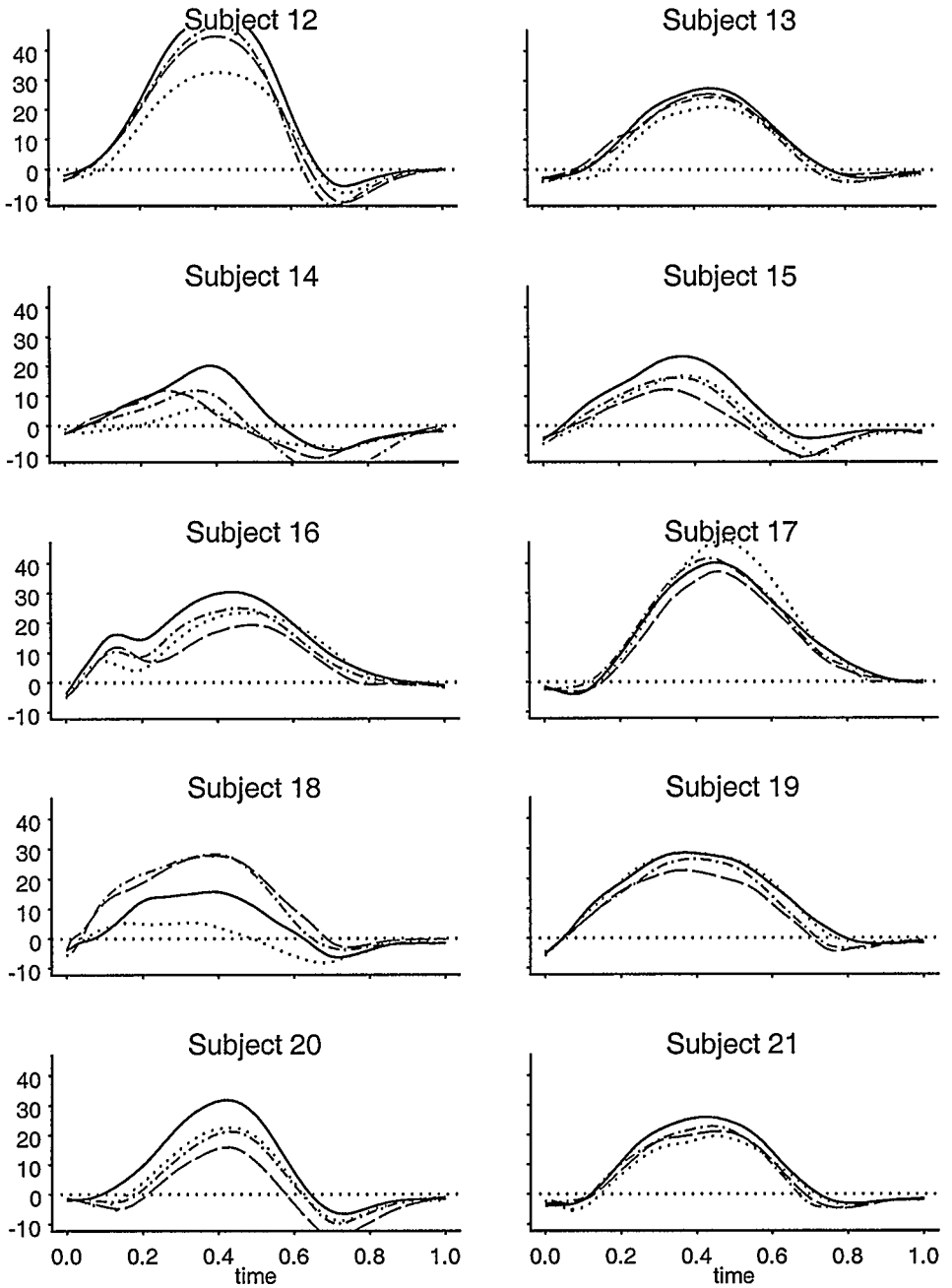

Figure 2.11: (b) Subject and condition variation for abduction moment in the knee joint (Variable 4): Subjects 12 to 21. Solid line: C1(Condition 1); dotted: C2; dot-dashed: C3; dashed: C4.

Chapter 3

Data Analysis

In the previous chapter, we have outlined the procedure for preparing the foot orthotics data for analysis. The classical summary statistics can be used on functional data. Our goal is to investigate whether there are differences between the effects of shoe insert conditions. If there are, we need to investigate how the conditions differ. Some standard statistical methods are introduced in this chapter to help us reach the goal.

3.1 Functional paired t-test

To achieve the goal addressed above, for simplicity, we first consider the ordinary paired t-test to compare two conditions.

There are several naive approaches to handling this kind of testing problem. The first naive approach is to treat each sample curve as a long multivariate vector, and then use a multivariate technique such as Hotelling's T^2 test. However, this approach has two serious drawbacks: it completely ignores the continuity of the values assigned at neighboring time points, and the dimensionality is typically much larger than the sample size [FL]. The second naive approach in common practice is to locate the maximum value or maximum slope of the response curve, and then perform a t-test on that value [Ar]. This approach results in loss of information because only one single value, instead of all the values contained in the response

curve, is used. Functional data analysis uses the entire curve or at least a union of intervals of interest to yield a more powerful overall testing procedure. An objective of this section is to propose a simple and powerful approach to properly combine the test statistics at different time points to obtain an overall test. This is then extended to the comparison of multiple groups of curves.

A large body of literature on longitudinal data analysis has developed various useful testing procedures (e.g., Diggle et al. [DL]; Hand and Crowder [HC]; Schmid [Sc]). The procedures can also be applicable to our functional data analysis setting. They usually treat longitudinal data as a multivariate vector and do not incorporate a dimensionality-reduction technique. For functional data analysis, the dimensionality is high, and hence dimensionality-reduction techniques are required. Although powerful for analyzing longitudinal data, traditional tests for high-dimensional problems need some tuning. Faraway [Fa] proposed smoothing on the functional data first and then using traditional analysis of variance (ANOVA).

According to Muendermann [Mu], the effects of posting (Condition 2) on most kinematic variables are significant and consistent across subjects. However, she also points out that these effects seem to be only present during the first half of the stance phase as maximum absolute foot inversion is significantly reduced, but posting does not affect foot inversion during the second half of the stance phase. Therefore, for the functional t-test, we chose foot inversion (Variable 1) and control and posting (Conditions 1 and 2) to verify the statement of Muendermann. At this stage, the data at hand were processed by the procedure outlined in Chapter 2 as follows. For a given subject with Condition 1 and Condition 2, 25 smoothed sample curves for each condition were registered, and a mean function over these registered curves

was obtained. Therefore, there are two sample mean functions resulting from the above step for the given subject. Furthermore, the mean difference function can be achieved as $\overline{y_{dif}}(t) = \overline{y_{c1}}(t) - \overline{y_{c2}}(t)$. Repeating this procedure for all subjects turns out twenty such functions, and they are plotted in the top panel of Figure 3.1. These functions were then registered over subjects (see Figure 3.1, bottom). Finally, the curves were discretized evenly by a grid of 101 points that are equal to those in the data. We chose such an evenly spaced discretization because it is simple and it captures all the essential features of interest in the curves.

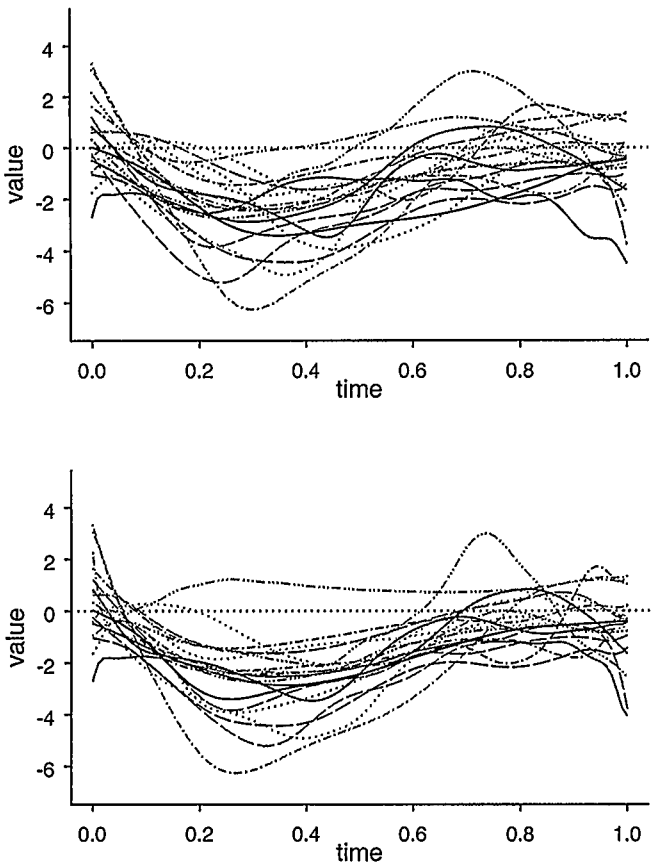

Figure 3.1: Mean difference curves. Top: unregistered. Bottom: registered.

A functional paired t-test (sample size n = 20) is then performed at each of the 101 time points on the response values from all the subjects. This process yields 101 resulting observed t-scores and corresponding p-values, which are plotted versus time in Figure 3.2. We use one-sided p-values based on what is seen from Figure 3.1 and Muendermann's statement about posting's (Condition 2) significantly reducing foot inversion. To see if there is reduction between the responses of two conditions, one can measure the proportion of those p-values below a specified criterion such as 0.05 within the whole time interval. What is seen from Figure 3.2 is generally in agreement with Muendermann's finding that the effect of posting on foot inversion is significant during the first half of stance phase. This figure shows that posting affects foot inversion significantly for the time interval (0.1, 0.7). The one-sided p-value is less than the significance level 82% of the time.

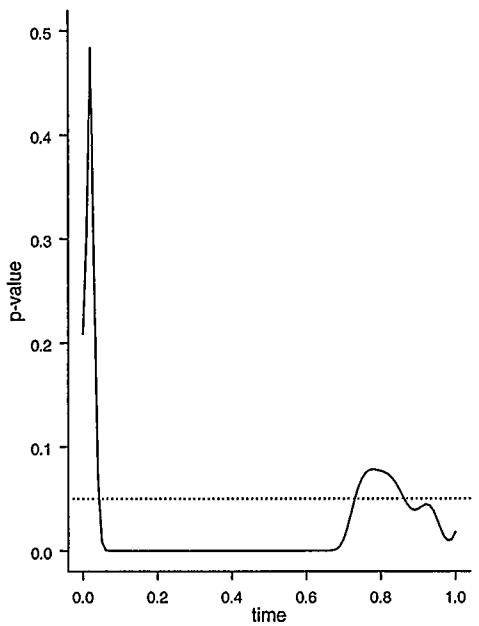

Figure 3.2: The p-value functions.

In the above procedure we used a paired t-test across subjects comparing each subject's mean curves (over 25 trials) under the two conditions. In addition, we can perform a t-test analysis separately for each subject, resulting in twenty separate p-value curves. For a given subject, at each time point there are 25 values for each condition. An ordinary two-sample t-test yields the corresponding p-value. We assume equal population variances for the two conditions and therefore use the pooled sample variance s_p^2 since, as shown in Figure 3.3, the pooled variance functions vary considerably over time.

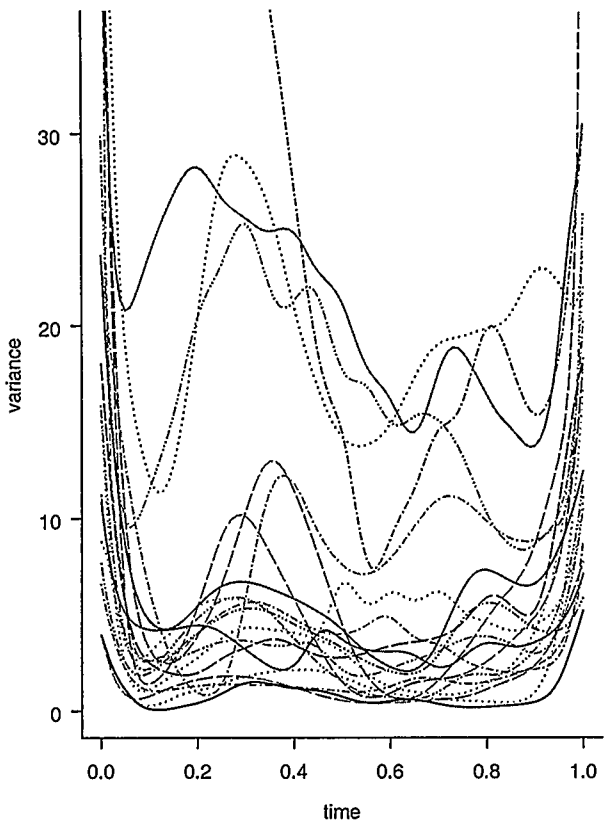

Figure 3.3: Pooled variance functions.

