Moments Based Functional Synchronization

GARETH M. JAMES[*]

Abstract

A significant problem with most functional data analyses is that of misaligned curves. Without adjustment, even an analysis as simple as estimation of the mean will fail. A common "synchronization" approach involves equating "landmarks" such as peaks or troughs. The landmarks method can work well but will fail if marker events can not be identified or are missing from some curves. It may also involve a manual identification of marker events. We develop automated alignment methods based on equating the "moments" of a given set of curves. These moments do not depend on the identification of markers. For example, the first moment is a measure of the average value of a curve in the *x*, or time, axis while the second moment measures its spread. We explore both linear and non-linear synchronization procedures. Finally, we discuss the advantages of utilizing, not only the "amplitude" information, which measures the general shape of the curves, but also the "warping" information, which measures the way the curves have been distorted in time. Illustrations are provided on functional analyses involving principal components, clustering, classification and regression.

Key Words : Moments, Curve Alignment, Procrustes, Landmark Registration.

1 Introduction

Over the last decade functional data analysis (FDA), which involves treating the entire curve or function as the unit of observation (Ramsay and Silverman, 1997), has become an increasingly important statistical approach in many different fields. Significant advances have been made in the analysis of functional data. However, one of the most commonly encountered problems, namely that of misaligned curves or functions, still possess significant difficulties. Figure 1a) provides a simple example. We have plotted ten curves produced using a mixture of two Gaussian densities with random, non-linear, shifting and stretching in the *x*-axis. The solid line through the curves represents the cross-sectional mean and illustrates how even such a simple misalignment can severely distort any analysis of the data. Figure 1b) shows that a high level of synchronization can be achieved using the approaches we discuss in this paper. The time transformations, or "synchronization functions", that were used to align the curves are plotted in Figure 1c). Some real world examples of this type of problem include growth curves where individuals experience changes at different ages and rates, microarrays observed over time where gene expression levels may have similar shapes but over different time periods and clinical trial data where patients are at different stages of disease progression upon enrollment in the study.

The problem of realigning such curves has been studied under different names in several fields. In the statistics literature it is referred to as curve registration (Silverman, 1995; Ramsay and Li, 1998) or, in the context of computing an average curve, structural averaging (Kneip and Gasser, 1988, 1992). It is also called curve alignment in biology and time warping in engineering (Sakoe and Chiba, 1978; Wang and Gasser,

---

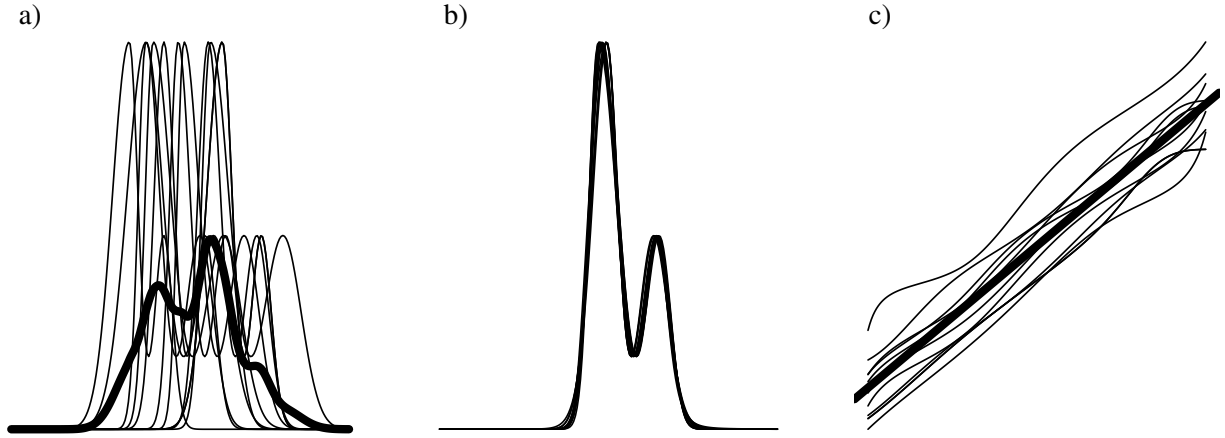[*]Marshall School of Business, University of Southern California

Figure 1: *a) Ten curves derived from a mixture of two Gaussian densities with random transformations of the x-axis, b) the synchronized curves and c) the transformations that were used to align each curve. The wider line in c) is the identity and corresponds to no transformation of Y.*

1997). Any set of curves can be decomposed into "amplitude" functions, which measure differences in the y-axis, and "warping" functions, which measure differences in location on the *x*-axis. Synchronization requires estimation of the warping functions. A number of approaches have been proposed. Marker, or landmark, registration (Kneip and Gasser, 1992) involves selecting common features in the data, such as peaks or troughs, and transforming time so that these features occur together. This method can work well when such features can be easily identified but tends to perform poorly if there are no landmarks or if marker events are missing from some curves. An alternative approach involves aligning curves using a target function. Silverman (1995) proposed registering curves using a simple shift in time such that the average squared distance between each curve and a target function is minimized. This idea was extended in Ramsay and Li (1998) using a Procrustes fitting procedure on a general nonlinear class of time transformations to provide maximal alignment to the target function subject to suitable smoothness of the transformations. The Procrustes approach is often very effective but depends heavily on the target function. Generally the cross-sectional mean is used which, as illustrated by Figure 1b), can provide misleading results if the curves are significantly misaligned. Other recent work in this area includes Wang and Gasser (1999), Kneip *et al.* (2000), Rønn (2001), and Gervini and Gasser (2004).

This paper has two goals. The first, is to develop a more robust synchronization method that incorporates the advantages of both the landmark and Procrustes approaches. Our method works to align the functions by equating the "moments" among all the curves. Recently James (2007) proposed definitions for the moments of a curve which are intended to measure the center, variance, skewness etc. of a function. Using these definitions one can calculate the first K moments for any set of curves. We develop theory that allows us to compute "average" moments for a population of curves and then align the curves by transforming time so as to approximately equate their moments. The fundamental idea is that, just as one can define the distribution of a random variable through its moments and equate two different distributions by transforming to equate the moments, we can also define the shape of a function through its moments and synchronize two curves by equating their moments. Our approach has the advantage over landmark registration that it does not rely on identifiable features. It can also outperform the Procrustes approach because it does not rely on a potentially

inaccurate target function so can be used on even very poorly synchronized curves.

The second goal is to illustrate the importance of incorporating both the synchronized curves and the warping functions into any functional data analysis. Numerous statistical analysis techniques have been adapted to the FDA paradigm. For example principal components analysis (PCA) (Rice and Silverman, 1991; James *et al.*, 2000; Rice and Wu, 2001), regression with both functional responses (Zeger and Diggle, 1994; Lin and Ying, 2001) and functional predictors (Ferraty and Vieu, 2002; Marx and Eilers, 1999; James, 2002; James and Silverman, 2005), linear discriminant analysis (James and Hastie, 2001; Ferraty and Vieu, 2003) and clustering (James and Sugar, 2003; Bar-Joseph *et al.*, 2003). However, throughout the majority of the FDA literature the curve synchronization problem is either ignored or else the warping functions are considered nuisance parameters that need to be estimated to align the curves but are then discarded from any further analysis. In practice, these functions often contain as much, or more, information about the underlying process that produces the data than the synchronized curves themselves. We provide specific examples, on growth curve, temperature and simulated data sets, of several methods of analysis that benefit from incorporating both sets of curves.

In Section 2 we outline our model for the unsynchronized curves and provide general definitions for the moments of an arbitrary function. Section 3 explores several moments based alignment approaches using both linear and nonlinear synchronization functions on real world and simulated data sets. In particular we detail the "Curve Alignment using Moments and Procrustes" (CAMP) method which utilizes the best of both the moments and Procrustes procedures. The results from four simulation studies, comparing the performance of CAMP with other synchronization methods, are reported in Section 4. Section 5 highlights the importance of utilizing both the synchronized curves and the warping functions in functional analyses ranging from PCA to supervised and unsupervised learning problems. Finally, a discussion is given in Section 6.

2   Curve Synchronization by Equating Moments

In this section we present a model for unsynchronized curves and provide definitions for the moments of an arbitrary function.

2.1   The Synchronization Model

Let $Y_1(t), Y_2(t), \ldots, Y_N(t)$ be the unsynchronized functions or curves with Y_i observed at $t_{i1}, \ldots, t_{in_i}$ where $t_{ij} \in [0, T]$. We model these curves as an "amplitude function", $Z_i(t)$, which is, stretched on the time axis according to a "warping function", $W_i(t)$, plus *iid* measurement errors. The model can be expressed as

$$Y_i(t_{ij}) = Z_i(W_i(t_{ij})) + \varepsilon_{ij}, \quad i = 1, \ldots, N. \tag{1}$$

Synchronization is only a meaningful procedure if the curves have some common underlying shape which means that the Z_i's must somehow be related to each other. We model the relationship as being of the form

$$Z_i(t) = \gamma_i + \theta_i z(t), \quad \theta_i > 0 \tag{2}$$

where $z(t)$ is a function common to all curves and γ_i and θ_i are considered random. Notice that, if we let $Y_i^S(t) = Z_i(W_i(t))$ denote the smooth version of $Y_i(t)$, then using (2) we can re-express (1) in the form,

$$z(t_{ij}) = \frac{Y_i^S(X_i(t_{ij})) - \gamma_i}{\theta_i}, \quad i = 1, \ldots, N \tag{3}$$

3

where $X_i = W_i^{-1}$ is the ith "synchronization function". We model $E\varepsilon_{ij} = 0, E\gamma_i = 0, E\theta_i = 1$ and $EX_i(t) = t$ for all t so that z can be taken as the "average" of the synchronized curves. Our aim is to recover estimates of the unobserved components, γ_i, θ_i, X_i and z from the observed curves, Y_i. These estimates can then be used in standard functional analyses such as principal components, clustering, classification or regression. The key to recovering the unobserved components is to obtain reliable estimates of the X_i's. Once this is achieved the other components can be estimated with relative ease.

