INCORPORATING PARAMETRIC EFFECTS INTO
FUNCTIONAL PRINCIPAL COMPONENTS ANALYSIS

BY

B. W. SILVERMAN

TECHNICAL REPORT NO. 448
MARCH 1994

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT DMS92-04864

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Incorporating Parametric Effects Into Functional Principal Components Analysis

By

B. W. Silverman

Technical Report No. 448

March 1994

Prepared Under the Auspices Of

National Science Foundation Grant DMS92-04864

Department of Statistics
Stanford University
Stanford, California
Incorporating parametric effects into functional principal components analysis

B. W. Silverman*
School of Mathematics, University of Bristol
University Walk, Bristol BS8 1TW, UK
24 March 1994

Abstract

The ideas of functional principal component analysis are extended to deal with data that are hybrids of 'functional' and 'parametric' effects. The parametric effects may be far more general than just the addition of a multiple of a given function, and may, for example, include shifts of the time axis. Given data, a Procrustes fitting method can be used to estimate the parametric effects. A number of possible ways of treating the estimated parameter values are discussed. The methods are illustrated by reference to temperature data at 35 Canadian weather stations.

1 Introduction

A number of recent papers (for example Ramsay and Dalzell, 1991, and Rice and Silverman, 1991, and works cited in those two papers) have considered versions of principal components analysis (PCA) for data that may be considered as curves rather than the vectors of classical multivariate analysis. Suppose that we have observations $X_1(t), \ldots, X_n(t)$ assumed to be drawn from a stochastic process $X$ on an interval $[0, T]$. In some (but by no means all) applications, the interval will be considered to be periodic.

*Research partially supported by NSF grant DMS 9209130 at Stanford University
1.1 The Rice–Silverman approach

The approach of Rice and Silverman (1991) performs a ‘functional’ principal components expansion on the given data. Smoothed principal components are extracted successively, based on a roughness penalty such as $\int u''^2$. The $\nu$th principal component $\hat{u}_\nu$ is estimated by maximising the sample variance of $\int u(t)X(t)dt$ minus $\alpha_\nu \int u''^2$ subject to the constraints that $\int u^2 = 1$ and the orthogonality constraints $\int \hat{u}_j u_\nu = 0$ for $j < \nu$. (Throughout the paper, integrals will be taken to be over the range $[0,T]$ on which the data are defined.) It has been found in practice to be desirable to choose smoothing parameters $\{\alpha_\nu\}$ that decrease as $\nu$ increases. An alternative approach to smoothed PCA is described by Silverman (1994).

The overall effect of this approach is to model a typical observation as an expansion

$$X(t) = \mu(t) + \sum \xi_\nu u_\nu(t)$$

(1)

where the principal component scores $\xi_\nu$ are uncorrelated variables with mean zero and variances $\sigma_\nu^2$, say. Just as in classical multivariate analysis, the principal components $u_\nu(t)$ often have interesting physical meanings. Rice and Silverman (1991) discussed the PCA of data obtained from the study of human gait. In this paper we shall concentrate on a set of data discussed by Ramsay and Dalzell (1991) consisting of the average monthly temperature for each of 35 Canadian weather stations (Canadian Climate Program, 1982).

The first four smoothed principal components for this data set (estimated by the method of Silverman, 1994) are shown in Figure 1. Interpreting the principal components is not always entirely straightforward, and one method that has been found to be helpful is to examine plots of the overall mean function and the functions obtained by adding and subtracting a suitable multiple of the principal component in question. Figure 2 shows that the first PC corresponds approximately to an overall increase in temperature; the second to an increase in the difference between summer and winter temperature; the third in some way to a time shift effect; and the fourth to an effect whereby the onset of spring is earlier and the autumn ends later. For our present purposes the third PC is of particular interest, because it does not really correspond to an effect naturally described in terms of adding a principal component function pointwise to a mean function, but rather to a transformation of the time axis.
Figure 1: First four smoothed principal components of Canadian temperature data