The mean difference functions from the upper panel of Figure 3.1 and their corresponding ± 2 standard deviation functions (dot-dashed lines) are shown in Figure 3.4. All subjects differ in Condition 1 and Condition 2 in most of the stance phase. Also, the direction of the difference is consistent, except for Subject 11. The ordinary two-sample t-test results in 101 pointwise p-values corresponding to each time point, and these are smoothed again. The percentage in Figure 3.5 represents the proportion of the interval for which the p-value is less than 0.05. Such a percentage for each subject is plotted in Figure 3.6, and it shows that 17 of 20 subjects have percentage exceeding 40%. For the remaining three subjects, there's not much difference between Condition 1 and Condition 2.

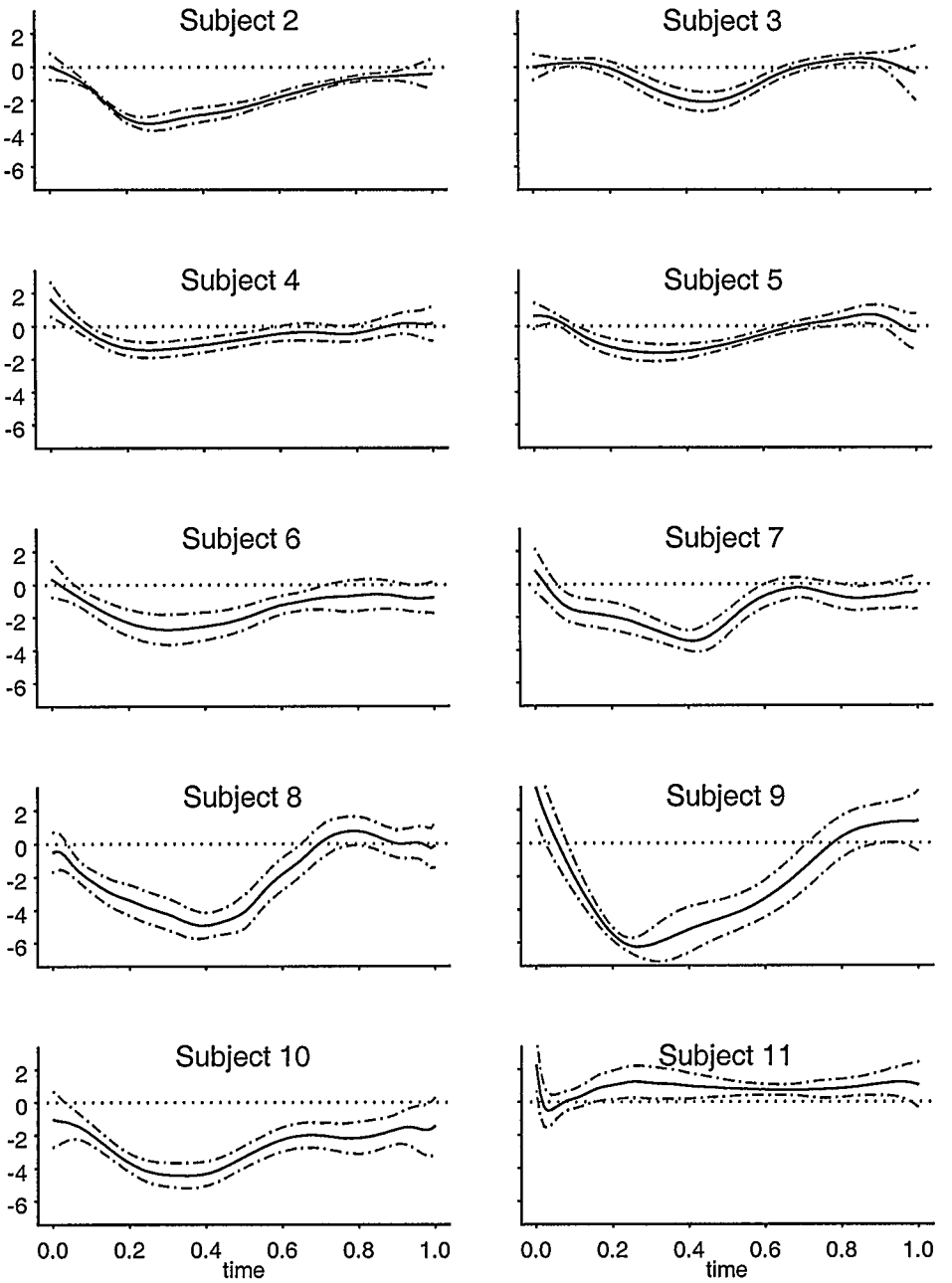

Figure 3.4: (a) Mean difference curve with ±2 standard deviation.

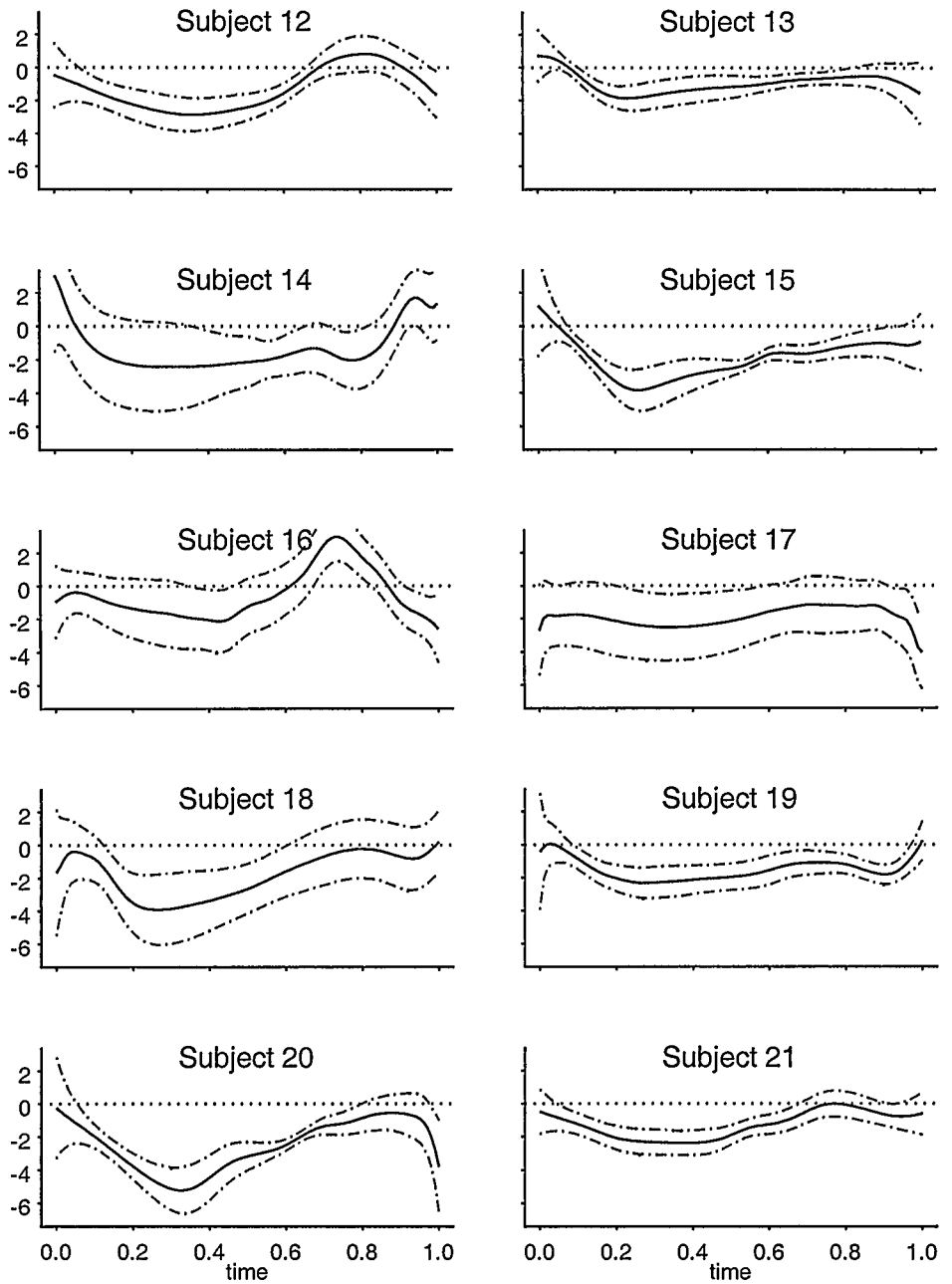

Figure 3.4: (b) Mean difference curve with ±2 standard deviation (continued).

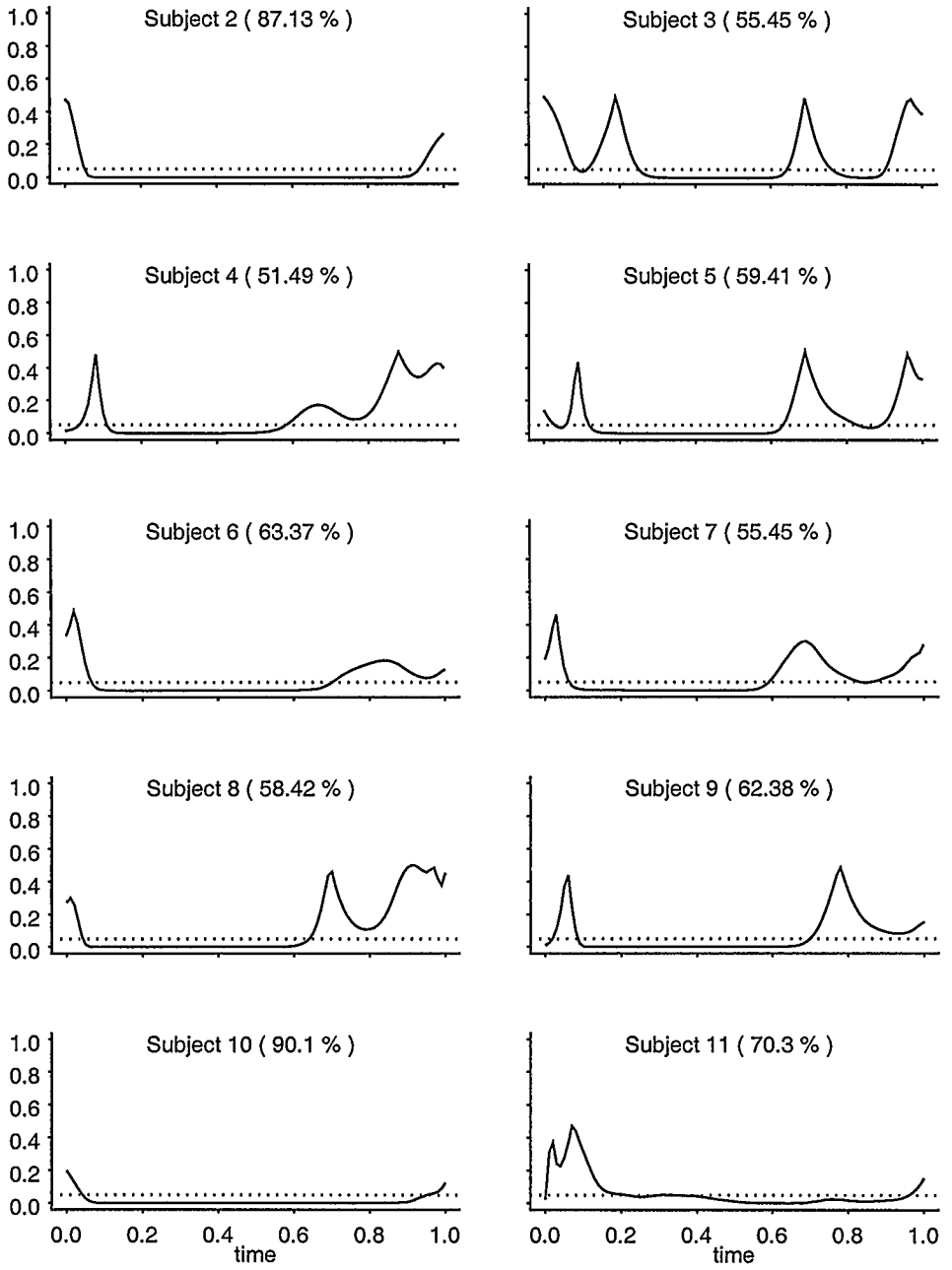

Figure 3.5: (a) Smoothed p-value functions.