The landmark and Procrustes curve alignment approaches can both be viewed as methods for computing z and synchronizing the curves to minimize distortion between the Y_i's and z. With landmark registration z is taken to be, for example, one of the Y_i's, and distortion between z and the Y_i's is measured via differences in certain marker events. With Procrustes, the cross-sectional average is used to estimate z and distortion is measured as the mean squared difference between z and the Y_i's. Both methods will suffer if z is poorly estimated. For landmark registration this will happen if marker events are missing or inconsistent among curves. Alternatively, for Procrustes, problems will occur if the cross-sectional average does not represent the typical shape of the synchronized curves. In this paper, rather than attempting to estimate z directly, we instead estimate its moments. We then measure distortion via differences between the moments of z and those of the synchronized curves. This approach has the advantage that the moments can be accurately estimated even in situations where there are no marker events or where the Y_i's are poorly synchronized.

2.2 Defining the Moments of a Function

In this section we outline the definitions for the moments of an arbitrary function, g, proposed in James (2007). They first introduce the concept of a "feature function", $I_g(t)$, for g and impose the constraints

$$I_g(t) \geq 0 \quad \text{and} \quad \int I_g(t)dt = 1.$$

I_g can be thought of as a weighting function with high weight placed on time points related to a particular feature. Depending on the properties of our data we may wish to concentrate on characteristics corresponding to local features, such as maximums or minimums, or alternatively ones that correspond to more global characteristics such as the slope at a given time. For a given choice of I_g, the first moment of g is defined as

$$\mu_g^{(1)} = \int t I_g(t)dt \tag{4}$$

and the kth moment by

$$\mu_g^{(k)} = \int \left(t - \mu_g^{(1)}\right)^k I_g(t)dt, \quad k \geq 2. \tag{5}$$

Equation (4) provides a measure of the center of g on the time axis. Alternatively, $\mu_g^{(2)}$ measures variability in g. Note that the variability is measured in relation to the time axis and not the y, or amplitude, axis. A curve could vary significantly in the y-axis, but still have a low value for $\mu_g^{(2)}$.

To better understand the properties of $\mu^{(k)}$ one may examine the relationship between the moments of a function $h(s)$ and those of the shape invariant function $h(\frac{s-a}{b})$. In this formulation, $h(s)$ is stretched, about $s = 0$, by a factor b and shifted to the right by a. Hence, since $\mu^{(1)}$ is a measure of the center of a function and $\mu^{(k)}$ is a measure of variability about the center, stretching by a factor b should multiply the first moment by b and the kth moment by b^k. For example, one would expect that $\mu_{h(\frac{s-a}{b})}^{(2)}$, which measures the variability of the transformed function, would equal $b^2 \mu_{h(s)}^{(k)}$. Similarly a shift of a should add a to the first moment

4

and leave the higher order moments, which are centered around the first moment, unchanged. This can be expressed mathematically as

$$\mu^{(1)}_{h\left(\frac{s-a}{b}\right)} = b\mu^{(1)}_{h(s)} + a \quad \text{and} \quad \mu^{(k)}_{h\left(\frac{s-a}{b}\right)} = b^k\mu^{(k)}_{h(s)}, \quad k \geq 2. \tag{6}$$

Theorem 1 shows that, provided one utilizes a certain family of feature functions, these properties will hold.

Theorem 1. *Suppose that $I_g(t)$ is chosen such that*

$$I_{g\left(\frac{s-a}{b}\right)}(t) \propto I_{g(s)}\left(\frac{t-a}{b}\right) \tag{7}$$

for all a, g and $b > 0$. Then (6) will hold for any function $h(s)$.

Condition (7) holds for many large classes of feature functions. We explore several of them here. The first involves weighting according to the absolute mth derivative of the curve i.e.

$$I_g^{(m)}(t) = \frac{|g^{(m)}(t)|}{\int |g^{(m)}(s)|ds}, \quad m = 0, 1, 2, \dots \tag{8}$$

where $g^{(m)}$ is the mth derivative of g. With $m = 0$ this function puts highest weight on large absolute values of g. With $m = 1$ most weight is placed on time points where g has a large slope and would be used when we are most interested in regions where a curve is changing rapidly. Setting $m = 2$ searches for points with greatest curvature etc. Equation (8) can be considered to be searching for global characteristics of a curve because $I_g^{(m)}$ is likely to spread its mass over all time points. Alternatively, one could adopt a more local approach where most of the weight is concentrated around the time points corresponding to a specific feature in the data. For example, as $r \to \infty$, (9) and (10) will increasingly concentrate their weight on the global maximum,

$$I_g^{\max}(t) \propto (g(t) - \min\{g(t)\})^r, \tag{9}$$

and the global minimum,

$$I_g^{\min}(t) \propto (\max\{g(t)\} - g(t))^r, \tag{10}$$

of $g(t)$. We may wish to search for local, as well as global, maximums and minimums. In this case one could utilize

$$I_g^{\text{local}}(t) \propto \begin{cases} \exp\left(-r\frac{|g^{(1)}(t)|}{\sqrt{|g^{(2)}(t)|}}\right) & g^{(2)}(t) \neq 0 \\ 0 & g^{(2)}(t) = 0. \end{cases} \tag{11}$$

This function places maximum weight on points where the first derivative is zero. However, $I_g^{\text{local}}(t)$ is also high for points with a low first derivative but a high second derivative. Thus, the function effectively searches for local maximums or minimums where g is changing most rapidly. As $r \to \infty$, $I_g^{\text{local}}(t)$ will place all its weight on the regions around local turning points.

The feature functions given by (8) through (11) all satisfy (7) and hence the moments corresponding to them all possess the desirable properties given by (6).

Corollary 1. *When utilizing $I_g^{(m)}, I_g^{\max}, I_g^{\min}$ or I_g^{local} condition (7) is satisfied, and hence (6) holds. In addition (7) is satisfied for any $I_g^{\phi}(t) \propto \phi(g(t))$ where $\phi(t)$ is an arbitrary function.*

Proofs of both Theorem 1 and Corollary 1 are given in James (2007). These classes of feature functions represent only a fraction of the possible choices one could utilize. In fact one of the strengths of our approach is the ability to design functions which best suit ones particular data. While our theoretical results are general to many feature functions, to maintain consistency, we have opted to use $I_g^{(1)}$ for all the data sets that we illustrate.

3 Curve Alignment Procedures

In this section we outline three different moments based procedures for synchronizing functional or curve data. While the methods differ somewhat in their implementation they all have the same four basic steps.

1. Produce a smoothed version of each curve, $\tilde{Y}_i$, using, for example, a smoothing spline.

2. Choose a feature function and calculate the first K moments of the $\tilde{Y}_i$'s i.e. $\mu_{\tilde{Y}_i}^{(1)}, \ldots, \mu_{\tilde{Y}_i}^{(K)}$, $i = 1, \ldots, N$.

3. Use the kth moment of the $\tilde{Y}_i$'s to estimate the corresponding moment of Z_i.

4. Choose X_i so as to equate $\mu_{\tilde{Y}_i(X_i)}^{(k)}$ with $\mu_{Z_i}^{(k)}$, either exactly or approximately, for all i and $1 \le k \le K$.

There is no restriction on the upper bound of K but in general we have found values of K between one and four provide the best results. Equating the first moment ensures that the curves all have the same center while equating the next two moments forces equal spread and skewness. By setting $\mu_{\tilde{Y}_i(X_i)}^{(k)}$ to $\mu_{Z_i}^{(k)}$ we guarantee that the synchronized curves maintain the same "average" shape as the original observed curves. Figure 2 provides an example of this approach on a set of ten simulated curves. Figure 2a) illustrates the original data while Figures 2b) through d) respectively show the synchronized curves after equating the first, first two, and first three moments. At each step the synchronization improves. The high level of alignment in Figure 2d) suggests the curves only differ in shape in their first three moments.

The fundamental idea behind our approach is that we wish to change the shape of the observed curves as little as possible, and hence to incorporate most of the shape information into the Z_i's, leaving in the X_i's only that which differs among the curves. Implicitly, this means that there is little shape information contained in the X_i's so they can be accurately estimated using the differences among the first few moments of the curves. Our approach can be successful in situations where, for example, the Procrustes method fails because our method only trys to estimate these moments rather than an entire target function.