Figure 2: Mean temperature curve and effect of adding and subtracting a suitable multiple of each principal component curve
1.2 The Ramsay–Dalzell approach

An important feature of the approach of Ramsay and Dalzell (1991) is the partitioning of the variability in $X$ into a 'parametric part' and a 'nonparametric part' by application of a suitable differential operator $L$. (Ramsay and Dalzell use the terms 'structural component' and 'residual component'.) The parametric part consists of components that are annihilated by $L$. Suppose, for example, that the observations are periodic on the interval $[0, 2\pi]$ and the differential operator $L$ is defined by $Lx = x'' + x$. Then the partitioning of the variability would be described by associating with each observation $X$ a vector of three values

$$X^{(1)} = \frac{1}{2\pi} \int X(t) dt, \quad X^{(2)} = \frac{1}{\pi} \int X(t) \cos t \, dt, \quad X^{(3)} = \frac{1}{\pi} \int X(t) \sin t \, dt.$$

The parametric part of the observation $X$ can then be thought of as being the curve

$$X^{\text{par}} = X^{(1)} + X^{(2)} \cos t + X^{(3)} \sin t$$

which will satisfy

$$LX^{\text{par}} = 0.$$

The nonparametric component is defined to be

$$X^{\text{nonpar}} = X - X^{\text{par}};$$

it is orthogonal to $X^{\text{par}}$ and satisfies $LX^{\text{nonpar}} = LX$. In Ramsay and Dalzell's approach, the function $LX$ is thought of as being the nonparametric part and the function $X^{\text{nonpar}}$ itself does not play a dominant role.

A key feature of the approach is that the functions $1, \sqrt{2} \sin t$ and $\sqrt{2} \cos t$ that are accounted for in the parametric part are implicitly defined by the choice of the differential operator $L$. Recent work of Dalzell and Ramsay (1993) explains how to specify a differential operator $L$ corresponding to any choice of functions spanning the 'parametric space', but the restriction to a parametric part consisting of a linear combination of functions added pointwise to one another remains.

Ramsay and Dalzell (1991) go on to examine the nonparametric part in terms of a principal components analysis of the observations $LX_i$; they use a somewhat different, but equivalent, terminology. Overall, their analysis
would allow a typical observation to be expressed as the sum of a mean function, a function drawn from the ‘parametric’ space, and a sum of multiples of principal component functions all lying in the ‘nonparametric’ space. The way in which their PCA is performed, on the functions \( L X_i \) rather than on the functions \( X_i^{nonpar} \) makes for a different expansion with somewhat different interpretability properties, but the broad overall idea of separating the variability remains.

2 More general transformations

The basic idea of this paper is to allow a more general model than a mean function with perturbations added to it. We shall consider a very general formulation in Section 6 below, but for the moment we concentrate on the simple case of shifts in time. For simplicity, we shall assume for the moment that functions of interest may be considered as periodic on the interval \([0, T]\).

Let us suppose that an observation can be modelled as

\[
X(t + \tau) = \mu(t) + \sum \xi_u u_u(t).
\]  

(2)

The model (2) of course only differs from (1) in that it allows for a shift in time \( \tau \) as well as for the addition of multiples of the principal component functions. Because of the periodicity, the shifted function \( X(t + \tau) \) may still be considered as being a function on \([0, T]\).

How should the parameter \( \tau \) be considered? It is helpful to distinguish three different possibilities:

**Model 1**: The \( \tau \) is merely a ‘nuisance parameter’ that is of no real value in the analysis. For example, the curve \( X \) may be obtained by switching on a recording instrument at an arbitrary time. Two \( X \) curves that differ only by a shift in time (i.e., by a difference in \( \tau \)) would under this model be considered as being essentially the same.