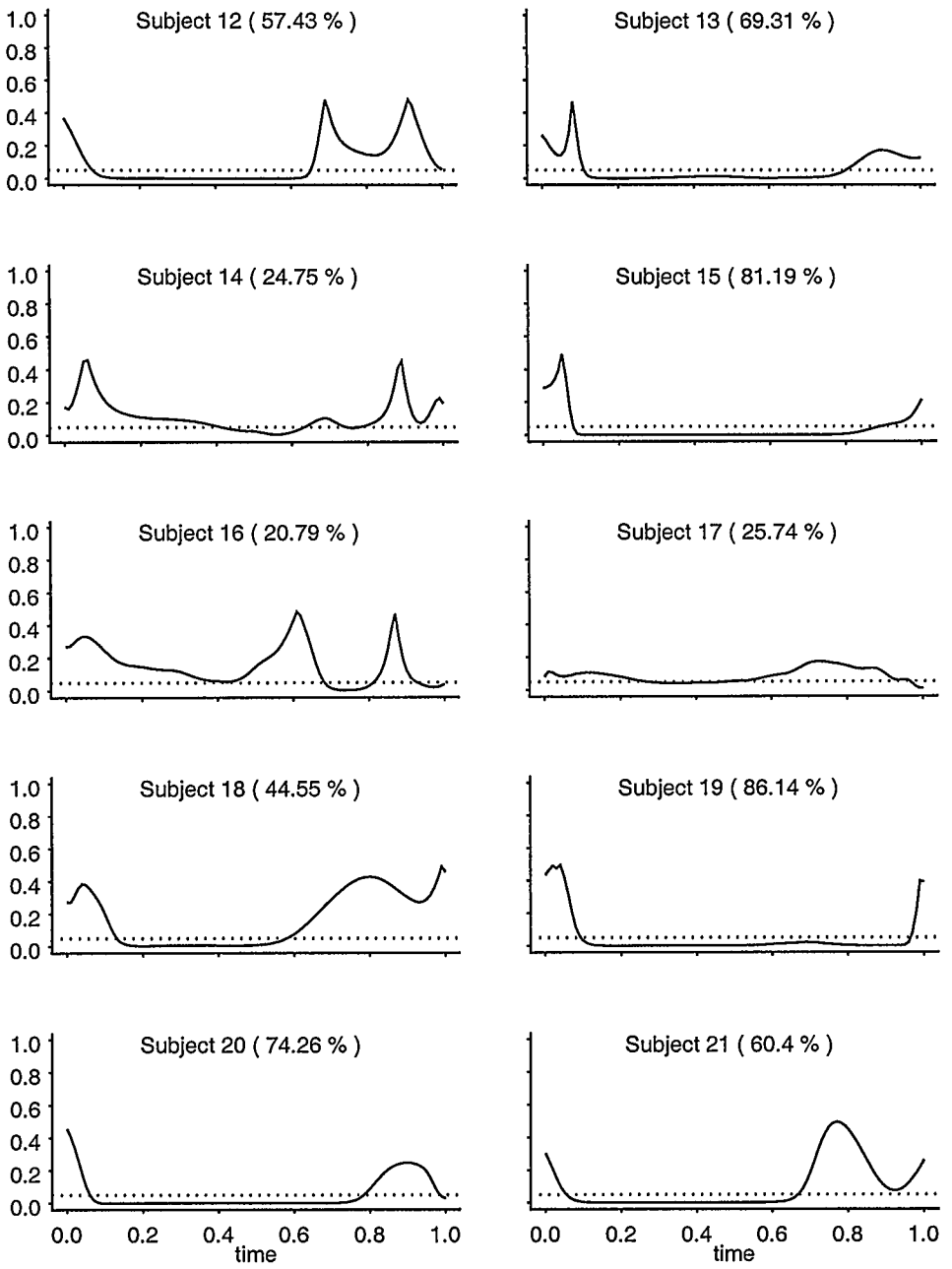

Figure 3.5: (b) Smoothed p-value functions (continued).

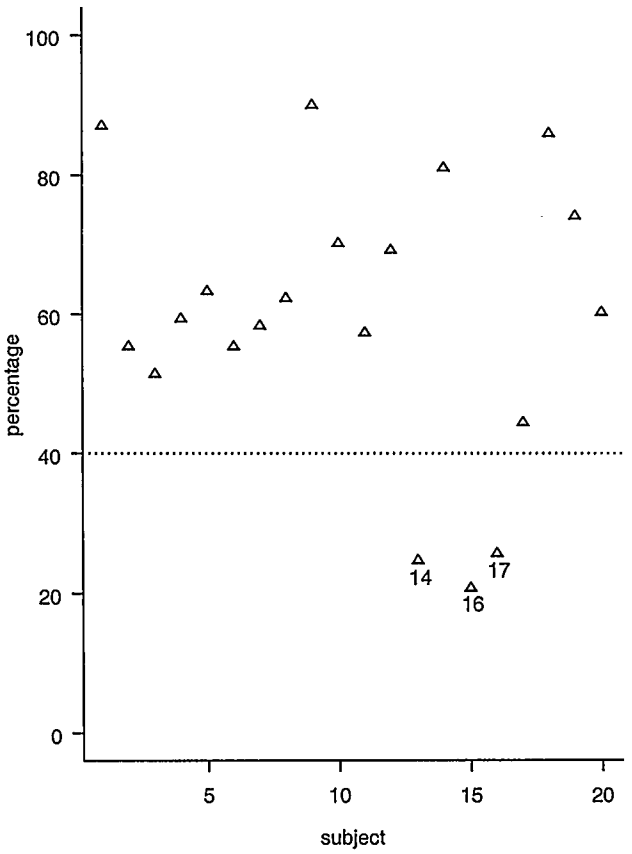

Figure 3.6: Percentage of time that the p-value is less than 0.05: all subjects.

3.2 Correlation analysis

Earlier studies found small and frequently not significant differences in kinematic variables when comparing different orthotics conditions [SC, NK]. In order to verify this finding, we carry out pointwise correlation analysis of two variables for each subject.

Foot inversion (Variable 1) and ankle inversion moment (Variable 4) are studied as responses because McClay [Mc] speculated that increased inversion moment at the ankle joint may be related to increased magnitude of foot inversion. In addition, due to the result that posting significantly reduces Variables 1 and 4 from Muendermann [Mu], we select control and posting (Conditions 1 and 2) for comparison. The bivariate data consist of 25 pairs of observed functions (x_i, y_i). The way in which these depend on one another can be measured by the pointwise sample *correlation* function

$$corr_{X,Y}(t_i) = \frac{cov\{X(t_i), Y(t_i)\}}{\sqrt{var\{X(t_i)\}var\{Y(t_i)\}}}, \tag{3.1}$$

where $cov\{X(t_i), Y(t_i)\}$ is the *covariance* function

$$cov\{X(t_i), Y(t_i)\} = (K-1)^{-1}\sum_{k}^{K}\{x_k(t_i) - \bar{x}(t_i)\}\{y_k(t_i) - \bar{y}(t_i)\},$$

and $k \in \{1, \cdots, 25\}$, $i \in \{1, \cdots, 101\}$. For each subject, the sample pointwise correlation function is computed as follows: first, for every trial, the two response functions are converted to 101 discrete values, equally spaced at the data time points; second, the correlation $corr_{X,Y}(t_i)$ is computed at each t_i using Equation (3.1); third, the correlation function is smoothed. The correlation functions obtained from the above steps for all subjects are shown in Figure 3.7. Disappointingly, there is no general trend seen in these correlation functions, which indicates that the pointwise sample correlation function is not a useful discriminant for comparing a bivariate response for different conditions. Figure 3.8 basically carries the same signal, presenting the correlation functions for two more variable response pairs for the first ten subjects. Again, it is hard to discriminate the correlations between three combinations.

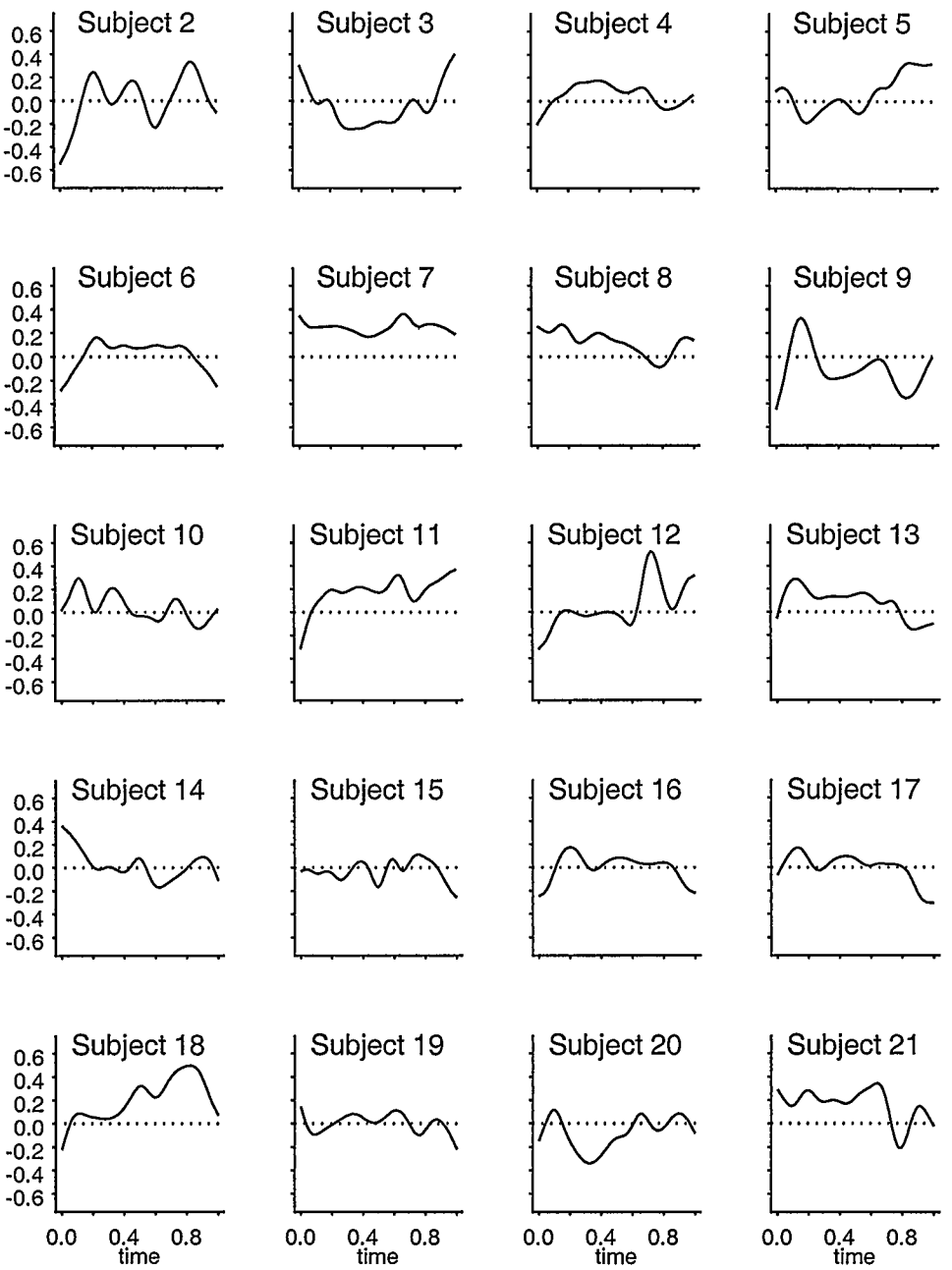

Figure 3.7: Sample correlation functions for all subjects with Variables 1 and 4 in Condition 2.

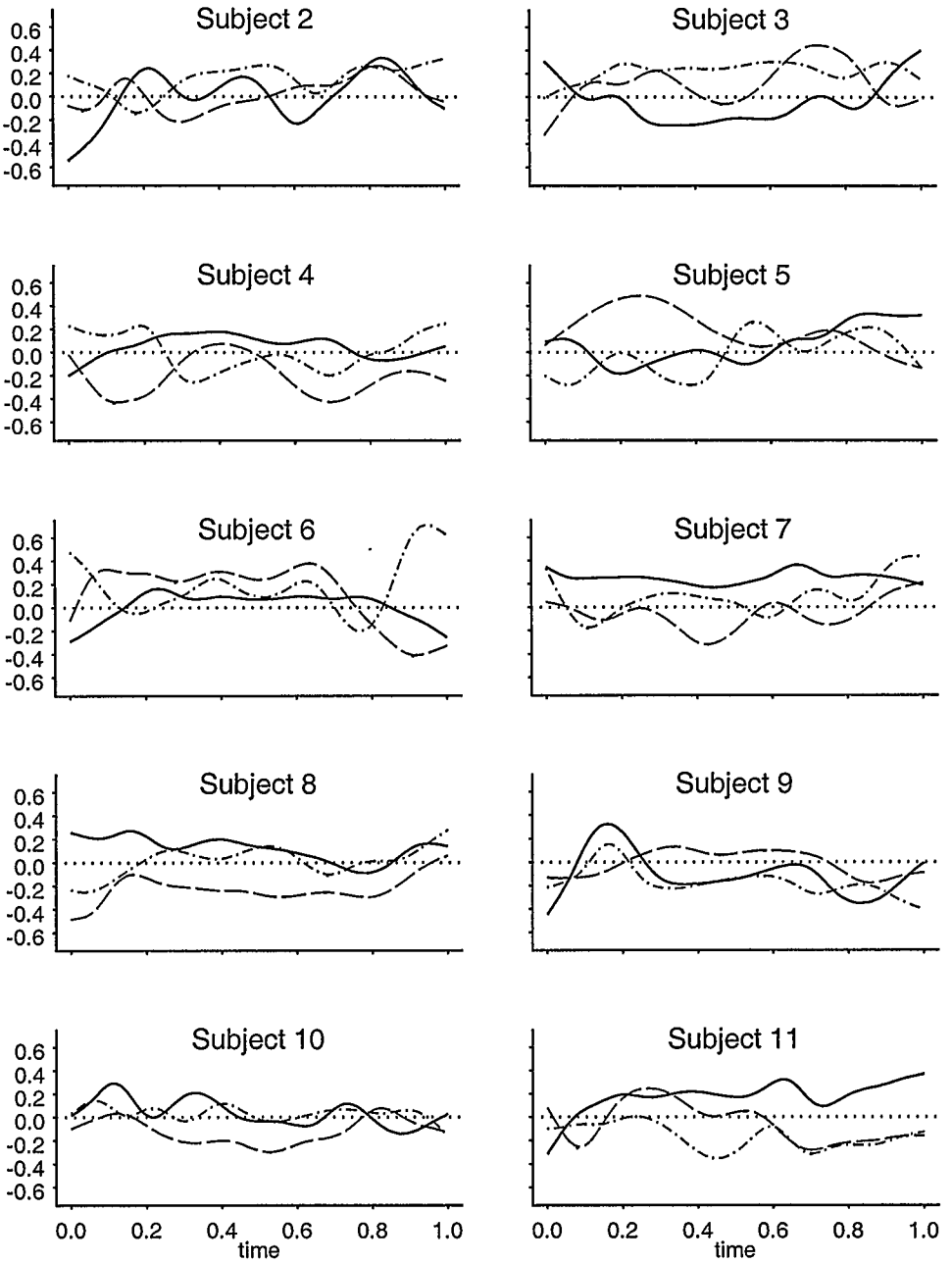

Figure 3.8: Sample correlation functions for first ten subjects with variable pairs of Variables 1 and 4 (solid line), Variables 1 and 6 (dot-dashed line), and Variables 4 and 6 (dashed line) in Condition 2.

