Alignment of functional data via self-modeling warping functions

Daniel Gervini* and Theo Gasser
Email: gervini@ifspm.unizh.ch
Department of Biostatistics, University of Zürich
Sumatrastrasse 30, 8006 Zürich, Switzerland

April 7, 2003

Abstract

This article introduces a new method of curve alignment, based on a semiparametric model for the warping functions. The warping functions are assumed to be linear combinations of q basis functions or components, which are themselves estimated from the data (hence the name "self-modeling warping functions"). With small values of q it is possible to achieve a warping flexibility comparable to that of fully nonparametric methods. This prevents self-modeling registration from overfitting time variability, which is a common problem of nonparametric procedures. As a convenient by-product, the q individual scores are often easily interpretable parameters that can be used for statistical inference.

KEY WORDS: Curve registration; Functional data analysis; Semiparametric models; Time warping.

*Supported by the Swiss National Science Foundation grant BE 20.63579.00.

1 Introduction

Data sets consisting of samples of functions have become increasingly common in recent years. Functional data arise in such diverse areas as medicine, economy, geology, meteorology, image analysis, etc. An excellent overview of many real-life applications of functional data analysis is Ramsay and Silverman (2002), and a mathematically more rigorous treatment of the topic is Ramsay and Silverman (1997). Some problems and aims of functional data analysis are similar to those of other areas of statistics (multivariate analysis in particular), although they present some peculiarities. Consider, for instance, the problem of estimating the mean function $\mu(t)$ from a sample of univariate functions $\{x_i(t)\}$. The naive estimator of $\mu(t)$, the cross-sectional sample mean $\bar{x}(t)$, will not always produce sensible results. This is illustrated in Fig. 1, where five leg growth velocity curves and their cross-sectional mean are given. A salient feature of these curves from a biomedical point of view is the pubertal spurt, which is the peak that occurs around age 14 for boys. The peak varies not only in intensity but also in timing. As a result, the pubertal spurt estimated by the cross-sectional mean is much less pronounced than any of the peaks in the sample. Why is this so? If τ_i denotes the time of maximal pubertal growth of person i, and τ_0 the time of maximal pubertal growth of $\bar{x}(t)$, then $x_i(\tau_0) < x_i(\tau_i)$ by definition, which implies that $\bar{x}(\tau_0) < \overline{x.(\tau.)}$. Thus, the cross-sectional mean *always* underestimates the mean amplitude of maximal pubertal growth (or any other local maximum, for that matter). This is a serious problem, since local extrema are often important features of the physical or biological process under examination, and it is precisely at those points where the statistician needs good estimates.

The bias of the sample mean at local extrema is due to the variability of the timings of these extrema. This problem is the rule and not the exception in functional data analysis. Ignoring time variability not only has an adverse effect on the estimation of the mean, but also on the estimation of the covariance and on methods derived from it, such as functional principal components. Early work on the area, which essentially consisted in straightforward adaptations of multivariate techniques (e.g. Rao, 1958), overlooked this problem completely. The specific effect of random time shifts on principal components was first characterized by Möcks (1986): in addition to the "real" components that explain amplitude variability (i.e. components that would appear in the decomposition even if there

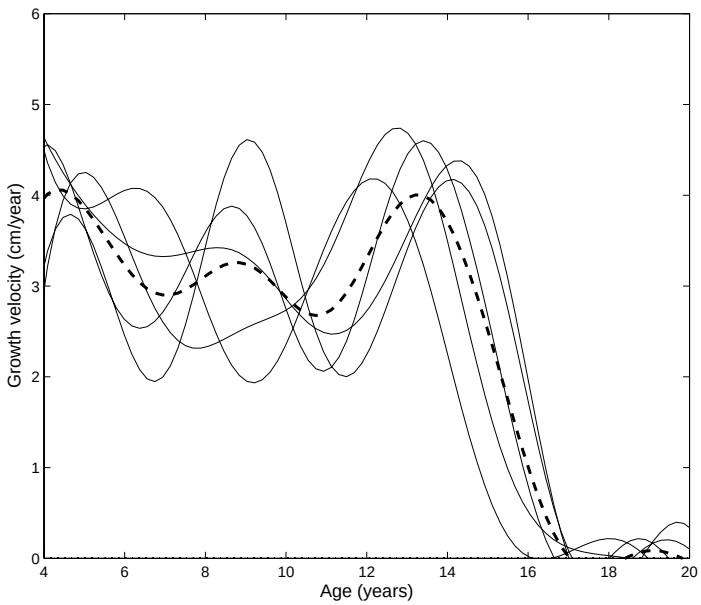

Figure 1: *Leg growth velocity curves of five boys (solid line) and cross-sectional mean (dashed line).*

was no time variability), the derivatives of some components also appear in the decomposition. This spurious inflation of the number of principal components leads to interpretability problems and misalocation of variance. Chapters 6 and 7 of Ramsay and Silverman (2002) show several examples of adequate treatments of time variability that have an important impact on the subsequent analysis. Therefore, although this paper will focus on estimation of the mean function, the reader should keep in mind that the benefits of properly handling time variability extend to other realms of functional data analysis, notably functional principal components.