**Model 2**: The \( \tau \) are parameters associated with particular observations, but they are not considered as being part of the ‘random model’. They are nevertheless of genuine interest.
Model 3: The $\tau$ are values that are of genuine interest, and furthermore they can be considered as random quantities that can be decomposed together with the functional part of the principal components analysis, for example by being modelled as

$$\tau = \sum \xi_\nu \tau_\nu. \quad (3)$$

Note that the values $\xi_\nu$ in (3) are the same as those in the expansion in (2). Thus the $\nu^{th}$ principal component of the variation can be considered as having two effects: the addition of the function $u_\nu(t)$ together with a contribution of $\tau_\nu$ to the time shift.

There are of course questions of identifiability in the models as we have formulated them. Apart from the obvious point that a shift of $T$ is equivalent to no shift at all, it is also clearly the case that a pure shift of a mean function $\mu$ could equally well be represented by an additive effect. We shall elaborate on this question of identifiability below. Model 2 is of course a ‘fixed effects’ model, while Model 3 is a ‘random effects’ model.

3 A Procrustes fitting approach

Suppose we have a set of data $X_i(t)$ such as the Canadian weather data. Let $\tau_i$ be the shift associated with $X_i$. Whichever model is being used, the first step in our analysis is to estimate the shifts $\tau_i$.

In the case of the temperature data, we perform this step by a Procrustes fitting approach (see, for example, Mardia, Kent and Bibby, 1979). Set each $\hat{\tau}_i$ to zero initially, and then cycle between the following two steps to convergence:

1. Define $\hat{\mu}(s)$ to be the sample mean of the shifted functions $X_i(s + \hat{\tau}_i)$. In some cases it may be appropriate to smooth the sample mean in the manner described in Rice and Silverman (1991) but the temperature data are sufficiently smooth as to make this unnecessary.

2. Update the $\hat{\tau}_i$ by a single Newton–Raphson step of the minimization of the sum of squares

$$R(\tau) = \sum_i \int \{X_i(s + \tau_i) - \hat{\mu}(s)\}^2 ds, \quad (4)$$
keeping \( \hat{\mu} \) fixed.

In order to carry out the second step, since the \( X_i \) need not be as smooth as the estimated mean \( \hat{\mu} \), we rewrite the residual sum of squares (4) as

\[
R(\tau) = \sum_i \int \{X_i(s) - \hat{\mu}(s - \tau_i)\}^2 ds.
\]  

(5)

Using the fact that \( 2 \int \hat{\mu}(u)\hat{\mu}'(u)du = \hat{\mu}(T)^2 - \hat{\mu}(0)^2 = 0 \), because of the periodicity assumptions, it follows that

\[
\frac{\partial R}{\partial \tau_i} = \int X_i(s)\hat{\mu}'(s - \tau_i)ds
\]

and

\[
\frac{\partial^2 R}{\partial \tau_i^2} = -\int X_i(s)\hat{\mu}''(s - \tau_i)ds,
\]

and hence that step 2 above increments \( \hat{\tau}_i \) by

\[
\int X_i(s + \hat{\tau}_i)\hat{\mu}'(s)ds = \int \{X_i(s + \hat{\tau}_i) - \hat{\mu}(s)\}\hat{\mu}'(s)ds = 0.
\]

(6)

In practice the convergence of this algorithm is rapid. In the early stages of the algorithm it may be preferable to replace the denominator of the increment to \( \tau \) by its expected value \( \int \hat{\mu}\hat{\mu}'' = -\int \hat{\mu}^2 \). One identifiability property is that a constant may be added to all the \( \tau_i \) without essentially changing the model; with this in mind the \( \hat{\tau}_i \) are adjusted to have sample mean equal to zero. Notice also that the exact method of estimating the \( \tau_i \) may well depend on the context. For instance, the Canadian weather data also include information on precipitation. However, the variability in the precipitation measurements is such that it is clearly best to estimate the time shift by reference to the temperature measurements only.