Integrated squared correlation function

The previous two figures only allow us to visually interpret the sample correlation functions. In order to provide a summary measure, Araki [Ar] proposed the integrated coefficient of determination (ICOD), which is equivalently defined as a uniformly averaged pointwise coefficient of determination,

$$(1/T)\int \rho^2(t)\,dt.$$

$\rho(t)$ can be estimated by the sample correlation $r(t)$.

Comparison of conditions

Now, we investigate the usefulness of the integrated coefficient of determination for comparing control and posting (Conditions 1 and 2) with a bivariate response. Variables 1 and 4 are used with respect to the ICOD primarily due to the finding in McClay [Mc] mentioned above. The two sets of ICOD values are plotted in Figure 3.9. Both of them are very close to zero, meaning that Variables 1 and 4 are essentially uncorrelated for both conditions. Also, due to the random trend shown by these two sets of ICOD, we conclude that ICOD is not an efficient measure in the sense of discriminating conditions. A two-sided paired t-test on the 20 pairs of ICOD values is performed and results in a p-value of 0.1131, indicating insignificant difference between these two conditions. This is consistent to the result of a Wilcoxon signed rank test, which gives a p-value of 0.1893.

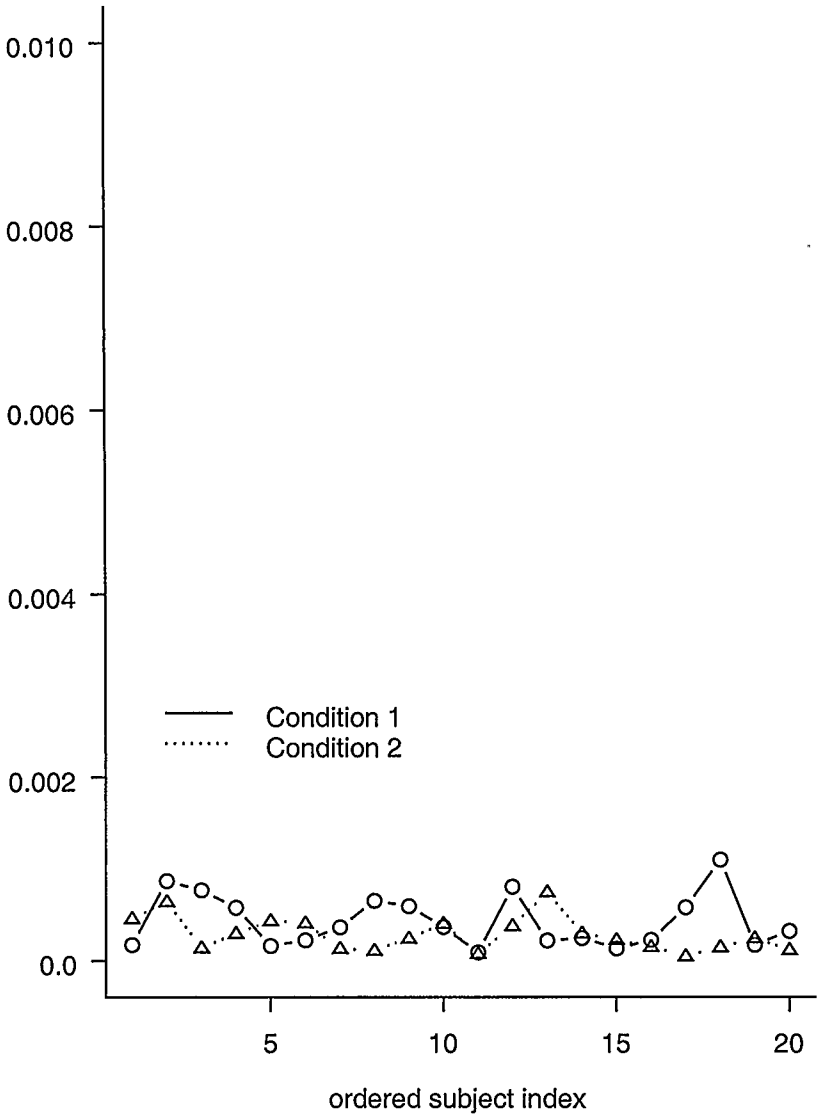

Figure 3.9: ICOD values of V1 and V4 in the comparison of two conditions.

Chapter 4

Functional Principal Components Analysis

The analysis considered previously gives a glimpse of ways in which the variability of our functional data set is interesting, but there is a need for more detailed and sophisticated ways of investigating variability. One of the ways is functional principal components analysis (FPCA), and it is the major theme of this dissertation.

There are several reasons to consider the FPCA technique. First, after smoothing and registering the data, we want to explore further to see the functional features of variability, no matter whether these features are surprising or not. It is also necessary to indicate the complexity of the data, in the sense of finding types of curves and their characteristics. Principal components analysis meets these ends desirably. Perhaps due to these reasons, this technique was considered in the early literature on classical functional data analysis (FDA) as the first choice. Second, the covariance structure can be detected by principal components analysis in a more informative way, and it overcomes the common problem that it is difficult to interpret the variance-covariance and correlation functions in the classical multivariate case.

In this section, we first introduce the PCA in multivariate statistics. With the usual questions about how classical PCA works in the functional context, we then discuss the modifications for FPCA. When applying FPCA in the foot orthotics data, we start with a univariate manner and then go on to its extension to the bivariate case.

4.1 PCA for classical multivariate data

Johnson and Wichern [JW] demonstrate the basic philosophy of principal components analysis as follows: to use a few uncorrelated linear combinations of the correlated response variables to explain their variance-covariance structure. The linear combinations are supposed to provide some useful interpretation themselves, and it is most useful if relatively few linear combinations explain most of the variability. Therefore, PCA's general objectives are twofold – data reduction and interpretation.

It is a central concept of PCA in multivariate analysis to consider the linear combinations of variable values,

$$f_i = \sum_{j=1}^{p} \beta_j x_{ij} = \langle \beta, x_i \rangle , \qquad i = 1, \cdots , N, \tag{4.1}$$

where $\{f_i\}$ is an *uncorrelated* set of linear combinations of the observed values x_{ij} of the jth variable with weighting coefficients β_j. The value of the linear combination f_i is called *principal component score*, and it helps in describing what the linear combination or principal component means in the sense of variation characteristics of replicates. The weighting coefficients are chosen so as to highlight the components of variation in the data. The following steps demonstrate how principal components analysis can be defined through sets of normalized weights that maximize variation in the linear combinations f_i:

1. Find the first principal component, which was the weight vector $\xi_1 = (\xi_{11}, \cdots , \xi_{p1})'$ for the linear combination f_{i1} with maximum mean square

$$N^{-1} \sum_i f_{i1}^2 = N^{-1} \sum_i \langle \xi_1, x_i \rangle^2 ,$$

subject to the normalization constraint

$$\sum_j \xi_{j1}^2 = \|\xi_1\|^2 = 1.$$

This step motivates identifying the most important mode of variation in the variables. It is convenient to restrict attention to the weight vectors of unit length because there is a problem of indeterminacy in the sense that $N^{-1}\sum_i f_{i1}^2$ could otherwise be increased arbitrarily.

2. Find successive weight vectors as follows: for the mth step, a new weight vector ξ_m is computed for new values f_{im} with maximum mean square again, subject to the constraint $\|\xi_m\|^2 = 1$ and the $m-1$ additional constraint(s)

$$\sum_j \xi_{jk}\xi_{jm} = \langle \xi_k, \xi_m \rangle = 0, \qquad k < m. \tag{4.2}$$

Using these steps, we are investigating the most important modes of variation again, but the amount of variation declines on each step because the weights defining the variation are required to be orthogonal to those from previous steps. In this way, we see a new component of variation at each step. Usually, the number of these steps carried out need not be up to the number of variables, p, since we expect that the first few principal components account for most of the total variability, and thus comprise most of the information in the data.

The weight vectors defined by the above procedure are eigenvectors of the sample covariance matrix.

4.2 Defining PCA for functional data

How does PCA carry over to FPCA? Instead of using variable values x_{ij}, FPCA computes the linear 'combination' f_i of function values $x_i(s)$, and f_i denotes the ith principal component of the variable function over continuous values indexed by s. Integration over s, rather than summation over j in Equation (4.1), is then used to define the ith principal component

$$f_i = \int \beta(s)\, x_i(s)\, ds = \langle \beta, x_i \rangle, \tag{4.3}$$

where $\beta(s)$ is a weighting function here. The counterpart of the normalized weight vector ξ_j is then $\xi(s)$, called the normalized weight function. As for multivariate PCA, the weight functions $\xi(s)$ are also required to satisfy the same kind of constraints mentioned in Section 4.1. However, this time the notation $\|\xi\|^2$, called the unit sum of squares in classical PCA, is used to denote the squared norm $\int \xi(s)^2 ds$ of the function $\xi(s)$. Meanwhile, the orthogonality constraints $\langle \xi_k, \xi_m \rangle$ in Equation (4.2) now denote

$$\int \xi_k(s)\, \xi_m(s)\, ds = 0, \qquad k < m.$$

In FPCA, we use the sample covariance function $v(s,t)$ to replace the covariance matrix for classical multivariate data. We write

$$v(s,t) = N^{-1} \sum_i \{x_i(s) - \bar{x}(s)\} \{x_i(t) - \bar{x}(t)\},$$

where $\bar{x}(\cdot)$ is the mean function of $x_i(\cdot)$.

The operation, in PCA, of finding eigenvalue-eigenvector pairs of the sample covariance matrix is replaced, in FPCA, by solving the eigenequation

$$\int v(s,t)\, \xi(t)\, dt = \langle v(s,\cdot), \xi \rangle = \rho \xi(s),$$

where the integration term is a covariance operator,

$$V\xi = \int v(s,t)\,\xi(t)\,dt. \tag{4.4}$$

Therefore, the eigenequation can be written as

$$V\xi = \rho\xi. \tag{4.5}$$

Eigenanalysis problems always concern the maximum number of different eigenvalue-eigenvector pairs. This number in multivariate eigenanalysis is equal to the rank of V, which is also the number of principal components limited by the number of variables, p. However, in the functional context, the counterpart of p is the number of function values. The number of "principal components" is thus infinite. As long as the functions x_i are linearly independent, the rank of the covariance operator V is $N-1$ due to the subtraction of the mean function $\bar{x}$ from N values.

4.3 Computational methods for functional PCA

Ramsay and Silverman [RS2, Chapter 6] illustrate two methods for solving the eigenequation problem above. One approach uses basis expansions in the estimation of the functional principal component curves. It works with a small number of parameters equal to the number of basis functions. The other approach, called the discretization method, was first introduced by Rao [Ra1, Ra2] and Tucker [Tu]. This method, essentially classical in nature, is simply to discretize the function curves to a fine grid of time points equally spaced in interval τ. Then, one can find a solution to the eigenequation $Vu = \lambda u$ with N eigenvectors u. The eigenvectors u are then converted to eigenfunctions by using any convenient interpolation method. Further,

smoothing with basis expansions is applied to the resulting eigenfunctions. Finally, Equation (4.3) is used to compute principal component scores.

The discretization approach is more convenient in S-PLUS, even though there is computational disadvantage in treating a large matrix. We used the discretization method for the foot orthotics data.

4.4 Applying functional PCA to our data

In this section, we apply functional PCA, using the method described in Section 4.3, on the foot orthotics data for both subjective variability and condition variability, separately. We first consider univariate functional PCA with one variable at a time. Foot inversion and ankle inversion moment (Variables 1 and 4) are used separately. The data functions were estimated by B-spline functions using the same parameters as in the preliminary steps in Chapter 2 and in the eigenfunction estimation.