In practice there are several issues that must be resolved to implement our moment equating strategy. The first problem involves the estimation of the $\mu_{Z_i}^{(k)}$'s since Z_i is unobserved. However, this calculation is greatly simplified if the feature function is chosen such that,

$$I_{a+bg}(t) = I_g(t) \tag{12}$$

for all a, b and g. Provided (12) holds then, under the model given by (3), it is easily seen that $\mu_{Z_i}^{(k)} = \mu_z^{(k)}$ for all i and k. Condition (12) can be shown to hold for many feature functions including I^{max}, I^{min} and $I^{(m)}$ for $m \ge 1$. Hence, in general, we can restrict our selves to estimating $\mu_z^{(1)}, \ldots, \mu_z^{(k)}$. We show how to estimate these moments in Section 3.1. A second, related problem, involves identifiability issues. All synchronization approaches, such as the landmark or Procrustes methods, suffer from a fundamental identifiability problem between X and Z. Namely, for any observed curve $Y(t)$ there exist an infinite number of pairs $\{X(t), Z(t)\}$ such that $Y(X(t)) = Z(t)$. Theorem 2 formally summarizes this idea.
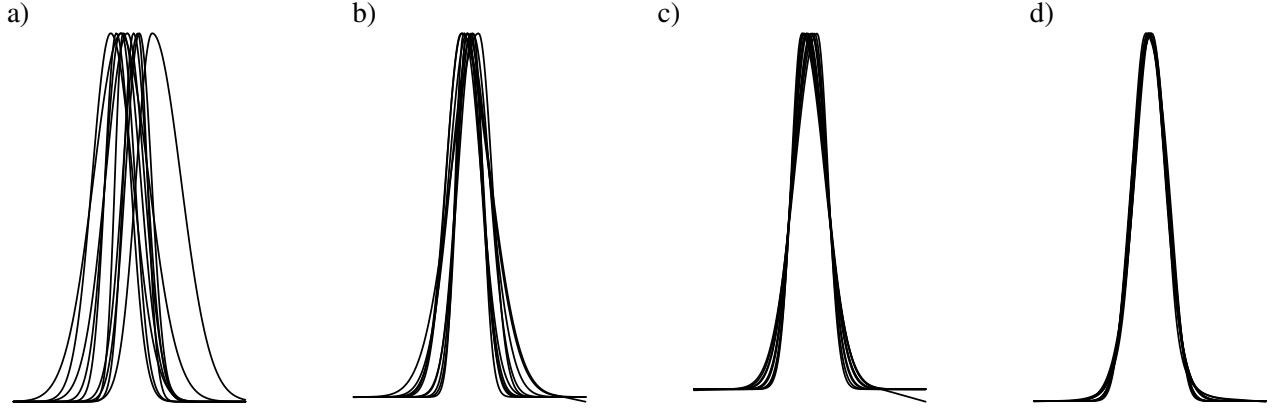
Figure 2: *a) Unsynchronized data, b) equating the first moment, c) equating the first two moments and d) equating the first three moments.*

Theorem 2. *Suppose that $Y(t)$ is a continuous function defined on $[t_0, t_n]$ with local minimums and maximums at $t_1, t_2, \ldots, t_{n-1}$. Let $Z(s)$ be any continuous function defined on $[s_0, s_n]$ with local minimums and maximums at $s_1, s_2, \ldots, s_{n-1}$ and*

$$Y(t_i) = Z(s_i) \quad i = 0, 1, \ldots, n. \tag{13}$$

Then for any such Y and Z there exists a continuous strictly increasing function $X(s)$ such that $Y(X(s)) = Z(s)$ for $s_0 \leq s \leq s_n$.

See James (2007) for a proof of this result. One important consequence of Theorem 2 is that, without constraints on X_i and Z_i, the shape of Y_i is an unidentifiable combination of the synchronization function, and the amplitude curve. We address the identifiability problem by placing constraints on the X_i's and/or the Z_i's via equation (2). The final problem relates to the best way to estimate the X_i's such that the moments of the curves are equated.

We discuss several approaches to these problems in the following three sections. The first approach, which we outline in Section 3.1, involves restricting the parametric form of $X(t)$ to the class of linear functions. Under this restriction we develop a number of useful theoretical results. In particular we show that one can guarantee identifiability and produce consistent estimates of the $\mu_z^{(k)}$'s. In Section 3.2 we discuss a second approach which utilizes a general non-parametric class of monotone increasing synchronization functions. In this case we place additional constraints on the variability of the X_i's to ensure identifiability. Finally, in Section 3.3, we outline an alignment method which utilizes a mixture of both the moments and Procrustes approaches.

3.1 Linear Synchronization Functions

Restricting to linear synchronization functions i.e. $X(t) = \alpha + \beta t$ has a number of advantages. First, it becomes possible to express the moments of z as functions of $\mu_{Y_i^S}^{(k)}$. Theorem 3 summarizes the relationship.

Theorem 3. *Suppose $X(t) = \alpha + \beta t$. Then using the model given by (3) and a feature function satisfying (7) and (12),*

$$\mu_z^{(k)} = \left(E \sqrt[k]{\mu_{Y_i^S}^{(k)}} \right)^k, \quad k \geq 1,$$

where the expectation is taken over Y_i^S.

Theorem 3 shows that for feature functions such as I^{max}, I^{min} and $I^{(m)}$ the moments of z can be recovered from the moments of the Y_i^S's. In particular, it suggests the form of an estimator for $\mu_z^{(k)}$, namely,

$$\widehat{\mu_z^{(k)}} = \left(\frac{1}{N} \sum_{i=1}^{N} \sqrt[k]{\mu_{\tilde{Y}_i}^{(k)}} \right)^k, \tag{14}$$

where $\tilde{Y}_i$ is a smoothed version of Y_i. Since $\mu_{\tilde{Y}_i}^{(k)}$ can always be computed using (4) and (5), equation (14) provides a simple method for estimating $\mu_z^{(k)}$, for any value of k. Consider the following mild assumptions on the convergence of $\tilde{Y}_i$:

A-1 $\tilde{Y}_{i,n}$ and Y_i^S are both uniformly continuous functions, sampled at n time points, such that $\tilde{Y}_{i,n}(t) \to Y_i^S(t)$ *a.s.* for all i and t as $n \to \infty$.

A-2 $I_g(t)$ is a uniformly continuous feature function such that, whenever A-1 holds, $I_{\tilde{Y}_{i,n}}(t) \to I_{Y_i^S}(t)$ *a.s.* for all i and t as $n \to \infty$.

Then, under A-1 and A-2, Theorem 4 shows that $\widehat{\mu_z^{(k)}}$ is a consistent estimator.

Theorem 4. *Provided A-1 and A-2 hold, then, under the conditions given in Theorem 3, $\widehat{\mu_z^{(k)}} \to \mu_z^{(k)}$ a.s. as n and N approach infinity.*

A-2 will hold for most feature functions, including I^{max}, I^{min} and $I^{(m)}$. A-1 will also hold for both kernel and spline smoothers provided the smoothing parameter is appropriately adjusted as a function of n and that Y_i^S is a smooth function (Priestley and Chao, 1972; Silverman, 1984, 1985).

The second advantage of using a linear synchronization function is that one can ensure identifiability in the sense that there is one and only one X_i that satisfies

$$\mu_{Y_i^S(X_i)}^{(k)} = \mu_z^{(k)}, \quad k = 1, \ldots, K \tag{15}$$

for $K = 2$. In fact, for data generated from a linear warping function, equating the first two moments will automatically equate all the moments. Theorem 5 summarizes this result.

Theorem 5. *Under the assumptions of Theorem 3, the linear synchronization function, $X_i(t) = \alpha_i + \beta_i t$, that satisfies (15) for $K = 2$ is unique and is given by*

$$\alpha_i = \mu_{Y_i^S}^{(1)} - \beta_i \mu_z^{(1)}, \quad \beta_i = \sqrt{\frac{\mu_{Y_i^S}^{(2)}}{\mu_z^{(2)}}}. \tag{16}$$

8
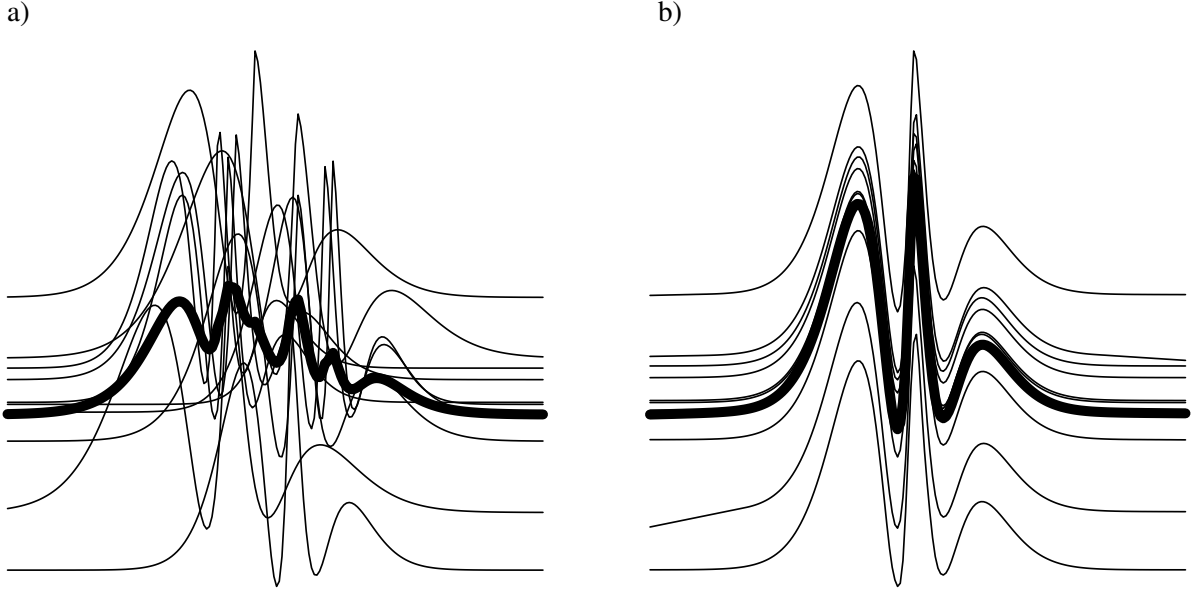
a)   b)