The algorithm as set out above is easily implemented by a Fourier series approach. The Canadian temperature data, for example, are observed at 12 points that we regard as equally spaced on the unit circle. A real Fourier expansion to 12 terms then gives a smooth interpolant of the data; we shall regard these interpolants as being the observed functional data \( X_i \). It is straightforward to find the Fourier series expansion of any shifted function \( Y_i \). (This will in general be of length 13, because the coefficient of \( \sin 12\pi t \)
Figure 3: Estimated time shifts for the Canadian weather stations, measured in days

will no longer necessarily be zero.) The differentiation and inner product operations in the algorithm are simply expressed in Fourier series terms. If any smoothing is required in the calculation of  is, it is again very natural to do this by tapering the mean of the Fourier coefficients of the Yi.

The estimated for the Canadian temperature data are shown in Figure 3. The way in which temperature patterns in different places show systematic differences in time (of over three weeks between the extreme cases) is of clear meteorological interest, and can be explained by a number of factors, such as the proximity to large areas of water and the effects of ocean currents.

4 Accounting for the estimated shifts

4.1 Separating the shifts from the PCA

Of course, if the shifts were of no interest then it would suffice to perform a PCA (possibly smoothed) on the shifted data .

In the case of Model 2 above, one could similarly perform a PCA
on the shifted data, but possibly relate the principal component scores to the estimated shifts. In the Ramsay-Dalzell paradigm this could be considered as allowing the shifts to constitute the 'parametric part' of the model. One extension of the original Ramsay-Dalzell methodology is that the specification of 'time shift' as a parametric variable is explicit, and is not connected to a particular differential operator. Furthermore parametric variables need no longer correspond to multiples of functions that are added pointwise to the mean function, but can be specified by other operations such as time shifts.

Once the PCA has been carried out, it can be instructive to calculate the correlations of the various principal component scores with the parametric components. An analysis of this kind will be carried out in Section 5.2 below.

4.2 Including the shifts in the PCA

In this subsection we concentrate on Model 3 and indicate how the shifts may be included directly in the principal components analysis. We regard the data as being pairs \((Y_i, \tau_i)\), where the \(Y_i\) are the shifted temperature curves \(X_i(\cdot + \hat{\tau}_i)\) and the \(\tau_i\) are the estimated values of the shift parameter. Note that because of property (6) above, the functions \(Y_i - \hat{\mu}\) are all orthogonal to the function \(\hat{\mu}'\). There is some connection with the work of Rice and Silverman (1991), who considered the analysis of bivariate curve data of the form \((X_i(t), Y_i(t))\), though the data we consider here consist of pairs where only one component of the bivariate vector is a function.

How should the PCA of these hybrid data be performed? In order to answer this question, it is helpful to digress briefly. Suppose that \(\eta\) is any function orthogonal to \(\hat{\mu}'\) and that \(\|\eta\|\) is small. Define \(z\) to be the function obtained by adding \(\eta\) to \(\hat{\mu}\) and shifting by \(\tau\), so that

\[
z(s) = \hat{\mu}(s - \tau) + \eta(s - \tau),
\]

so that, to first order in \(\eta\) and \(\tau\),

\[
z(s) - \hat{\mu}(s) \approx -\tau \hat{\mu}'(s - \tau) + \eta(s - \tau).
\]

Thus, by the orthogonality of \(\eta\) and \(\hat{\mu}'\) we have

\[
\int \{z(s) - \hat{\mu}(s)\}^2 ds \approx \int \eta^2 + \hat{\mu}'^2 \tau^2
\]  

(7)
where $C^2 = \int \hat{\mu}^2$.

We can conclude that if we regard the norm of the pair $(\eta, \tau)$ as being $\int \eta^2 + C^2 \tau^2$ then the norm will agree, to first order, with the usual square integral norm of the effect of applying the pair to $\hat{\mu}$. With this calculation in mind, we perform our PCA of the pairs $(Y_i, \tau_i)$ relative to this norm, which we shall call the canonical norm of the hybrid data PCA.