4.4.1 Subject variation

Although our main purpose is to detect the variability between conditions, it is worthwhile to explore subjective variability. For each subject, the 25 registered sample curves for control (Condition 1) are summarized with the mean function. Twenty summary mean functions for all the subjects are obtained and then registered again to the mean function of these twenty. Figure 4.1 presents such registered mean functions for Variables 1 and 4, respectively.

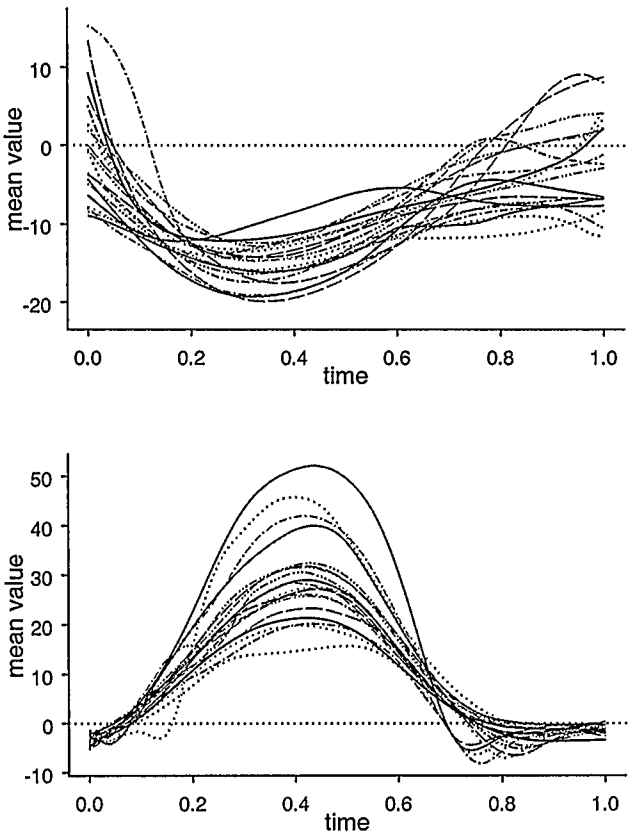

Figure 4.1: Registered mean functions for all subjects. Top: Variable 1. Bottom: Variable 4.

Plotting principal components as perturbations of the mean

Just as in classical multivariate analysis, the weight functions commonly have physical meanings. A discussion on PCA of the data obtained from the study of human gait was developed by Rice and Silverman [RS4]. Figure 4.2 displays the first two smoothed weight functions for Variable 1 in the left panels. The percentages indicate the amount of variation accounted for, and the first two PCs account for 83.4% of the total variation. Each weight function has the task of defining the most important mode of variation in the curves subject to each mode being orthogonal to

those modes defined on previous steps. Note that the weight functions are defined only to within a sign change. Although the first weight function ξ_1 for Variable 1 is positive for the entire time interval, the weight emphasizes the beginning and end of the stance phase. This means that the greatest variation between subjects will be found by heavily weighting the two ends, with only a light contribution from the middle phase of the step. The second weight function ξ_2 highlights the central phase at about time 0.2 to 0.7, and weights negatively for the remaining phase after time 0.7.

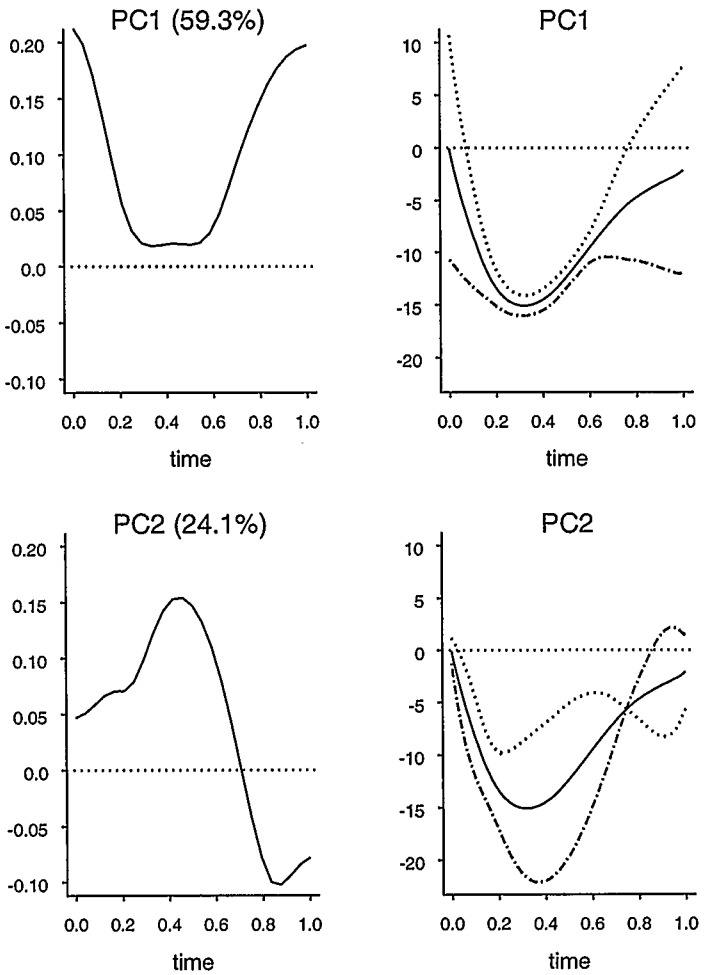

Figure 4.2: Univariate FPCA for subjective variation: Condition 1, Variable 1, all 20 subjects. Left: smoothed weight functions. Right: mean functions with offset curves.

Interpreting the principal components is sometimes challenging, particularly for the latter PCs due to the increased complexity in the weight structure. We now consider techniques for understanding them. One way of interpreting the first principal component (PC1), for example, is to examine a plot of the overall mean curve and two curves (offset curves) obtained by adding and subtracting a multiple of the

weight function for PC1, the counterpart of the loadings in classical PCA [Si1]. Essentially, this shows the effects of PC1 on the "average" case. Subsequent principal components are interpreted in the same way.

The right panels of Figure 4.2 separately show such plots for the first two PCs of Variable 1. In each case, the solid curve is the overall mean function, and the dotted and dashed curves show the effects of adding and subtracting a multiple of each corresponding weight function (plotted in the left panels), respectively. The effect of the first PC is approximately to add or subtract a relatively small constant to the mean function between time 0.2 to 0.6, which means that the mean response of all subjects is uniformly affected for this time interval. For the remaining stance phase, the offset curves indicate more and more departures from the mean function close to the two ends. The second principal component explains 24.1% of the total variation. We note that its effect is confined at about time 0.7, showing similar and then opposite effect with different magnitude to that of the first component for stance phase before 0.7 and after, respectively. The third and fouth principal components account for only 9.7% and 4.1% of total variation, respectively, and are considered negligible; therefore, they are not shown in the figure.

Similarly, the same settings for Variable 4 are plotted in Figure 4.3. Just like the first weight function for Variable 1, ξ_1 for Variable 4 is positive for the entire interval, but emphasizes the central rather than the two ends of the stance phase. A strong mode of variation, 91.1%, is explained by the first component. At the time interval of (0.2, 0.7), the variation is large and becomes the maximum at the peak, reflecting a great amount of subjective variability. The second component contributes ignorable mode of variation, but it is interesting to note that the weight function has negative

influence for intervals (0, 0.2) and (0.5, 0.6) and positive influence for intervals (0.2, 0.5) and (0.6, 0.9). This suggests that the component represents a contrast effect on roughly every other quarter of the entire stance phase.

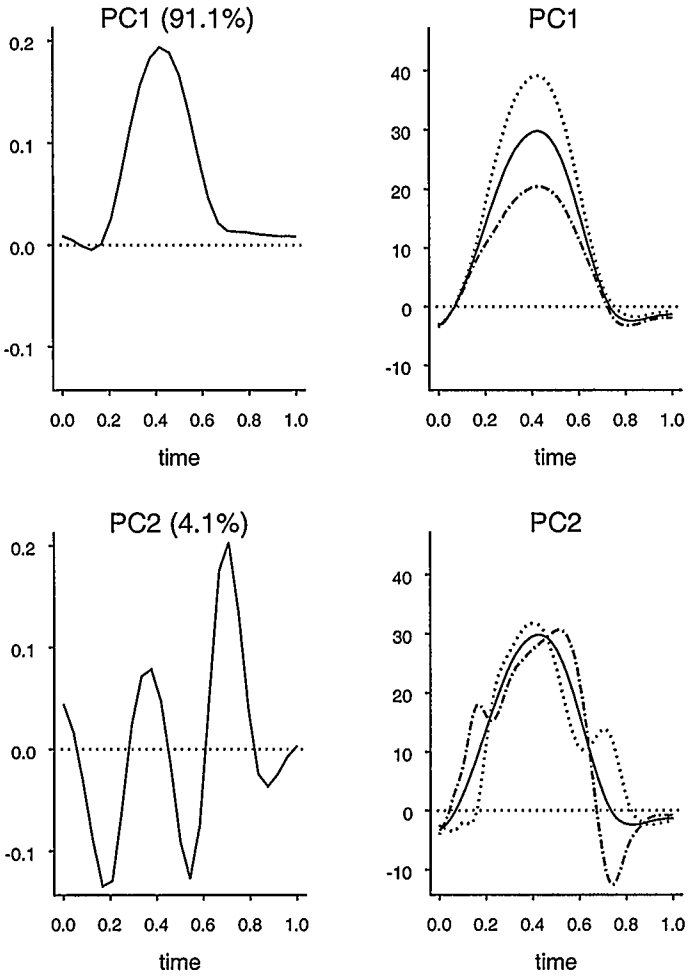

Figure 4.3: Univariate FPCA for subjective variation: Condition 1, Variable 4, all 20 subjects. Left: smoothed weight functions. Right: mean functions with offset curves.

Figure 4.4 compares Variable 1 and Variable 4 in one plot. This plot makes it easy to detect the interesting findings of PCs for these two variables. As we saw in

their corresponding weight functions, PC1 for each variable emphasizes the stance phase in entirely different intervals. Also, for Variable 4, PC1 strongly dominates over PC2, while for Variable 1, contributions between PC1 and PC2 are less different than for Variable 4.

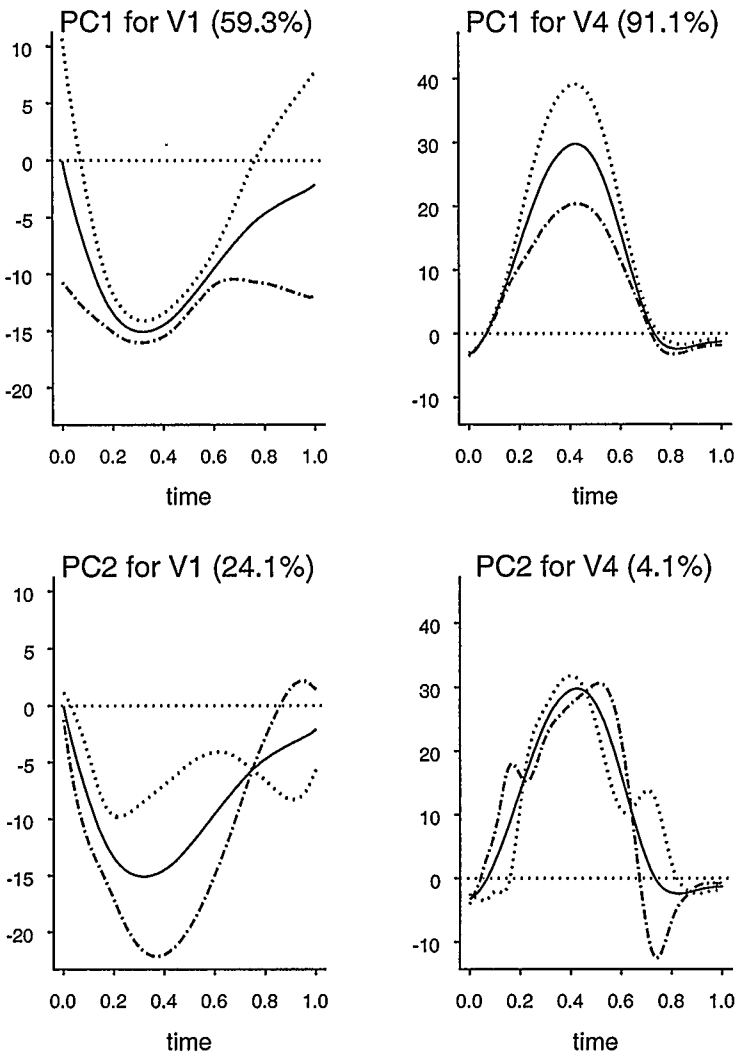

Figure 4.4: Univariate FPCA for subjective variation: Condition 1, all 20 subjects. Left: Variable 1. Right: Variable 4.