Figure 3: *a) Unsynchronized data, b) synchronized data using a linear warping function.*

Theorem 5 guarantees that, given $\mu_z^{(1)}$ and $\mu_z^{(2)}$, $X_i(t)$ will be unique and also suggests that an obvious strategy for estimating $X_i(t)$ would be to replace $\mu_{Y_i^S}^{(k)}$ by $\mu_{\tilde{Y}_i}^{(k)}$ and to replace $\mu_z^{(k)}$ by $\widehat{\mu_z^{(k)}}$ in (16). This provides a computationally efficient, simple closed form solution. We call this synchronization approach "Linear Curve Alignment by Moments (Linear CAM)". Figure 3 provides an example of the Linear CAM approach on a simulated data set. The raw curves are plotted in Figure 3a). These curves were generated from a common shape function. They were then randomly shifted vertically by a constant amount and linearly shifted and stretched horizontally by random amounts. Figure 3b) illustrates that a high level of synchronization can be achieved using Linear CAM. The wider line through the center represents the cross-sectional mean and is an accurate representation of the true $Z(t)$ from which the original data was generated. In addition to the computational advantages that Linear CAM possess, Theorem 6 shows that it can be used to produce consistent estimators for $z(t)$, as well as the $X_i(t)$'s, γ_i's and the θ_i's.

Theorem 6. *Suppose the Y_i's are observed at $t_1, \dots t_n$. Let $\hat{\alpha}_i = \mu_{\tilde{Y}_i}^{(1)} - \hat{\beta}_i \widehat{\mu_z^{(1)}}$, $\hat{\beta}_i = \sqrt{\dfrac{\mu_{\tilde{Y}_i}^{(2)}}{\widehat{\mu_z^{(2)}}}}$ and $\hat{X}_i(t) = \hat{\alpha}_i + \hat{\beta}_i t$.*

Also let

$$\overline{Z}(t) = \frac{1}{N}\sum_{i=1}^{N}\tilde{Y}_{i,n}(\hat{X}_i(t)), \qquad \widehat{\gamma}_i = \frac{2}{n(n-1)}\sum_{j=1}^{n-1}\sum_{l=j+1}^{n}\frac{Z_i^j\tilde{Y}_{i,n}(t_l)-Z_i^l\tilde{Y}_{i,n}(t_j)}{Z_i^j-Z_i^l}, \qquad \widehat{\theta}_i = \frac{1}{n}\sum_{j=1}^{n}\frac{\tilde{Y}_{i,n}(t_j)-\widehat{\gamma}_i}{Z_i^j}$$

where $Z_i^j = \overline{Z}\left(\frac{t_j-\hat{\alpha}_i}{\hat{\beta}_i}\right)$. Then, provided A-1, A-2 and the assumptions of Theorem 3 hold, $\hat{X}_i(t)$, $\overline{Z}(t)$, $\widehat{\gamma}_i$ and $\widehat{\theta}_i$ respectively converge to $X_i(t), z(t), \gamma_i$ and θ_i a.s. as $N, n \to \infty$.

3.2 Nonlinear Synchronization Functions

Linear CAM provides an extremely computationally simple and theoretically rigorous alignment procedure. The plots in Figure 3 illustrate that it can also produce accurate synchronizations provided that the warping

functions are close to linear. The curves in Figure 2c) provide another example of a Linear CAM synchronization. While there is a reasonable level of alignment it is clear from Figure 2d) that one can achieve better synchronization. The data in Figure 2 was generated using non-linear warping functions which accounts for the reduced performance of Linear CAM. As one might expect, in general, the accuracy of Linear CAM deteriorates as the warping functions become more non-linear because one can only equate the first two moments. Potentially, improved alignment can be achieved using a non-linear synchronization function and equating additional moments. Figure 2d) provides an example using the first three moments.

When $X_i(t)$ is modeled as a non-linear increasing function there will no longer be a unique solution to (15). Hence, some form of regularization must be imposed on the synchronization functions. In general, one would like to achieve as high a level of alignment as possible while minimizing the distortion to the shape of the Y_i's. An obvious measure of distortion is given by

$$P(X_i) = \int (X_i(t) - X_i(0) - t)^2 dt. \tag{17}$$

$P(X_i)$ penalizes deviations of X_i from the 45 degree line $X_i(0) + t$. A synchronization function that conformed exactly to $X_i(0) + t$ would involve only a constant shift of Y_i and hence no alteration to its shape. Hence the "Non-linear CAM" approach utilizes the synchronization function that minimizes

$$Q_i = \sum_{k=1}^{K} \lambda_k \left(\mu_{\tilde{Y}_i(X_i)}^{(k)} - \widehat{\mu_z^{(k)}} \right)^2 + P(X_i) \tag{18}$$

where the λ_k's determine the level of agreement between the moments. As the λ_k's converge to infinity, minimizing (18) produces the least variable X_i subject to (15) holding for the first K moments. Alternatively, as the λ_k's converge to zero minimizing (18) will only cause constant shifts in Y_i and no alteration to its shape.

To minimize (18) we first reexpress X_i as

$$X_i(t) = \alpha_i + \int_0^t \exp(f_i(s)) ds$$

which involves no loss of generality because a synchronization function must be monotone increasing. We then model $f_i(s)$ as coming from a finite q-dimensional basis, $\mathbf{b}(s)$, i.e. $f_i(s) = \mathbf{b}^T(s)\eta_i$ where η_i are the associated basis coefficients. We utilize a b-spline basis. For a given value of η_i one can then compute X_i and hence the $\mu_{\tilde{Y}_i(X_i)}^{(k)}$'s. Therefore, for any particular values of the $\widehat{\mu_z^{(k)}}$'s and choice of the λ_k's, (18) can be minimized over η_i using any standard non-linear optimization function. The optimization is relatively efficient because one can calculate the analytic derivatives of Q_i.

An additional difficulty when using non-linear synchronization functions involves the estimation of $\mu_z^{(k)}$. Theorem 3 only holds exactly for linear synchronization functions. For certain feature functions, it is still possible to calculate general non-linear expressions for the moments of z. However, the equations are difficult to evaluate on real data sets. In practice we have found the estimates produced by (14) to be accurate enough to preserve the correct shape of the Z_i's provided the X_i's do not exhibit extreme non-linearity.

A direct comparison of the Linear and Non-linear CAM approaches is given in Figures 2c) and d). The Non-linear CAM, which in this case equated the first three moments, produces improved results because the curves were generated using non-linear warping functions. We have plotted the corresponding estimated linear and non-linear synchronization functions in Figure 4. Notice that, while the non-linear functions exhibit more curvature, the penalty $P(X_i)$ tends to force the X_i's closer to a 45 degree line which reduces distortion of the shape of the Y_i's.
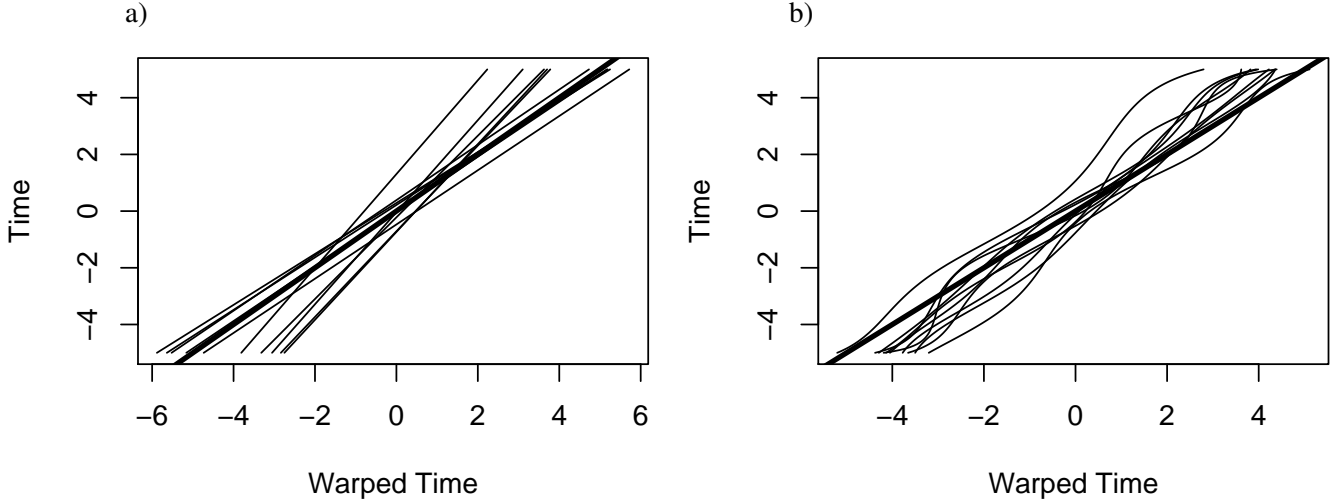
Figure 4: *a) Linear and b) non-linear synchronization functions for the data from Figure 2.*

3.3 Curve Alignment by Moments and Procrustes (CAMP)

To date, one of the most effective curve registration approaches is the Procrustes fitting procedure of Ramsay and Li (1998). This method involves defining a target function, T, and estimating curves X_i such that

$$\int (\tilde{Y}_i(X_i(t)) - T(t))^2 dt \tag{19}$$

is minimized subject to smoothness constraints on the X_i's. Ramsay and Li demonstrate that this approach can produce impressive curve alignment provided a reasonable target function can be identified. In most situations a predefined target is not given so the cross-sectional average of the unsynchronized curves is generally utilized. Hence, the usefulness of this target function will depend on the degree of agreement between the curves.