The easiest way of achieving this in practice is by a Fourier transform approach. Assume that the $Y_i$ have, or can be considered to have, exact Fourier expansions of length $N$, with $N$-vectors of coefficients $Y_i\hat{\eta}^*$. We then augment each of these vectors of Fourier coefficients by the coefficient $C\tau_i$. The sum of squares of the entries in each of the augmented vectors is then equal to $\int Y_i^2 + C^2 \tau_i^2$. We perform a standard PCA of the augmented data vectors, for example using the Splus routine prcomp. Each principal component vector that is extracted can then be re-expressed as a (function, shift) pair, by separating off the 'augmented' part of the vector and then applying an inverse Fourier transform to the remainder. The percentage of variability in each component due to 'shift' and to 'function addition' can be found by summing the squares of the components in the relevant parts of the principal component vectors. If it appropriate to use a smoothed version of PCA, the approach of Silverman (1994) can be applied immediately to these data, with the roughness penalty depending only on the Fourier coefficient part of the vectors.

Because of the way that the norm of the hybrid data is calculated, the PCA of the $(Y_i, \hat{\eta}_i)$ will correspond very closely to a standard functional PCA of the original data $X_i$. An important difference will be that the 'shift' part of each principal component will be made explicit without any need for interpretative skill. Furthermore any smoothing that is performed will not affect the 'shift' part of the data or the principal component vectors, because this part is explicitly excluded from the calculation of the roughness penalty. Note that the constraint that the canonical norm of a principal component is unity implies that the maximum possible absolute value of the variable $\tau$ is $C^{-1}$, and the quantity $C^2 \tau^2$ gives the proportion of the variability in the principal component accounted for by the 'parametric' part.
Figure 4: The effect of treating the shift as a separate quantity, included in the PCA. Each plot gives the effect on the mean function of the ‘function’ part of the PCA, and the shift part is given explicitly. A component that was entirely ‘shift’ would have a time shift value of ±5.3 days.

5 Application to the temperature data

5.1 Model 3 analysis

Figure 4 gives the result of applying the methodology we have described to the Canadian temperature data. Each plot gives the result of adding and subtracting a multiple of the ‘function’ part of the hybrid principal component, and the ‘shift’ component is given explicitly. The multiple used is the same in all cases, and the reason that principal component 3 appears to have a smaller effect is that the norm of the function part is much less than in the other components; most of the variability is taken up by the shift part. If the shift was as large as 5.3 days then 100% of the principal component would be accounted for by the shift and the function part would be identically zero.

Examination of this plot indicates that it might be of interest to consider an overall change in the temperature as being another ‘parametric’ effect, in
other words to model individual data curves as

$$X(t + \tau) = \mu(t) + \theta + \sum_\nu \xi_\nu u_\nu(t),$$

where $\theta$ is an overall temperature effect that may be considered in any of the three ways set out in Section 2 above. In order to follow the procedure set out for the case where the shift was the sole parametric variable, we estimate $\hat{\tau}$ and $\hat{\theta}$ to minimize the residual sum of squares

$$R(\tau, \theta) = \sum_i \int \{X_i(s) - \theta_i - \hat{\mu}(s - \tau_i)\}^2 ds.$$  \hspace{1cm} (9)

Of course, this does not require any additional Procrustes fitting beyond that already carried out above. Impose the identifiability constraint $\sum \theta_i = 0$; the minimum is then attained by finding $\hat{\tau}$ and $\hat{\mu}$ as above, and then setting $\hat{\theta}_i$ to be $T^{-1} \int \{X_i(s) - \hat{\mu}(s - \tau_i)\} ds$. An extension of the argument given in Section 4.2 above shows that it is appropriate to regard the canonical norm of the triple $(\eta, \tau, \theta)$ as being $\int \eta^2 + C^2 + T^2 \theta^2$. Some details will be given in Section 6 below.