Plotting principal component scores

Examining the scores f_{im} is also an important aspect of PCA. Although principal components are uncorrelated, their scatter plots sometimes reveal important structures in the data other than linear correlation. Figure 4.5 is such a scatter plot of the PC scores. Each subject is plotted as a circle, with subject number identified to the extreme values. The first two PC scores for Variable 1 reveal that there are two roughly distinct groups of the subjects, located on the two sides of the vertical line, separately. The highest score for PC1 and lowest score for PC2 both go to Subject 15. For Variable 4, Subject 12 has the highest PC1 score but lowest PC2 score. There is a distinct group centering at the average scores of PC1 and PC2. Therefore, we conclude that Variable 1 is better than Variable 4 in terms of discriminating between subjects.

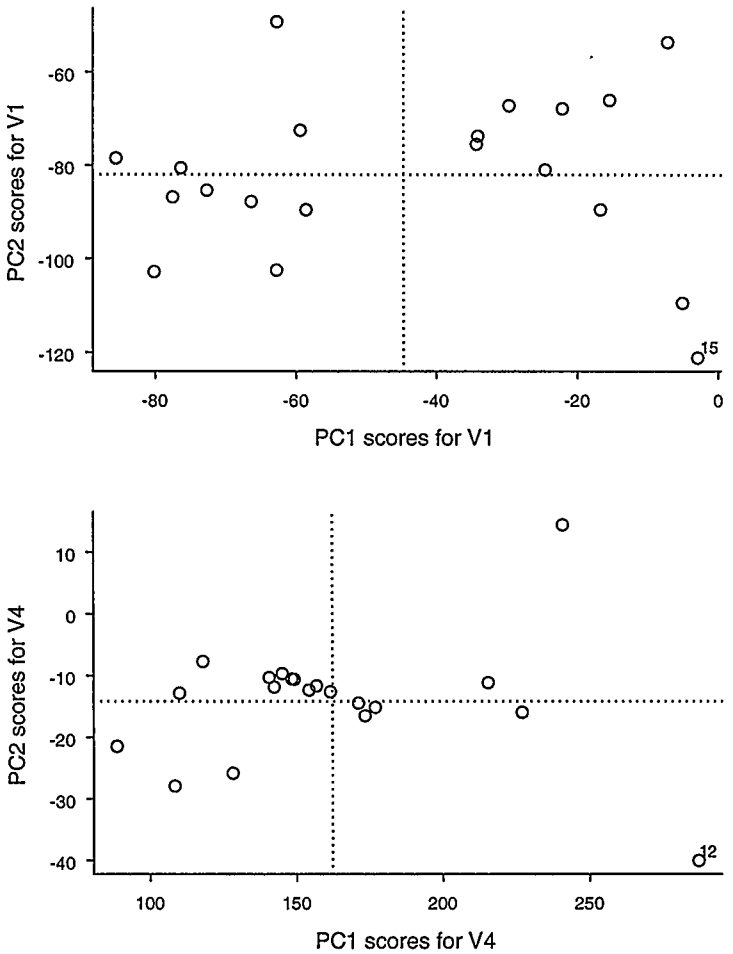

Figure 4.5: The scores on the first two PCs for Variables 1 and 4: PC2 vs PC1. Vertical dotted line: mean of PC1. Horizontal dotted line: mean of PC2.

4.4.2 Condition variation

In the above section, we investigated the variability due to twenty subjects for one condition. In this section, we investigate the more important problem of variability due to conditions for both within subject and across subjects.

Plotting weight functions

In Figure 4.6, we plot the weight functions for the first two principal components, separately for Variable 1 and Variable 4, using Subject 4. For the first PC, the weight functions ξ_1 for both variables have positive effect for most stance phase, and emphasize intervals (0.3, 1) and (0.2, 0.8) for Variable 1 and Variable 4, respectively. The greatest variability between conditions will be found by heavily weighting at about time 0.7 for Variable 1 (63.1% of total variation), and time 0.5 for Variable 4 (98.4% of total variation). For the second PC, ξ_2 for Variable 1 consists of a positive contribution for the interval (0.1, 0.7) and a negative contribution for the two ends, defining a second mode that accounts for 34.7% of the total variation. The second PC for Variable 4 only explains tiny portion (1.2%) of the total variation and is therefore negligible. Nevertheless, one thing draws our attention – its structure. Again, this structure suggests a contrast effect represented by this component, similar to what is seen in the lower left panel of Figure 4.3.

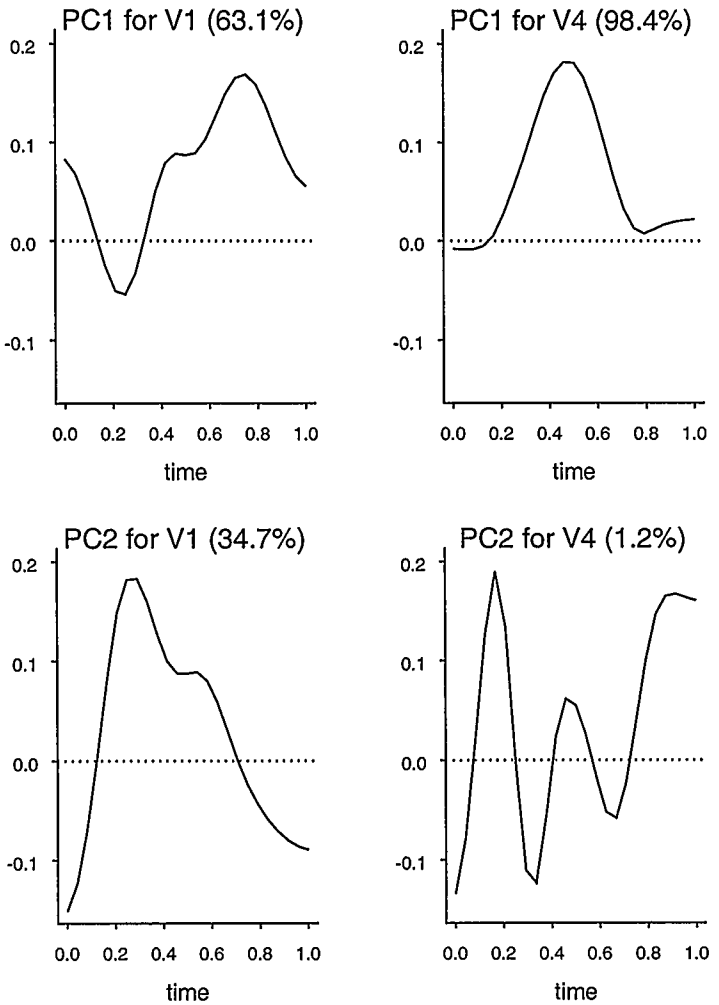

Figure 4.6: Univariate weight functions for condition variation: Subject 4, all 4 conditions. Left: Variable 1. Right: Variable 4.

Figure 4.7 illustrates the same thing as Figure 4.6 does, but for condition variation across subjects. The first PCs for both variables explain most of the variation, with the weight functions emphasizing the interval before time 0.8 for Variable 1 and the central interval for Variable 4. The weight structure of PC2 is more complicated

for both variables. The most obvious feature is that PC2 heavily weights at a short interval close to the end of the stance phase for both variables.

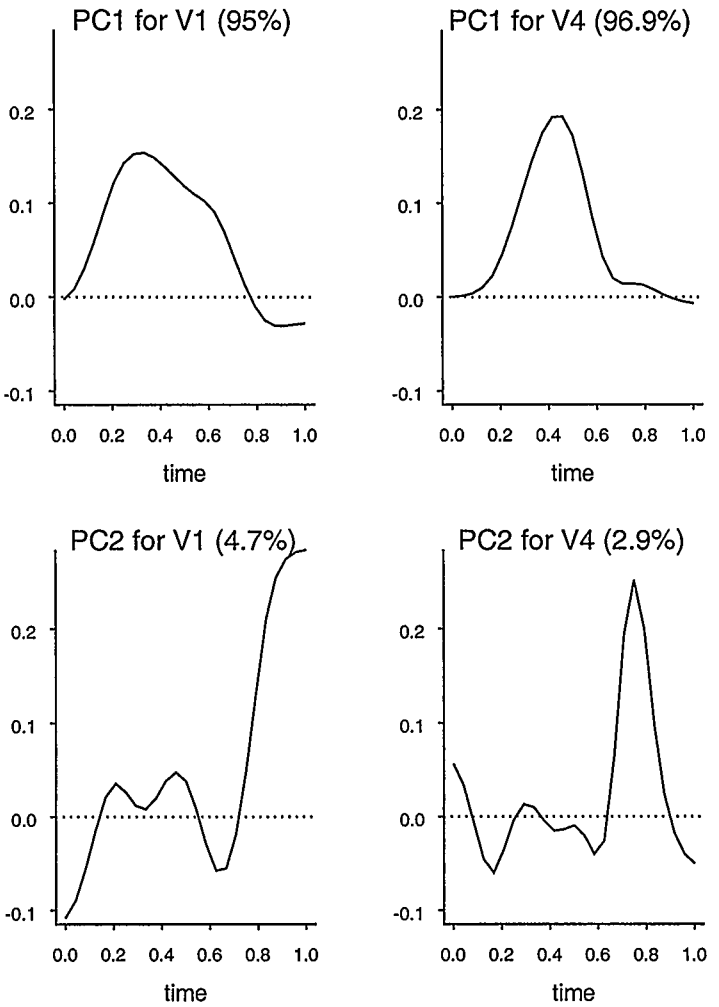

Figure 4.7: Univariate weight functions for condition variation: 4 conditions, each across all subjects. Left: Variable 1. Right: Variable 4.

Finally, we compare condition variability within subject to across subjects. Except PC1 for Variable 4, the shapes of other weight function curves in Figure 4.6 look quite different from the corresponding ones in Figure 4.7. This is also true for plots like Figure 4.6 for other subjects (not shown), but the shapes differ in various ways. This is due to a certain amount of subjective variability. However, for most subjects, the weight function curves of PC1 for Variable 4 are very similar. This indicates that for the most important mode of variation, the four conditions differ in a consistent way for most subjects in Variable 4.

Plotting principal component scores

Figure 4.8 plots PC1 scores against PC2 scores for Subject 4. For Variable 1, PC1 discriminates Condition 4 from the others, while PC2 discriminates Condition 2 from the others. For Variable 4, Condition 1 is distinct.

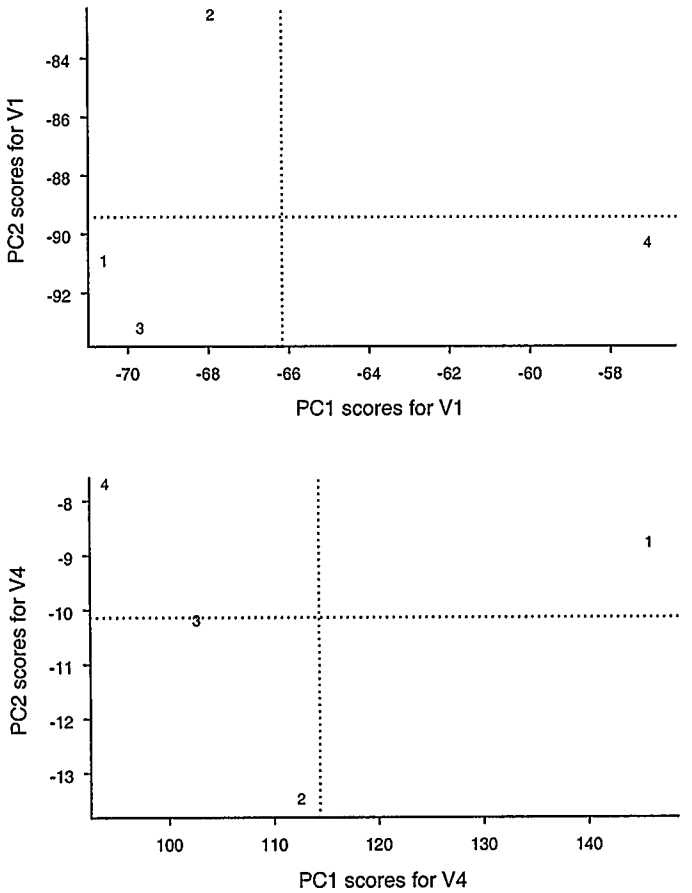

Figure 4.8: The scores on the first two PCs for Variables 1 and 4: PC2 vs PC1. Vertical dotted line: mean of PC1. Horizontal dotted line: mean of PC2.

Figure 4.9 plots the PC scores for all the individual subjects and the ones across subjects. For both variables, most scores are near their means. For the scores across subjects, we notice that the highest PC1 score goes to Condition 2 for Variable 1 and to Condition 1 for Variable 4. Nevertheless, it is hard to distinguish any condition group. In conclusion, neither of these two variables is a good discriminant for identifying conditions when plotting principal component scores.