To formalize the discussion, consider a sample of curves $x_i : \mathbb{R} \to \mathbb{R}$ that follow the model

$$x_i(t) = \mu(v_i(t)) + \varepsilon_i(t),\ t \in T \subset \mathbb{R}.$$

The non-random function μ is called "structural mean", $\{\varepsilon_i(t)\}$ are stochastic processes that account for amplitude variability, and $\{v_i(t)\}$ are stochastic processes that account for time variability. Assuming $\mathbb{E}\{\varepsilon_i(t)\} = 0$ for all $t \in T$, we have that $\bar{x}(t) \to \mathbb{E}\{\mu(v(t))\}$, which in general is different from $\mu(t)$. Therefore, the cross-sectional mean is not a consistent estimator of the structural mean. But there is a "simple" remedy for this. If the processes $\{v_i(t)\}$ were observable,

a consistent estimator of μ would be the average of the so-called registered or aligned functions $x_i^\star = x_i(v_i^{-1}(t))$. In practice, of course, the v_i's are not observable. A number of methods to estimate the v_i's and μ have been proposed in the literature. They are: landmark registration (Gasser et al., 1990, and Kneip and Gasser, 1992), continuous monotone registration (Ramsay and Li, 1998), dynamic time warping (Wang and Gasser, 1999) and local regression (Kneip et al., 2000).

Landmark registration, when feasible, is probably the best method because it gets to the heart of the problem. The method involves identifying the timings of specific features of the curves (such as maximal pubertal growth, in the example given above), and then aligning the curves so that all these events occur at the same time for each curve. In the example of Fig. 1, the so-called warping functions $w_i(t) = v_i^{-1}(t)$ would be constructed so that $w_i(\bar{\tau}) = \tau_i$, which implies that $\overline{x^\star}(\bar{\tau}) = \overline{x_\cdot(\tau_\cdot)}$. This solves the problem of underestimation of the pubertal peak. Unfortunately, landmark identification may be a difficult task. For instance, the analysis of growth data in Gasser et al. (1991) uses eight landmarks: the four zero-crossings and the four local extrema of a typical growth acceleration curve. Individual acceleration curves, however, show a variable number of zero crossings, depending on the quality of smoothing and individual characteristics. The eight landmarks have to be carefully determined for each of the 232 curves of the sample, which is a tedious, time consuming and error-prone task.

Continuous monotone registration, on the other hand, is a fully nonparametric method that does not require landmark identification. The warping functions are modeled using B-spline basis functions, with the individual coefficients found by minimization of a lack-of-fit measure. This allows great flexibility for the warping functions, but makes the method susceptible to "time variability overfitting": that is, attempting to eliminate variability that is not of time nature (see Ramsay and Li, 1998, for an example). The solution proposed by the authors is to minimize an ad-hoc measure of lack of fit instead of the more natural integrated squared error. However, the main reason behind time variability overfitting is that warping flexibility can only be attained by using a relatively large number of basis functions. This is necessarily accompanied by some degree of overfitting, which can be ameliorated by adding a roughness penalty term to the objective function, but this brings on the problem of choosing the penalty parameter and

considerably increases computing time.

What we propose in this article is a nonparametric method of registration that does not involve landmark identification, offers considerable warping flexibility and avoids the problem of overfitting. The idea is to model the warping functions as linear combinations of a small number of simple functions: $w_i(t) = t + \sum_{j=1}^{q} s_{ij}\phi_j(t)$. The component functions $\{\phi_j\}$ are estimated from the data, hence the name "self-modeling" warping functions. They are linear combinations of B-spline basis, but the coefficients are estimated using the combined data set, since the components are common to all subjects. Only q individual-specific scores need to be estimated (for the growth example, $q = 3$ or 4 suffices). This approach, then, avoids the problem of overfitting and, at the same time, offers enough flexibility for the warping functions to capture the most important modes of time variation.

Although the "self-modeling" decomposition of the warping functions is reminiscent of a functional principal component decomposition, we will impose certain restrictions on the components $\{\phi_j\}$, which may seem capricious without the proper motivation. Therefore, the rationale behind the "self-modeling" method is given in the next section, and the method itself is formally introduced in section 3.