The effect of treating both the time shift and the overall mean temperature as parametric variables is illustrated in Figure 5. This presentation makes it very clear that the first PC is almost entirely due to the mean temperature, and the third PC almost entirely to the time shift.

A comparison between Figures 2 and 5 is instructive. The estimated variance of each of the principal component scores (given in the last two columns of Table 1) is virtually identical. It is interesting that the greatest discrepancy is in the third principal component, the one in which the non-linear 'time shift' effect is greatest. In Table 1 we also give the percentage of variability in each principal component due to the functional part and to the various parametric variables, as judged by the contribution that each makes to the squared norm of the corresponding principal component weight vector. It can be seen at once from the table that the first principal component is mainly 'overall temperature' and that the third is mainly 'shift'. It is worth noting that it is not true that the vast majority of the shift is accounted for by the third PC, because one needs to weight by the variances in order to apportion the shift between the various principal components. If we do this, we find that the first four principal components respectively account for 0.3%, 38.5%, 60.0% and 1.2% of the variability in the $\hat{\tau}_i$. For the overall
Figure 5: Effect of treating both shift and overall temperature as parametric variables. A component that was entirely 'time shift' would correspond to a time shift of ±5.3 days, while a component that was entirely a change in overall mean temperature would correspond to a mean temperature effect of ±1 degree.
Table 1: Statistics of the principal component analysis taking account of the shift and overall temperature. The final column gives the variances of the corresponding PCA of the original curve data, without removing any parametric effects.

<table>
<thead>
<tr>
<th>PC</th>
<th>Percentage of squared norm due to function</th>
<th>Overall temp</th>
<th>Shift</th>
<th>Variance in this analysis</th>
<th>Variance in standard analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14.5</td>
<td>85.5</td>
<td>0.0</td>
<td>42.06 (89.7%)</td>
<td>42.01</td>
</tr>
<tr>
<td>2</td>
<td>76.9</td>
<td>12.1</td>
<td>11.1</td>
<td>3.80 (8.1%)</td>
<td>3.84</td>
</tr>
<tr>
<td>3</td>
<td>17.0</td>
<td>2.0</td>
<td>81.0</td>
<td>0.81 (1.72%)</td>
<td>0.69</td>
</tr>
<tr>
<td>4</td>
<td>91.9</td>
<td>0.9</td>
<td>7.1</td>
<td>0.18 (0.38%)</td>
<td>0.18</td>
</tr>
</tbody>
</table>

temperature effects \( \hat{\theta} \); the first principal component accounts for 98.7% of the variation and the second for almost all of the remainder.

5.2 Model 2 analysis

Suppose now that we proceed more in the manner of Ramsay and Dalzell (1991) and remove the parametric variables from the principal components analysis altogether. Of course, this reduces the sum of squares considerably; shifting the curves \( X_i \) by \( \hat{r}_i \) and subtracting the constant \( \hat{\theta} \), from each one reduces the the residual sum of squares \( \sum f(X_i - \hat{\mu})^2 \) to 20.9% of its original value.

A smoothed PCA of the transformed data curves was carried out. The first two principal components have estimated variances 8.94 and 0.44, and account for 99% of the variability in the transformed data. Hence the percentage of variability in the original data remaining unexplained by these two components and the parametric variables is 0.2%. A plot of the principal components is given in Figure 6. It can be seen that a large score on the first component corresponds to an increased range of temperatures between summer and winter, and on the second to early onset of spring and late onset of autumn—corresponding to the interpretations of the second and fourth PCs, respectively, of the previous analyses.