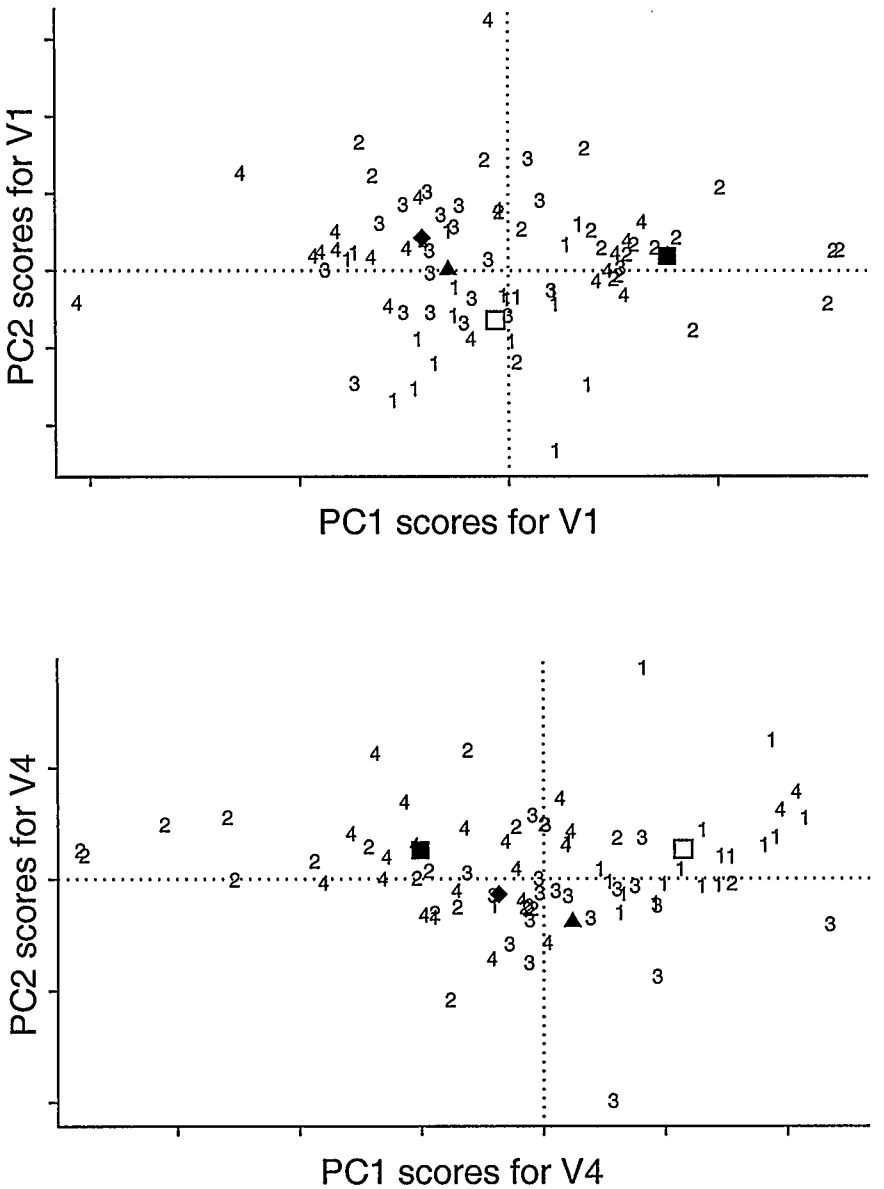

Figure 4.9: The scores on the first two PCs for Variables 1 and 4: PC2 vs PC1. Vertical dotted line: mean of PC1. Horizontal dotted line: mean of PC2. Square: C1 across subjects; bold square: C2 across subjects; bold triangle: C3 across subjects; bold diamond: C4 across subjects.

4.5 Bivariate functional PCA

Investigating the concurrent variation of more than one response function is also of interest. For example, we want to know how two or more variables in the foot orthotics data vary mutually. In the bivariate case, a typical principal component weight function is now defined by $\xi = \left(\xi^A, \xi^B\right)'$, where ξ^A and ξ^B denote the variation in two different variables. To compose two weight functions together, the weighted linear combination in Equation (4.1) becomes

$$f_i = \left\langle \xi^A, x_i^A \right\rangle + \left\langle \xi^B, x_i^B \right\rangle . \tag{4.6}$$

The next task is to elicit the solutions of the eigenequation system (see Equation (4.5)) as we carried out in the univariate analysis. To solve this system, we write

$$\begin{aligned} \int v_{AA}(s,t)\xi^A(t)dt + \int v_{AB}(s,t)\xi^B(t)dt &= \rho\xi^A(s) \\ \int v_{BB}(s,t)\xi^B(t)dt + \int v_{BA}(s,t)\xi^A(t)dt &= \rho\xi^B(s), \end{aligned}$$

where v is defined as the cross-covariance function here. In practice, we first conduct this calculation by discretizing each response function, x_i^A and x_i^B, and then concatenate the resulting vectors to form a single long vector. This is done for each i, and a classical PCA is performed on the resulting system. Finally, the resulting principal component weight vectors are separated into the parts corresponding to x_i^A and x_i^B, representing those two variables. The divided weight vectors are expanded by employing the same basis functions as used in the univariate case.

Visualizing the results on our data

We want to study how two variables behave jointly in both subjective and condition variation cases. Figure 4.10 plots the bivariate (Variables 1 and 4) principal component weight functions for subjective variability. This figure shows that the first PCs in both variables are almost identical to those of the univariate case (see Figure 4.2 and Figure 4.3). PC1 accounts for 60.6% of the total variation, while PC2 accounts for 20.1%.

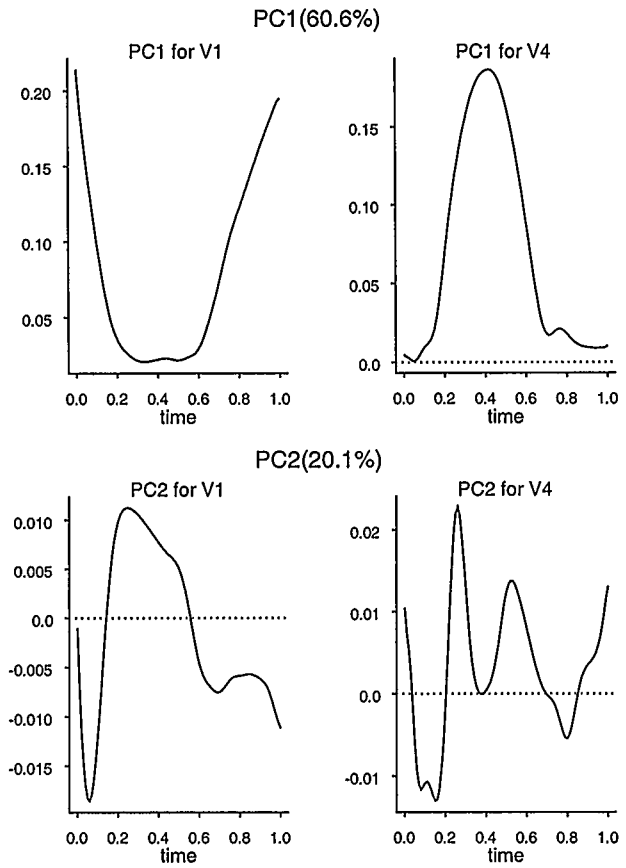

Figure 4.10: Bivariate PC weight functions for subjective variation: Condition 1, all subjects. Left: Variable 1. Right: Variable 4.

Figure 4.11 illustrates the bivariate PCA for condition variation across subjects. Again, the shapes of weight function curves for PC1 in both variables are very similar to those of the univariate analysis (see Figure 4.7). The first two PCs explain 81.4% and 17.7% of the total variation, separately. For both PCs, Variable 1 contributes much more variation than Variable 4 does.

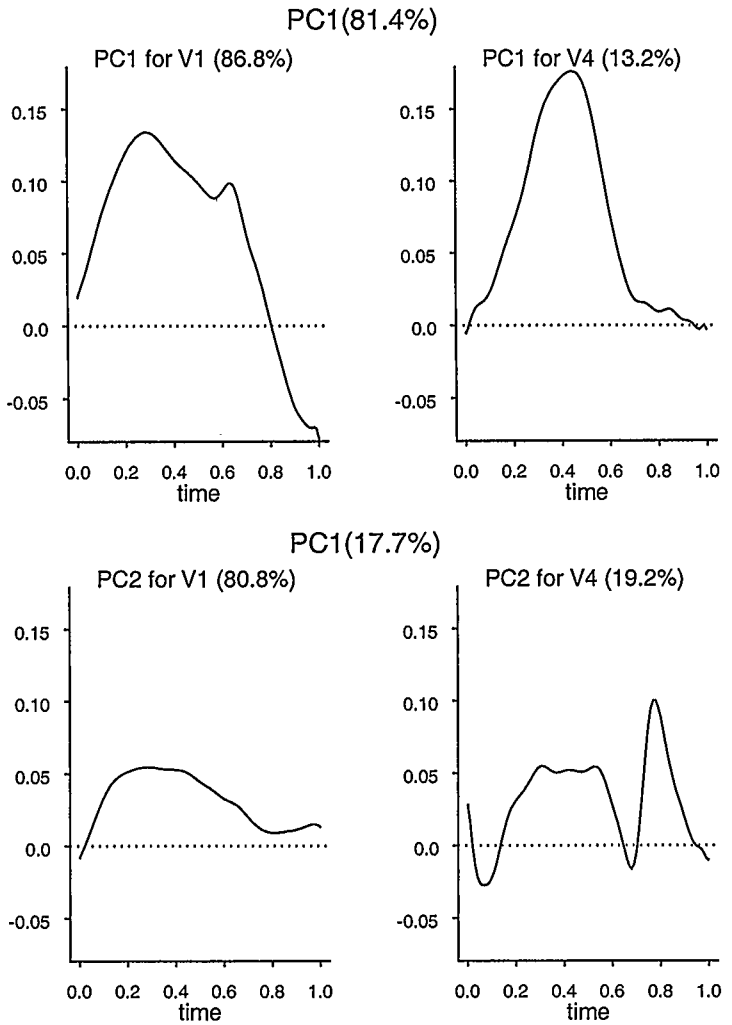

Figure 4.11: Bivariate PC weight functions for condition variation: 4 conditions, each across all subjects. Left: Variable 1. Right: Variable 4.

As for the univariate case, we plot the second bivariate PC scores against the first ones in Figure 4.12. As in Figure 4.9, Condition 2 and Condition 1 differ from the other conditions for Variable 1 and Variable 4, respectively.

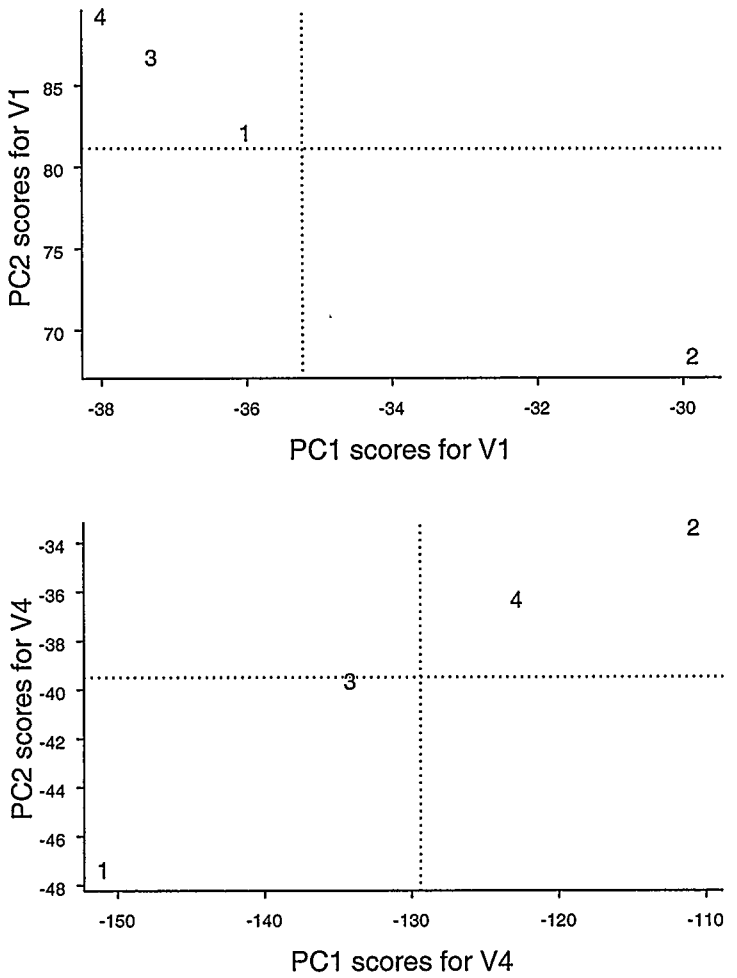

Figure 4.12: The scores on the first two bivariate PCs for condition variation: PC2 vs PC1.

In the bivariate case, the optimal way to exhibit the outcome relies upon the particular context. For cases like the gait data in Ramsay and Silverman [RS2], it is adequate to display individual weight vectors separately. One other approach to demonstrating principal components in the bivariate case is considered especially effective. This approach is to plot the two variables against each other for one PC at a time [RS2]. For observations obtained in an equally spaced interval, the positions of the mean function values $\left(x^A(t), x^B(t)\right)$ are first plotted in the (x, y) plane, indicated by symbols like dots. Then, each dot is connected to the point $\left(x^A(t) + C\xi_m^A(t), x^B(t) + C\xi_m^B(t)\right)$ by an arrow. The constant C is chosen arbitrarily.