2 Motivation

The kind of component functions $\{\phi_j\}$ that we are looking for are, essentially, non-negative bell-shaped curves. This is motivated by landmark registration. For simplicity, consider the situation where we have only two landmarks per curve, τ_{i1} and τ_{i2}, on an interval $T = [a, b]$. Let $\tau_{01} = \bar{\tau}_{\cdot 1}$ and $\tau_{02} = \bar{\tau}_{\cdot 2}$ be the average landmarks. The simplest landmark registration method defines warping functions such that $w_i(a) = a$, $w_i(\tau_{01}) = \tau_{i1}$, $w_i(\tau_{02}) = \tau_{i2}$, $w_i(b) = b$, and linearly interpolates in between. Then, the centered warping function $w_i(t) - t$ is

$$w_i(t) - t = \begin{cases} (\tau_{i1} - \tau_{01})\left(\frac{t-a}{\tau_{01}-a}\right) & a \le t \le \tau_{01} \\ (\tau_{i1} - \tau_{01})\left(\frac{\tau_{02}-t}{\tau_{02}-\tau_{01}}\right) + (\tau_{i2} - \tau_{02})\left(\frac{t-\tau_{01}}{\tau_{02}-\tau_{01}}\right) & \tau_{01} \le t \le \tau_{02} \\ (\tau_{i2} - \tau_{02})\left(\frac{b-t}{b-\tau_{02}}\right) & \tau_{02} \le t \le b. \end{cases}$$

We can write $w_i(t) - t = s_{i1}\phi_1(t) + s_{i2}\phi_2(t)$, where $s_{ij} = (\tau_{ij} - \tau_{0j})$ and

$$\phi_1(t) = \begin{cases} \frac{t-a}{\tau_{01}-a} & a \leq t \leq \tau_{01} \\ \frac{\tau_{02}-t}{\tau_{02}-\tau_{01}} & \tau_{01} \leq t \leq \tau_{02} \\ 0 & \tau_{02} \leq t \leq b \end{cases}, \ \phi_2(t) = \begin{cases} 0 & a \leq t \leq \tau_{01} \\ \frac{t-\tau_{01}}{\tau_{02}-\tau_{01}} & \tau_{01} \leq t \leq \tau_{02} \\ \frac{b-t}{b-\tau_{02}} & \tau_{02} \leq t \leq b. \end{cases}$$

These functions are triangles peaking at the mean landmarks τ_{01} and τ_{02}. Each coefficient s_{ij} is the deviation of the landmark τ_{ij} with respect to τ_{0j}. This is a decomposition of warping functions as combinations of peculiar basis functions $\{\phi_j\}$, each of which is associated with a landmark. Note that if $B(t) = [\beta_1(t), \ldots, \beta_4(t)]^\top$ denotes the vector of the four B-spline basis functions of order 2 (piecewise linear) with knots $\{a, \tau_{01}, \tau_{02}, b\}$, then $\phi_j(t) = c_j^\top B(t)$, with $c_1 = (0, 1, 0, 0)^\top$ and $c_2 = (0, 0, 1, 0)^\top$. Therefore, although ϕ_1 and ϕ_2 are not themselves orthogonal, their coefficients on the B-spline basis are.

Roughly speaking, our proposal will attempt to back-engineer this decomposition. The self-modeling components will be restricted to linear combinations of B-spline basis functions with non-negative and orthogonal coefficients. This will produce approximately bell-shaped functions with localized support (as in Fig. 6), which can be thought of as associated with "hidden landmarks". Consequently, the scores will correlate with time variation at specific parts of the domain, which makes them easily interpretable and suitable for statistical inference, as the example of section 5 will show.

3 Self-modeling registration

3.1 The model

Let $\{x_i : i = 1, \ldots, n\}$ be a sample of functions $x_i : T \subset \mathbb{R} \to \mathbb{R}$ following the model

$$x_i(t) = a_i \mu(v_i(t)) + \varepsilon_i(t), \ t \in T, \tag{1}$$

where:

1. $\mu : T \to \mathbb{R}$ is a differentiable function.

2. $\{\varepsilon_i\}$ are i.i.d. stochastic processes with $\mathbb{E}\{\varepsilon(t)\} = 0$ for all $t \in T$.

3. $\{a_i\}$ are i.i.d. with $\mathbb{E}(a) = 1$.

4. $\{v_i\}$ are i.i.d. monotone increasing stochastic processes, $v_i : T \to T$, such that $\mathbb{E}\{v^{-1}(t)\} = t$ for all $t \in T$.

5. a, v and ε are stochastically independent.

This non-linear shape invariant model reasonably accounts for both amplitude and time variability, and is identifiable (see Appendix A.1 for a proof).