Conceptually, the Procrustes method and the moments approach of this paper are both attempting to estimate z and then calculate the X_i's to provide maximum alignment of the curves with z. The difference lies in the method of estimation. The Procrustes procedure uses T, the cross-sectional average, while the moments approach only attempts to compute the first few moments of z. Both methods have potential advantages. The cross-sectional average works well when the alignment between curves is close enough to produce a reasonable estimate of z but can break down if the data exhibits a significant lack of alignment. Alternatively, since the moments approach does not attempt to directly estimate z it can still produce reasonable estimates of the moments even in situations where the cross-sectional average performs poorly. However, some of the finer detail in z may be lost by only estimating the first few moments. Hence, a natural procedure involves attempting to utilize the best of both methods to produce more accurate estimates of z and hence better alignment of the curves. We call this approach "Curve Alignment by Moments and Procrustes (CAMP)". To implement CAMP we first estimate the target function. Since Linear CAM essentially involves no computational burden we estimate T using the cross-sectional average of the linear synchronized curves. We then calculate the estimated moments of z, in the same way as with CAM, and compute the X_i
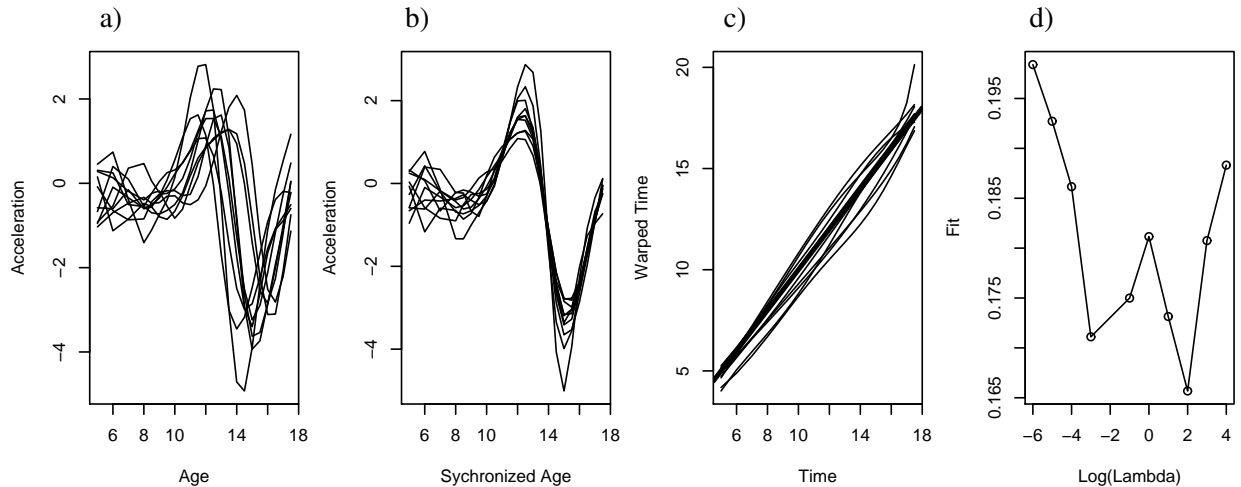
11

Figure 5: *Acceleration curves from the Berkeley growth curve study. a) Unsynchronized curves, b) after alignment using CAMP, c) the estimated warping functions, d) degree of misalignment for different values of the λ_k's.*

which minimizes

$$Q_i = \sum_{k=1}^{K} \lambda_k \left(\mu_{\tilde{Y}_i(X_i)}^{(k)} - \widehat{\mu_z^{(k)}} \right)^2 + \lambda_P \int (\tilde{Y}_i(X_i) - T(t))^2 dt + P(X_i). \tag{20}$$

Equation 20 incorporates the Procrustes criterion, (19), with the CAM criterion given by (18). As λ_P approaches zero CAMP reduces to CAM and the estimation of z is performed purely in terms of its first K moments. Alternatively, as the λ_k's approach zero CAMP reduces to a version of the Procrustes algorithm and z is estimated based solely on the cross-sectional average. However, when positive weight is placed on both sets of penalty terms the estimation of z utilizes a mixture of the two, often producing improved levels of synchronization over that from either approach used separately.

Figure 5 illustrates the CAMP fit applied to the acceleration curves of ten boys from the Berkeley growth curve study (Tuddenham and Snyder, 1954). Figure 5a), which plots the unsynchronized curves, shows a clear trend of positive and then negative acceleration during the teenage years. However, the onset times, and spread, of these growth spurts can differ by several years. Figure 5b) gives the curves after synchronization using CAMP. A significant degree of alignment, especially during the latter time periods, is achieved. Interestingly, Figure 5c), which plots the synchronization functions, shows that the alignment can be achieved with relatively little warping of time. Finally, Figure 5d) plots a measure of curve synchronization as a function of the λ_k's. Here we measure synchronization using the root median squared deviation between the aligned curves and their cross-sectional average. As the λ_k's approach zero CAMP produces a Procrustes fit and the level of alignment deteriorates. Similarly, as the λ_k's approach infinity CAMP places all weight on equating the moments and again the alignment deteriorates. A clear optimum is achieved using a mix of both criterion.

4 Simulation Study

In this section we compare the performance of CAMP with several other approaches over four sets of simulations. For each simulation 50 data sets of ten curves each were generated from a given distribution. Five different synchronization methods were then applied to each data set. We compared CAMP to the

12

Method	Simulation							
	One		Two		Three		Four	
	Fit	Var	Fit	Var	Fit	Var	Fit	Var
Peak Alignment	58.6	0.00	80.6	0.00	24.4	0.00	50.9	0.00
Procrustes	12.5	0.72	36.7	2.95	16.2	0.18	34.1	1.31
Linear CAM	20.4	1.16	30.0	2.68	42.7	0.31	26.0	2.40
Non-Lin CAM	13.6	0.43	29.6	2.62	40.1	0.05	14.8	1.76
CAMP	6.9	0.45	18.7	2.56	9.8	0.17	7.8	0.70

Table 1: Results from four simulations on five different alignment methods. Standard errors on the fits ranged from 0.5 to 1.9.

linear and non linear CAM methods as well as the Procrustes approach. We also used a simple landmark alignment procedure where the curves were shifted to align the highest peak. For each set of simulations the λ parameters were chosen by selecting the values that provided maximum alignment on a preliminary data set. The quality of fit was measured by the root mean squared deviation of each synchronized curve from the sample mean curve, $\bar{Z}(t)$. The results are summarized in Table 1. Two numbers are provided for each simulation-method pair. The first gives the average root mean squared deviation on the synchronized curves as a percentage of that for the unsynchronized curves. The second number is the square root of $P(X_i)$ given by (17). This term measures the variability in the warping functions and hence gives a measure of the degree of warping that each method requires to synchronize the curves.