Figure 4.13 illustrates such a technique for the case of condition variability across subjects. In the Variable 1 - Variable 4 plane, the mean cycle displays the general configuration of the the gait cycle. Now, one can explain the principal component effect of variation using this illustration. In the upper panel, most arrows are approximately in the x-direction. As we have already seen from Figure 4.11, this means that Variable 1 has more effect of variation than Variable 4. Due to the strong positive effect from Variable 4 at its peak, which is roughly located at the interval (0.35, 0.5), those arrows are pulled to the North-East direction. In the last 20% of the cycle, arrows point to the left on the x-axis because PC1 of Variable 1 has negative effect of variation at that interval. Moreover, the second PC demonstrates a similar situation. Again, Variable 1 shows stronger effect, weighting heavily during the first half of the cycle. Arrows at the interval (0.7, 0.95) point to the y-direction because of stronger mode contributed from Variable 4.

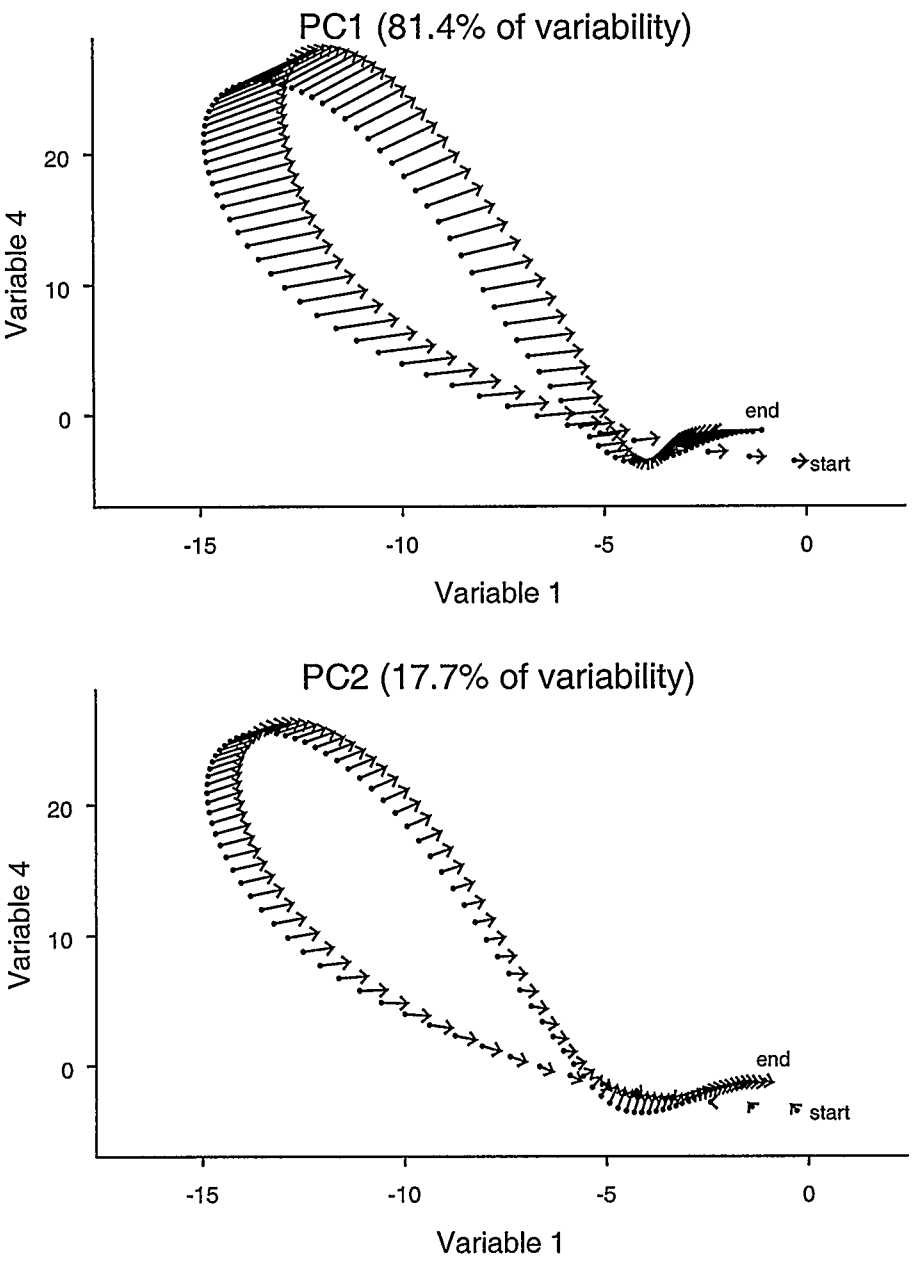

Figure 4.13: Plots of Variable 4 vs Variable 1: mean cycle and the effects of adding a multiple of each of the first two principal component cycles in turn.

The extension to the multivariate case is possible when more than two variables of interest are considered. To do this, one simply concatenates more function vectors to an even longer single vector. Of course, we might encounter even more difficulty in visualizing and interpreting the results than what we faced in the univariate and bivariate cases.

4.6 Extended PCA usage and other approaches

Ramsay and Silverman [RS2, Chapter 8] addressed explicit types of variation in an extraordinary way to make FPCA more discriminating and informative. They used the temperature data as illustration, assessing a small shift of time for each temperature record and investigating its variation. The record-to-record temperature variability then becomes more recognizable. Furthermore, an approach called *principal differential analysis* was introduced in a later chapter. There, they brought up the question of how to incorporate derivative information in examining components of variation.

One of the other approaches to functional PCA, called *regularized principal components analysis*, incorporates techniques of smoothing or regularization into FPCA itself [RS2, Chapter 7]. This method processes data in a different order from that of the approach discussed above. Instead of smoothing the raw data first, it keeps the data unsmoothed before applying ordinary PCA. Smoothing with roughness penalty approach is then performed on the principal component weight function. Ramsay and Silverman claim that the differences between these two methods depend on the degree of smoothing applied to the data and to the principal component functions.

However, the method using smoothed principal components is weakly consistent with the method of estimated eigenvalues and eigenfunctions as used in this thesis [PS].

Chapter 5

Discussion and Conclusions

The purpose of this dissertation is to apply functional data analysis (FDA) methods to empirically determining the effect of foot orthotics on several relevant biomechanical variables. Even with some amount of subjective variability, the data are sufficient to allow us to achieve some conclusive results.

Results from cross-sectional t-tests over all subjects for comparison of Conditions 1 and 2 have shown significant difference. We found that it is particularly useful to measure the proportion of the corresponding p-values below a specified critical value. Further, it was recommended by Araki [Ar] to investigate the locations of parts of the stance phase discriminated by that critical value. Then, pointwise t-tests were considered for within-subject comparison of those two conditions. For most subjects, there is a time interval over which the response differs significantly.

The correlation between certain variables is also of study interest. A pointwise correlation function between two variables for each subject was adopted for the analysis. No obvious general pattern has been shown; instead, it reflects quite an amount of subject-to-subject variability. This is also true for some other variable pairs. The integrated coefficient of determination, ICOD, was proposed to compare a bivariate response for two conditions. No evidence of difference was speculated. Although the method used for correlation analysis here employs the whole curve information, which is better than using only a single value such as the global maximum, it might be better to divide the whole interval to several sections depending upon regions

showing strong local features. Moreover, due to not being able to capture the shape the of correlation function, ICOD is still not an optimal statistic for this situation. In addition, one might explore a better measure for multivariate functional data, instead of correlation for bivariate data.

We applied principal components analysis (PCA) to explore both subject variability and condition variability, after transforming our data to functional form and registering them to a better display. For both univariate and bivariate analysis, weight functions $\xi(s)$ were derived to define the most important mode of variation in the curves; in addition, the scores f_i were examined and plotted to identify distinct conditions or groups of subjects. Extending bivariate FPCA to the multivariate case is as simple as bivariate FPCA itself, but data interpretation will be more challenging.

This thesis provides evidence that there is difference between some of the tested foot orthotics (Conditions 1 and 2) for the variable (Variable 1) analyzed. However, there are insignificant findings in the correlation analysis. This suggests that stronger instruments are required for this analysis. Also, the certain amount of subjective variability is one thing that causes undesired results. Here are several avenues suggested for future research. First, one should take account of greater within-day repeatability; that is to analyze the data within session instead of across sessions. This way allows us to keep more statistical power. Second, due to the time issue regarding the registration process, we did not use the whole set of data. Rather, only a sample of 25 was used per subject. Again, it is a suboptimal method that forces the loss of power.

The intention of this dissertation is to provide a general framework of how we can

apply FDA on such types of data. The existing statistical technology is not quite sufficient for FDA; therefore, the field of FDA offers many challenges and research opportunities. We hope that this study will motivate additional research in this area and add some new knowledge to this field.

Bibliography

[Ar] Araki, Y. (2002) *Functional Data Analysis of Human Gait.* M.Sc. thesis, Department of Mathematics and Statistics, University of Calgary.

[CM] Chaudhury, P. and Marron, S. (1999) Size for exploration of structures in curves. *Journal of the American Statistical Association* 94:807-823.

[DL] Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1994) *Analysis of Longitudinal Data.* Clarendon Press, Oxford.

[EM] Eilers, P.H.C. and Marx, B.D. (1996) Flexible smoothing with B-splines and penalties. *Statistical Science* 11(2):89-121.

[Fa] Faraway, J.J. (1997) Regression analysis for a functional response. *Technometrics* 39:254-261.

[FL] Fan, J. and Lin, S. (1998) Test of significance when data are curves. *Journal of the American Statistical Association* 93:1007-1021.

[FS] Friedman, J. and Silverman, B.W. (1989) Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics* 31:3-39.

[GS] Green, P.J. and Silverman, B.W. (1994) *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall, London.

[HC] Hand, D. and Crowder, M. (1996) *Practical Longitudinal Data Analysis.* Chapman and Hall, London.

[JW] Johnson, R.A. and Wichern, D.A. (1998) *Applied Multivariate Statistical Analysis.* Prentic-Hall, New Jersey, 4th edition.

[KG] Kneip, A. and Gasser, T. (1992) Statistical tools to analyze data representing a sample of curves. *Annals of Statistics* 20:1266-1305.

[Mc] McClay I. (2000) The evolution of the study of the mechanics of running. Relationship to injury. *J. Am. Podiatr. Med. Assoc.* 90:133-148.

[Mu] Muendermann, A. (2002) *Posting versus Custom-Molding of Foot Orthotics: Effects on Biomechanical Variables and Comfort.* Ph.D. thesis, Department of Medical Science, University of Calgary.

[NC] Nigg, B.M., Cole, G., Stergiou, P. and Stefanyshyn, D. (2000) The use of pressure measurements to determine the effect of shoe orthotics on knee joint moments. In *Abstracts of the Emed Millennium Meeting, Munich, Germany,* page 34.

[NK] Novick, A. and Kelley, D.L. (1990) Position and movement changes of the foot with orthotic intervention during loading response of gait. *J. Sports Phys. Ther.* 11(7):301-311.

[PS] Pezzulli, S. and Silverman, B.W. (1993) Some properties of smoothed principal components analysis for the functional data. *Computational Statistics* 8:1-16.

[Ra1] Rao, C.R. (1958) Some statistical methods for comparison of growth curves. *Biometrics* 14:1-17.

[Ra2] Rao, C.R. (1987) Prediction in growth curve models (with discussion). *Statistical Science* 2:434-471.

[RB] Ramsay, J.O. and Bock, R.D. (2002) Functional data analysis for human growth. Unpublished manuscript, McGill University.

[RD] Ramsay, J.O. and Dalzell, C.J. (1991) Some tools for functional data analysis. *Journal of the Royal Statistical Society B* 53:539-572.

[RL] Ramsay, J.O. and Li, X. (1998) Curve registration. *Journal of the Royal Statistical Society B* 60:351-363.

[RS1] Ramsey, F.L. and Schafer, D.W. (1996) *The Statistical Sleuth: A Course in Methods of Data Analysis.* Duxbury Press, Belmont, Calif.

[RS2] Ramsay, J.O. and Silverman, B.W. (1997) *Functional Data Analysis.* Springer-Verlag, New York.

[RS3] Ramsay, J.O. and Silverman, B.W. (2002) *Applied Functional Data Analysis: Methods and Case Studies.* Springer-Verlag, New York.

[RS4] Rice, J.A. and Silverman, B.W. (1991) Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society B* 53:233-243.

[Sc] Schmid, C.H. (1996) An EM algorithm fitting first-order autoregressive models to longitudinal data. *Journal of the American Statistical Association* 91:1322-1330.

[SC] Smith, L.S., Clarke, T.E., Hamill, C.J. and Santopietro, F. (1986) The effects of soft and semi-rigid orthoses upon foot eversion in running. *J. Am. Podiatr. Med. Assoc.* 76:227-233.

[Si1] Silverman, B.W. (1995) Incorporating parametric effects into functional principal components analysis. *Journal of the Royal Statistical Society B* 57:673-689.

[Si2] Simonoff, J.S. (1996) *Smoothing Methods in Statistics.* Springer-Verlag, New York.

[Tu] Tucker, L.R. (1958) Determination of parameters of a functional relationship by factor analysis. *Psychometrika* 23:19-23.