In Table 2 we give the sample correlations between the matrix between the various effects. The scores on the two principal components are called 'range' and 'long summer'. Several effects are of interest. There is very little correlation between the time shift and the average temperature. Average
Figure 6: The effects on the overall mean temperature of adding and subtracting multiples of the first two principal components of the data corrected for overall mean and time shift

Table 2: Sample correlations between the two parametric variables and the first two principal components of the transformed data

<table>
<thead>
<tr>
<th></th>
<th>Average</th>
<th>Shift</th>
<th>‘range’</th>
<th>‘long summer’</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>1</td>
<td>-0.003</td>
<td>-0.756</td>
<td>0.439</td>
</tr>
<tr>
<td>Shift</td>
<td>-0.003</td>
<td>1</td>
<td>-0.311</td>
<td>-0.547</td>
</tr>
<tr>
<td>‘range’</td>
<td>-0.756</td>
<td>-0.311</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>‘long summer’</td>
<td>0.439</td>
<td>-0.547</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
temperature is quite strongly negatively correlated with range, and is positively correlated with summer length. There are negative correlations between the time shift and both the principal component effects.

Clearly there is the possibility of interplay between the two approaches we have described. For example, the original analysis provides a suggestion of which quantities might be important parametric variables, and the model 3 analysis allows a clearer interpretation of the original analysis. Finally the model 2 analysis gives a way of separating the parametric effects altogether from effects displayed in the functional part of the analysis. Clearly it is now possible to define new parametric variables to describe some of these effects. For example, it is probably more natural to describe the early onset of spring and the late onset of autumn by a deformation of the time axis rather than an additive effect. In the next section we set out a general framework that allows parameters of this kind to be considered.

6 A general framework

In order to set the above discussion in a very general framework, let us posit a parametric model where the parametric part of the model for $X$ consists of a transformation $G_{\theta}(\mu)$ where $\theta$ is a vector of real parameters. We assume that the transformation $G_{\theta}$ has any necessary smoothness properties, and that the mapping $\theta \mapsto G_{\theta}$ is itself smooth. We shall also assume that if $\theta = 0$ then $G_{\theta}$ is the identity.

Our model for a random function $X$ will then be

$$G_{\theta}^{-1}X = \mu + \sum \xi_{\nu}u_{\nu},$$

(10)

where $\mu$ is a mean function and $u_{\nu}$ are principal component functions. This is a generalization of (2) and (8) above, which both arise for suitable choices of the transformation families $\{G_{\theta}\}$.

Given a data set of observed curves $X_i$, we first obtain estimates $\hat{\theta}_i$ of the parameters corresponding to each observation. The usual way of proceeding will be by a Procrustes (or explicit) minimization of the residual sum of squares $R(\theta_1, \ldots, \theta_n) = \sum_i \|X_i - G_{\theta_i}\hat{\mu}\|^2$, where $\| \cdot \|$ is the $L^2$ norm, and $\hat{\mu}$ is an estimate, possibly smoothed, of the mean of the functions $G_{\theta_i}^{-1}X_i$.

In order to ensure identifiability of the parameters $\theta$ and $\mu$ we constrain
the \( \theta_i \) to have mean zero. The precise way in which it is appropriate to enforce this constraint will depend on the context, but if a Procrustes fitting procedure of the form set out above is used, it may be possible to build it into the Newton–Raphson step, or alternatively the estimated parameters can be adjusted by subtracting their sample mean before \( \hat{\mu} \) is re-estimated. In some cases it may be appropriate to define \( R(\theta) \) using some other norm, for example a weighted norm that places more emphasis on some deviations from \( \hat{\mu} \) than others.

Because \( G_\theta \) is posited to be a smooth transformation, for small \( \theta \) we will have the approximation

\[
G_\theta \hat{\mu}(t) \approx \hat{\mu}(t) + \theta^T g(t)
\]  

(11)

where \( g \) is a particular vector of functions, the functional derivative of \( G_\theta \) evaluated at \( \hat{\mu} \) with \( \theta = 0 \). We shall elaborate on the calculation of the vector \( g \) in particular cases below.