Simulation one used the distribution from Figure 2 and provided a simple test case. The curves were generated using a standard Gaussian density with non-linear warping functions. Despite this non-linearity the Linear CAM approach still produced a considerable level of synchronization and significant improvements over the peak alignment method. The Non-linear CAM gave a further improvement as did Procrustes which worked well because the sample mean on the original curves was fairly representative of the data. However, CAMP clearly dominated the other methods. For Simulation two we generated data from a distribution similar to that illustrated in Figure 3 except that vertical shifts between curves were removed. This data used a linear warping function so the Linear and Non-linear CAM's produced similar results. Procrustes did not perform as well because the sample mean curve was not representative of the general shape of the individual curves. Again CAMP provided a considerable improvement over the other approaches. Simulation three used curves equal to a sin function for values of x between 0 and 2π and equal to 0 for other values of x. The functions were then shifted and stretched in time and the resulting curves were plotted between 0 and 2π. The CAM methods had difficulty with this data because, as a result of the warping, some of the functions only contained partial sin curves which affected the moment estimates. The Procrustes method performed well but, despite the CAM problems, CAMP was still able to provide the best fit. Finally, Simulation four generated curves from the Gaussian mixture illustrated in Figure 1. The performance of the Procrustes method was somewhat inferior to the other approaches because the cross-sectional mean produced a poor representation of the data. The warping functions in this simulation were non-linear so the Non-linear CAM produced improved results over the Linear CAM. As with the other simulations, CAMP produced the highest levels of synchronization. Notice that, among the various methods, CAMP typically has one of the lowest variances in the warping functions, suggesting that it is able to achieve high levels of synchronization with less distortion of the shape.
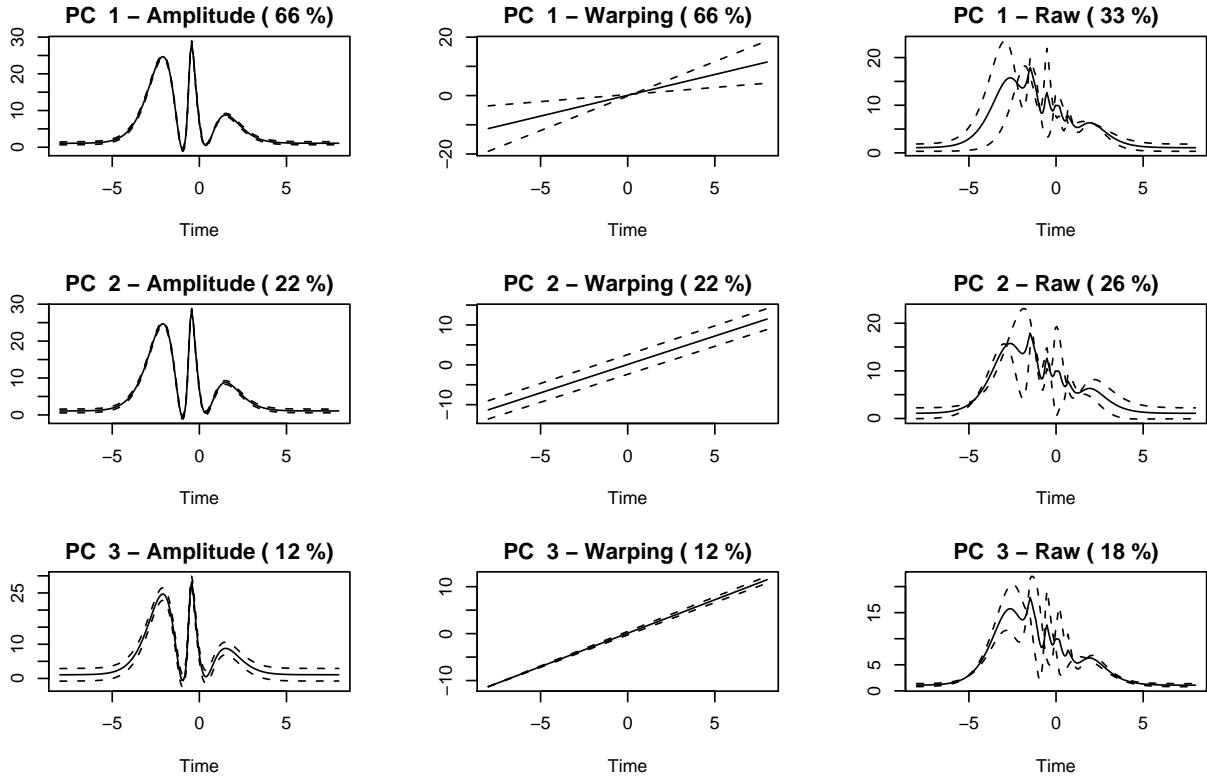
Figure 6: *The first three principal components calculated on data similar to that in Figure 3. The solid lines represent the mean curves while the dashed lines correspond to each principal component multiplied by its associated standard deviation and added or subtracted from the mean. The percentage of variability associated with each component is given in parentheses.*

5 Applications of Synchronized Curves

Any observed curve, Y_i, can be viewed as a composition of an amplitude and a warping function and using a moments based approach we have derived the CAM and CAMP methods for performing a decomposition of Y_i into these two parts. Once a set of curves has been aligned, one often disregards the synchronization functions. However, such an approach is inefficient because a significant proportion of the information provided by Y_i may be contained in the synchronization function. In this section we illustrate methods for utilizing the moments and warping functions, as well as the corresponding amplitude functions, to produce superior results in PCA, clustering, classification and regression settings.

5.1 Principal Components on Functional Data

Perhaps the simplest form of analysis on curve data is that of estimating the mean function using the cross-sectional average. Unfortunately, even such a simple analysis has been shown to produce estimates that fail to properly represent the underlying shape for unsynchronized curves (Kneip and Gasser, 1992; Gasser and Kneip, 1995). The mean curve from Figure 3a) provides a particularly extreme example. Theorem 6 shows that this problem can be solved by taking the average of the aligned curves, $\overline{Z}(t)$, which is a consistent estimator of the mean curve.

A related, but somewhat less well studied problem, is that of estimating principal component curves on unsynchronized data. As with standard PCA, principal component curves provide the main modes of variation of a set of functions. Just as with the mean, estimating principal components using the unsynchronized curves can provide misleading estimates of the type of variability. However, simply computing the principal components on the aligned curves also makes little sense because such an approach ignores the variability associated with warping over time. A method that simultaneously estimates the types of variation in the amplitude and warping functions is needed. One simple solution involves sampling both the amplitude and warping functions over a grid of *m* equally spaced time points, combining the two sets of observations together and performing principal components on the resulting *N* by 2*m* matrix. The first *m* components of each eigenvector will give the variability in "amplitude space" while the remainder correspond to "warping space". Unfortunately, this approach may not produce reasonable answers because there is no reason that the two spaces need be on comparable scales. Instead we decompose the variability in the unsynchronized curves into that associated with warping of time and that associated with amplitude shifts. This decomposition is achieved by calculating the mean variance over all time points for both the unsynchronized and synchronized curves and taking the variability associated with warping as the difference between the two. We then standardize the data so that the average variability over all time points in the warping space is directly proportional to the amount of variability associated with warping. Finally, we apply PCA to the *N* by 2*m* matrix produced from the standardized data. This standardization has the effect of placing both sets of curves on a comparable scale. Furthermore, the warping curves are upweighted in situations where most variability results from a lack of synchronization and vice versa when the curves are well synchronized.

Figure 6 provides an example of this approach applied to data simulated from the same distribution as that of Figure 3. Each row corresponds to a principal component. The first column illustrates variability associated with changes in amplitude while the second column corresponds to warping over time. The data was simulated by taking a generating function, applying a linear shift and stretching of time and then a constant vertical shift in the y-axis. The principal components break down neatly into these three effects. The first corresponds to linear stretching of each curve with no amplitude variation on the y-axis. The second indicates constant shifts over time and the third a constant amplitude shift with no warping. Notice that about 88% of the variability in these curves can be attributed to warping rather than differences in amplitude. All of this warping variability would be missed if one simply performed principal components on the synchronized curves. For comparison purposes, the final column provides the principal components calculated directly from the unsynchronized data. Attempting to perform a principal components analysis on the unsynchronized curves clearly produces nonsense.

5.2 Clustering and Classification with Functional Data

As was demonstrated in the previous section, it is often the case that a curve can be more meaningfully represented using both its warping and amplitude functions. The Berkeley growth curve data provides a good illustration in both clustering and classification contexts. Figure 7a) plots two dimensional summaries of the acceleration curves for the 39 boys and 54 girls participating in the study. We have plotted the difference between maximum and minimum acceleration, for each individual, on the x-axis and their average level of acceleration on the *y*-axis. There is some evidence of differences between the genders but the distinction is not clear. If we treat this as a two class classification problem the cross-validated error rate using 3-nearest neighbors is 20.4%. Alternatively, if we use *k*-means to cluster the data without reference to gender 18.4% of individuals are assigned to the "incorrect" cluster. The range and average acceleration only utilize amplitude information from the *y*-axis and fail to account for differences in the time courses of each curve. Alternatively, Figure 7b) plots the first two moments for each acceleration curve. These moments indicate
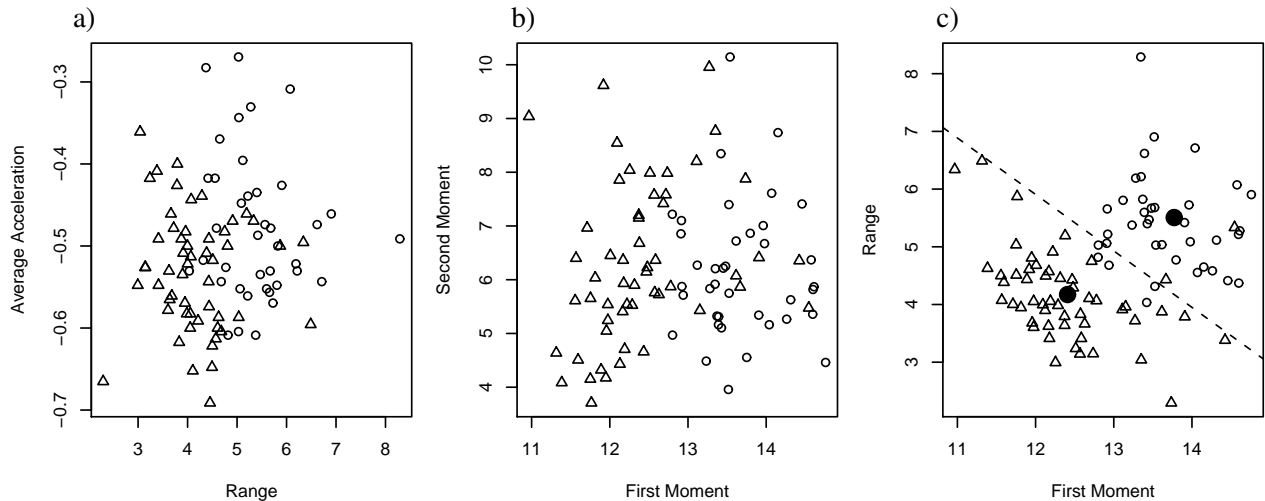
Figure 7: *Summaries of acceleration rates of growth for* 39 *boys and* 54 *girls from the Berkeley growth curve data. Girls are represented by triangles and boys by open circles. The closed circles and dashed line in c) represent the cluster centers and resulting cluster boundaries using k-means.*

the mean and variance of each curve on the time or *x*-axis. Again, there is some evidence of differences between genders. In the classification setting the cross-validated error rate is 17.2% and when clustering the misassignment rate is between 12.9% and 47.3%, depending on initialization. Together, Figures 7a) and b) represent the amplitude and warping components of each curve but individually both fail to adequately separate the genders. Figure 7c) combines the two together by plotting the first moment versus the range. Now the separation becomes considerably clearer. When using *k*-nearest neighbors the cross-validated error rate drops to only 6.5% and using *k*-means 7.5% of individuals are misassigned.

Figure 8 provides another example of this effect. Here we have plotted summaries of the smoothed daily temperatures, in Montreal, for every year from 1961 to 1994 (Ramsay and Silverman, 2002). Figures 8a) and b) respectively plot the absolute first and second moments for each year versus the minimum yearly temperature. Again, the plots provide information about both the amplitude and warping effects. Most years appear to have relatively similar temperature patterns. However, this analysis reveals several outliers. Figure 8a) shows that 1981 and 1989 both had unusually low minimum temperatures as well as significantly shifted seasons. From Figure 8b) we see that 1963, 68, 70 and 81 all had low minimums and increased time variability while 1964 had a very high minimum with low time variability. These outliers could potentially be noted by examining the amplitude or warping effects individually but the effect becomes significantly clearly when they are combined.

5.3 Regression with Functional Data

A similar application to the classification problem involves performing regression using functional predictors to predict the value of a continuous scalar variable. A common approach is to model the scalar response as a linear function of the observed predictor curve. However, in reality the response may more accurately be modeled as a linear function of both the warping and amplitude curves. As an illustration we produced 200 "predictor" curves from the distribution given in Figure 3. These curves were all generated by shifting and stretching a baseline function in the time axis and adding a random constant vertical shift. In addition we computed 200 responses by taking a linear combination of the time shift and stretch terms as well as
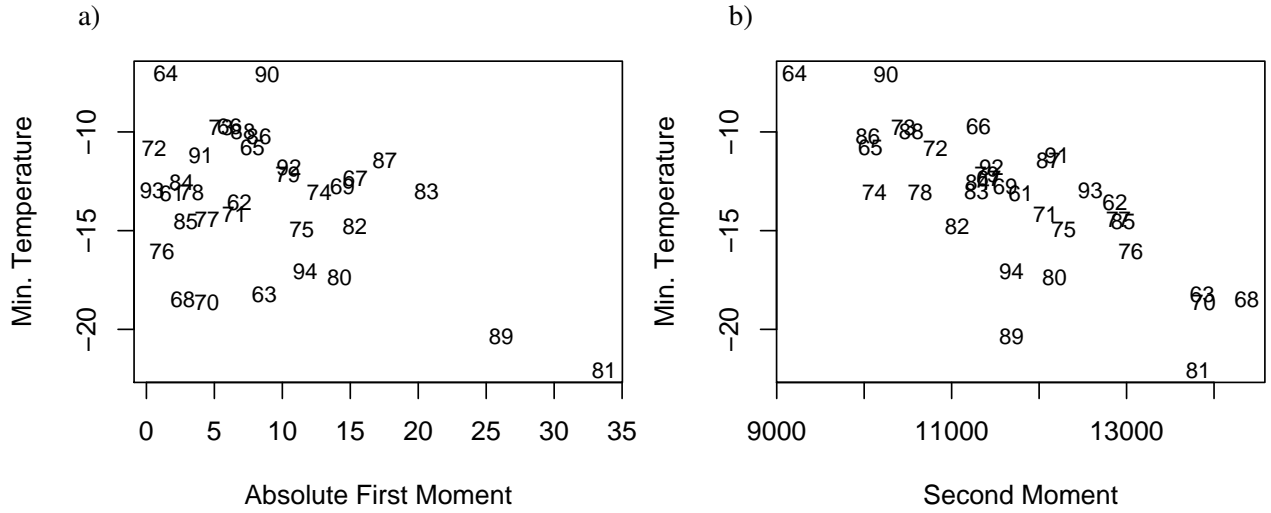
16

a)

b)

Figure 8: *Summaries of the yearly temperature curves in Montreal from 1961 to 1994.*

the vertical shift. Hence, the responses were linear functions of the warping and amplitude curves but not of the observed curves. We then compared the cross-validated predictions using only the observed curves with those using both the estimated warping and amplitude functions. The predictions from the observed curves were produced by computing the least squares fit of a b-spline basis to each curve and using the resulting coefficients as predictors in a linear regression. The predictions from the warping and amplitude curves were produced using the first two moments, as predictors for the warping function, and the b-spline coefficients from the synchronized curves, as predictors for the amplitude function. In both cases, degrees of freedom for the b-spine basis were chosen to provide the optimal fit. Using both the warping and amplitude curves produced a significantly improved fit with the root mean squared error between the responses and predictions only one fifth that from using the observed curves directly.

6 Discussion

In this article we develop a general moments based approach to the problem of synchronization of functional or curve data. The key idea is that one can use the moments of the Y_i's to accurately estimate the first few moments of the Z_i's, or equivalently of z, even in situations were the misalignment of the curves prevents one from using the cross-sectional average to directly estimate z. Hence, our approach has the advantage over target based methods that it can work well even when a target can not be accurately estimated. It also has an advantage over landmark based methods in that it can synchronize data with no, or inconsistent, marker events.

These methods could be extended in several directions. One obvious possibility would be to utilize more than one feature function. Let $I_g^1, \ldots, I_g^L$ represent L different feature functions and let $\mu_g^{(1,k)}, \ldots, \mu_g^{(L,k)}$ represent the corresponding kth moments. Then we could replace (20) by

$$Q_i = \sum_{l=1}^{L} \sum_{k=1}^{K_l} \lambda_{l,k} \left(\mu_{\tilde{Y}_i(X_i)}^{(l,k)} - \widehat{\mu_z^{(l,k)}} \right)^2 + \lambda_P \int (\tilde{Y}_i(X_i) - T(t))^2 dt + P(X_i). \tag{21}$$

This approach would allow one to take advantage of both local features in the curves, such as minimums and

17

maximums, as well as more global features, such as the absolute first derivative. When utilizing (21), CAMP can take full advantage of any landmarks that are present in the data while still maintaining the ability to synchronize the functions towards a common target function. In this sense CAMP possesses all the strengths of the landmark and Procrustes methods but eliminates most of their weaknesses. Another possible extension would be to higher dimensional functions. Although, in this article, we have only discussed one-dimensional curves, the moments approach could potentially be extended to multidimensional data. The definition of the feature function, I_g, could easily be expanded to such data and hence the moments also. Equating the lower order moments could then be achieved in a similar fashion to the one-dimensional case. The most significant challenge would seem to be dealing with higher order moments on high dimensional data were the number of cross product terms could become unmanageable.