To first order in \( \theta_i \), then, we have

\[
\frac{\partial R}{\partial \theta_i} = \int \{ X_i(t) - G_{\theta_i} \hat{\mu}(t) \} g(t) dt,
\]

(12)

so that the residuals \( X_i(t) - G_{\theta_i} \hat{\mu}(t) \) will be orthogonal to the vector of functions \( g(t) \). Thus our procedure partitions the variability into that parallel and orthogonal to this vector of functions, but of course the vector of functions is not defined explicitly.

Once the parameters \( \theta_i \) have been estimated, we can proceed in any of the ways set out in 2 above, either building the parameters into a combined analysis, or performing a separate PCA on the residual functions. In order to carry out a 'Model 3' analysis, the data will be of the form \( (Y_i, \theta_i) \) where \( Y_i \) is the \( i \)th residual function, usually most naturally defined as \( G_{\theta}^{-1} X_i - \hat{\mu} \).

To first order, these residuals will still be orthogonal to \( g(t) \).

One of the objects of a Model 3 analysis is to provide a more interpretable version of a standard functional PCA performed on the original data. Just as in the special cases set out above, this can be achieved by choosing a suitable norm with respect to which to carry out the PCA. To do this, suppose that \( y(t) \) is orthogonal to \( g(t) \) and that \( \theta \) is small. We then have the canonical norm

\[
\| G_\theta \hat{\mu} + y - \hat{\mu} \|^2 \approx \int y^2 + \theta^T A \theta
\]
where $A$ is the matrix $\int g(t)g^T(t)dt$ of integrated squares and cross-products of the vector $g$. Performing the PCA with respect to the canonical norm will then give results comparable, to first order, with those obtained from a standard functional PCA, but with the parametric effects made explicit.

### 6.1 A special case

Suppose, now, that $G_{\theta}$ consists of a time shift by a function $b_{\theta}$ and a function addition $a_{\theta}$, so that

$$G_{\theta}\mu(b_{\theta}(t)) = \mu(t) + a_{\theta}(t)$$

and

$$G_{\theta}\mu(s) = \mu(b_{\theta}^{-1}(s)) + a_{\theta}(b_{\theta}^{-1}(s)).$$

Suppose that $G_{\theta}$ is differentiable as a function of $\theta$ so that for small $\theta$ we have $b_{\theta}(t) \approx t + \theta^T b(t)$ and $a_{\theta}(t) \approx \theta^T a(t)$ for vector functions $a$ and $b$ which are the gradients of $a_{\theta}$ and $b_{\theta}$ at $\theta = 0$. Then for small $\theta$ we will have

$$G_{\theta}\mu(s) \approx \mu(s - \theta^T b(s)) + \theta^T a(s) \approx \mu(s) - \theta^T b(s) \mu'(s) + \theta^T a(s),$$

so that the functional derivative vector

$$g(s) = a(s) - \mu'(s)b(s).$$

As a simple example, consider the time shift/overall mean model on a periodic interval of length 1, as considered in Section 5 above. This corresponds to the special case where $\theta = (\theta_1, \theta_2)$, $a_{\theta}(t) = \theta_1$ and $b_{\theta}(t) = \theta_2$. Therefore we have

$$\mathbf{a}(s) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \text{ and } \mathbf{b}(s) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

so that

$$g(s) = \begin{pmatrix} 1 \\ -\mu'(s) \end{pmatrix}.$$  

By the periodicity assumption, $\int \mu'(s)ds = 0$, and so the matrix $A$ is given by

$$A = \begin{pmatrix} \int ds & -\int \mu'(s)ds \\ -\int \mu'(s)ds & \int \mu'(s)^2ds \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & C^2 \end{pmatrix},$$

and $\theta^TA\theta = \theta_1^2 + C^2\theta_2^2$ as used in Section 5.1 above.
7 Acknowledgments

I am extremely grateful to Jim Ramsay for a number of stimulating discussions and for providing the weather data; to Richard Olshen for his continuing encouragement; and to Guy Nason for his very helpful remarks on a previous draft.

References