A   Appendix

A.1   Proof of Theorem 3

First note that if $X_i(t) = \alpha_i + \beta_i t$ then $W_i(t) = \frac{t-\alpha_i}{\beta_i}$. Hence, by (7) and (12),

$$\mu^{(1)}_{Y_i^S(s)} = \mu^{(1)}_{\gamma_i + \theta_i z(W_i(s))} = \mu^{(1)}_{z\left(\frac{s-\alpha_i}{\beta_i}\right)} = \beta_i \mu^{(1)}_z + \alpha_i$$

and

$$\mu^{(k)}_{Y_i^S(s)} = \mu^{(k)}_{\gamma_i + \theta_i z(W_i(s))} = \mu^{(k)}_{z\left(\frac{s-\alpha_i}{\beta_i}\right)} = \beta_i^k \mu^{(k)}_z, \quad k \geq 2.$$

However, if $EX_i(t) = t$ for all t then $E\alpha_i = 0$ and $E\beta_i = 1$. Hence,

$$E\mu^{(1)}_{Y_i^S(s)} = \mu^{(1)}_z E\beta_i + E\alpha_i = \mu^{(1)}_z$$

and

$$\left(E\sqrt[k]{\mu^{(k)}_{Y_i^S(s)}}\right)^k = \left(\sqrt[k]{\mu^{(k)}_z} E\beta_i\right)^k = \mu^{(k)}_z$$

A.2   Proof of Theorem 4

First note that, under A-1 and A-2, for large enough n, $\sup_t \left| I_{\tilde{Y}_{i,n}}(t) - I_{Y_i^S}(t) \right| < \delta$ $a.s.$ where $\delta > 0$ can be made arbitrarily small. Hence, for large enough n,

$$\left| \mu^{(1)}_{\tilde{Y}_{i,n}} - \mu^{(1)}_{Y_i^S} \right| = \left| \int_0^T t(I_{\tilde{Y}_i}(t) - I_{Y_i^S}(t))dt \right| \leq \int_0^T t \left| I_{\tilde{Y}_i}(t) - I_{Y_i^S}(t) \right| dt \leq \delta \int_0^T t \, dt \quad a.s.$$

Hence $\mu^{(1)}_{\tilde{Y}_{i,n}} \to \mu^{(1)}_{Y_i^S}$ $a.s.$ and we can prove in a similar manner the same result for $\mu^{(k)}_{\tilde{Y}_{i,n}}, k \geq 2$. Therefore

$$\lim_{N,n \to \infty} \widehat{\mu^{(k)}_z} = \lim_{N \to \infty} \lim_{n \to \infty} \left(\frac{1}{N} \sum_{i=1}^N \sqrt[k]{\mu^{(k)}_{\tilde{Y}_i}} \right)^k = \lim_{N \to \infty} \left(\frac{1}{N} \sum_{i=1}^N \sqrt[k]{\mu^{(k)}_{Y_i^S}} \right)^k = \mu^{(k)}_z \quad a.s.$$

The last step follows from Theorem 3, the law of large numbers and Slutsky's Theorem.

A.3 Proof of Theorem 5

Equation (15) implies $\mu^{(1)}_{Y_i^S(X_i)} = \mu_z^{(1)}$ and $\mu^{(2)}_{Y_i^S(X_i)} = \mu_z^{(2)}$. Hence, by (6)

$$\frac{1}{\beta_i}\mu^{(1)}_{Y_i^S} - \frac{\alpha_i}{\beta_i} = \mu_z^{(1)} \quad \text{and} \quad \frac{1}{\beta_i^2}\mu^{(2)}_{Y_i^S} = \mu_z^{(2)}.$$

Solving these two equations for α_i and β_i gives the unique solution characterized by (16).

A.4 Proof of Theorem 6

By Theorem 4 and Slutsky's theorem $\hat{\alpha}_i$ and $\hat{\beta}_i$ respectively converge to α_i and β_i a.s. Hence $\hat{X}_i(t) \to X_i(t)$ a.s. To prove convergence of $\bar{Z}(t)$ first note that as $n \to \infty$, $\tilde{Y}_{i,n}$ converges to Y_i^S a.s. so we only need show convergence for $\frac{1}{N}\sum_i Y_i^S(\hat{X}_i(t)) = \frac{1}{N}\sum_i \gamma_i + \frac{1}{N}\sum_i \theta_i z(W_i(\hat{X}_i(t)))$. But

$$W_i(\hat{X}_i(t)) = \frac{\hat{\alpha}_i + \hat{\beta}_i t - \alpha_i}{\beta_i} = \mu_z^{(1)} + \left(t - \widehat{\mu_z^{(1)}}\right)\sqrt{\frac{\mu_z^{(2)}}{\widehat{\mu_z^{(2)}}}}$$

so $z(W_i(\hat{X}_i(t)))$ is constant as a function of i and converges to $z(t)$. Therefore

$$\bar{Z}(t) \to \frac{1}{N}\sum_i \gamma_i + z\left(\mu_z^{(1)} - \left(\widehat{\mu_z^{(1)}} - t\right)\sqrt{\frac{\mu_z^{(2)}}{\widehat{\mu_z^{(2)}}}}\right)\frac{1}{N}\sum_i \theta_i \to 0 + z(t)\cdot 1 = z(t) \quad a.s.$$

by the law of large numbers. Finally, note that for linear synchronization functions

$$\gamma_i = Y_i^S(t) - \theta_i z\left(\frac{t - \alpha_i}{\beta_i}\right) \quad \text{and} \quad \theta_i = \frac{Y_i^S(t) - \gamma_i}{z\left(\frac{t-\alpha_i}{\beta_i}\right)}$$

for all t. Therefore we can set up two simultaneous equations by setting t equal to t_j and then t_l with $j < l$. Solving these equations gives

$$\gamma_i = \frac{z\left(\frac{t_j-\alpha_i}{\beta_i}\right)Y_i^S(t_l) - z\left(\frac{t_l-\alpha_i}{\beta_i}\right)Y_i^S(t_j)}{z\left(\frac{t_j-\alpha_i}{\beta_i}\right) - z\left(\frac{t_l-\alpha_i}{\beta_i}\right)}.$$

But this holds for any t_j and t_l so we can reexpress γ_i and θ_i as

$$\gamma_i = \frac{2}{n(n-1)}\sum_{j=1}^{n-1}\sum_{l=j+1}^{n}\frac{z\left(\frac{t_j-\alpha_i}{\beta_i}\right)Y_i^S(t_l) - z\left(\frac{t_l-\alpha_i}{\beta_i}\right)Y_i^S(t_j)}{z\left(\frac{t_j-\alpha_i}{\beta_i}\right) - z\left(\frac{t_l-\alpha_i}{\beta_i}\right)}, \quad \theta_i = \frac{1}{n}\sum_{j=1}^{n}\frac{Y_i^S(t_j) - \gamma_i}{z\left(\frac{t_j-\alpha_i}{\beta_i}\right)}$$

Noting that Z_i^j converges to $z\left(\frac{t_j-\alpha_i}{\beta_i}\right)$ a.s and $\tilde{Y}_{i,n}$ converges to Y_i^S a.s. shows that $\widehat{\gamma}_i$ and $\widehat{\theta}_i$ also converge a.s.

References

Bar-Joseph, Z., Gerber, G. K., Gifford, D. K., Jaakkola, T. S., and Simon, I. (2003). Continuous representations of time-series gene expression data. *Journal of Computational Biology* **10**, 341–356.

Ferraty, F. and Vieu, P. (2002). The functional nonparametric model and applications to spectrometric data. *Computational Statistics* **17**, 545–564.

Ferraty, F. and Vieu, P. (2003). Curves discrimination: a nonparametric functional approach. *Computational Statistics and Data Analysis* **44**, 161–173.

Gasser, T. and Kneip, A. (1995). Searching for structure in curve samples. *Journal of the American Statistical Association* **90**, 1179–1188.

Gervini, D. and Gasser, T. (2004). Self-modeling warping functions. *Journal of the Royal Statistical Society, Series B* **66**, 959–971.

James, G. M. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society, Series B* **64**, 411–432.

James, G. M. (2007). Curve alignment by moments. *Annals of Applied Statistics* **1**, 480–501.

James, G. M. and Hastie, T. J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society, Series B* **63**, 533–550.

James, G. M., Hastie, T. J., and Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **87**, 587–602.

James, G. M. and Silverman, B. W. (2005). Functional adaptive model estimation. *Journal of the American Statistical Association* **100**, 565–576.

James, G. M. and Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association* **98**, 397–408.

Kneip, A. and Gasser, T. (1988). Convergence and consistency results for self-modeling nonlinear regression. *Annals of Statistics* **16**, 82–112.

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

Kneip, A., Li, X., MacGibbon, K. B., and Ramsay, J. O. (2000). Curve registration by local regression. *The Canadian Journal of Statistics* **28**, 1, 19–29.

Lin, D. Y. and Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association* **96**, 103–113.

Marx, B. D. and Eilers, P. H. (1999). Generalized linear regression on sampled signals and curves: A P-spline approach. *Technometrics* **41**, 1–13.

Priestley, M. B. and Chao, M. T. (1972). Non-parametric function fitting. *Journal of the Royal Statistical Society, Series B* **34**, 285–392.

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

Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis*. Springer.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis*. Springer.

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, Ser. B* **53**, 233–243.

Rice, J. A. and Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57**, 253–259.

Rønn, B. B. (2001). Nonparametric maximum likelihood estimation for shifted curves. *Journal of the Royal Statistical Society, Series B, Methodological* **63**, 2, 243–259.

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **26**, 43–49.

Silverman, B. W. (1984). Spline smoothing: The equivalent variable kernel method. *The Annals of Statistics* **12**, 898–916.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *R. Statist. Soc. B* **47**, 1–52.

Silverman, B. W. (1995). Incorporating parametric effects into functional principal components analysis. *Journal of the Royal Statistical Society, Sec. B* **57**, 673–689.

Tuddenham, R. D. and Snyder, M. M. (1954). Physical growth of California boys and girls from birth to eighteen years. *University of California Publications in Child Development* **1**, 183–364.

Wang, K. and Gasser, T. (1997). Alignment of curves by dynamic time warping. *The Annals of Statistics* **25**, 1251–1276.

Wang, K. and Gasser, T. (1999). Synchronizing sample curves nonparametrically. *The Annals of Statistics* **27**, 2, 439–460.

Zeger, S. L. and Diggle, P. J. (1994). Semiparametric models for longitudinal data with applications to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**, 689–699.