Note that we can rewrite (1) as $x_i(t) = \mu(v_i(t)) + b_i\eta(v_i(t)) + \varepsilon_i(t)$, with $b_i = a_i - 1$ and $\eta = \mu$. This is a one-component model for amplitude variability, where the component η is forced to coincide with the mean function. It is a simple model that works well for the example given later. A more sophisticated alternative is to consider an arbitrary η, and perhaps several components, but we will not pursue this avenue.

Regarding the warping functions $w_i = v_i^{-1}$, the model we propose is

$$w_i(t) = t + \sum_{j=1}^{q} s_{ij}\phi_j(t), \ t \in T, \tag{2}$$

where:

1. $\mathbf{s}_i = (s_{i1}, \ldots, s_{iq})^\top$ are i.i.d. such that $\mathbb{E}(\mathbf{s}) = 0$.

2. $\{\phi_j\}$ are nonrandom functions. Given a B-spline basis $\beta_1(t), \ldots, \beta_p(t)$, we assume that $\phi_j(t) = \sum_{k=1}^{p} c_{jk}\beta_k(t)$, with coefficients satisfying the following conditions:

   (a) $c_{jk} \geq 0$ for all j, k, and $\sum_{k=1}^{p} c_{jk}^2 = 1$ for all j.

   (b) The coefficient matrix $C = (c_{jk}) \in \mathbb{R}^{q\times p}$ has a block structure: there is a sequence of "block delimiters" $1 \leq K_1 < K_2 < \ldots < K_{q+1} \leq p+1$ such that $c_{jk} > 0$ for $K_j \leq k < K_{j+1}$ and $c_{jk} = 0$ for $k < K_j$ and $k \geq K_{j+1}$.

   (c) $c_{j1} = c_{jp} = 0$ for all j (in other words, $K_1 = 2$ and $K_{q+1} = p$, so that $K_2 \geq 3$ and $K_q \leq p - 1$).

Restriction 2(b) imposes a block structure on the coefficients, which guarantees that the components have connected supports. This rules out, for example, the case of a component consisting of two separate "bells" intersected by a second component. This restriction is necessary not only for visual satisfaction but, more importantly, for identifiability of the model (which is proved in Appendix A.2). Restriction 2(c) guarantees that the warping functions are "tied down" on the extremes of T. In what follows we will denote by $\mathcal{C}$ the space of coefficient matrices satisfying the restrictions 2(a)-2(c).

Model (2) offers great flexibility for the warping functions, and the number of parameters to be estimated is not too large. The actual dimension of $\mathcal{C}$ is not pq but only $p-q-2$, due to the restrictions. Since the coefficient matrix is common to all subjects, the number of basis functions p is not limited by the number of observations per individual. Only the number of components q is limited in this way, because the scores must be specifically estimated for each individual. But the number of components that are necessary for an adequate fit is typically small. This represents a big improvement over continuous monotone registration, which entails estimation of p coefficients per individual. The difference is most noticeable when the number of observations per subject is not too large. In the growth example of section 5 we have 30 observations per person, which limits the number of parameters that can be accurately estimated to 5 or 6 per individual. This is very restrictive for continuous monotone registration but not for self-modeling registration (we use $q \leq 4$). In situations like this, monotone registration must resort to data smoothing which enables the use of a larger number of basis functions, while self-modeling registration can still be safely applied to the raw data, avoiding smoothing bias.

3.2 Estimation

The parameters of models (1) and (2) can be estimated by integrated least squares, minimizing

$$F(\mu, \{a_i\}, C, \{\mathbf{s}_i\}) = \sum_{i=1}^{n} \int_a^b \{x_i(t) - a_i \mu(v_i(t))\}^2 dt. \tag{3}$$

In practice we only have discrete observations of $\{x_i(t)\}$ on a grid $t_1, \ldots, t_m$, so that (3) must be substituted with a discrete version:

$$F^\star(\mu, \{a_i\}, C, \{\mathbf{s}_i\}) = \sum_{i=1}^{n} \sum_{j=1}^{m} \{x_i(t_j) - a_i \mu(v_i(t_j))\}^2 (t_j^\star - t_{j-1}^\star), \tag{4}$$

where $t_0^\star = t_1$, $t_j^\star = (t_j + t_{j-1})/2$ for $j = 2, \ldots, m-1$, and $t_m^\star = t_m$. If the input grid is dense and the data is too noisy, it is advisable to first smooth the data and then minimize (4) using smoothed values $\hat{x}_i(t_j)$.

The minimizers $\hat{\mu}$ and $\{\hat{a}_i\}$ of (3) for given $\{w_i\}$, though, have closed expressions. Since

$$F(\mu, \{a_i\}, C, \{\mathbf{s}_i\}) = \int_a^b \sum_{i=1}^{n} \left(x_i(w_i(t)) - a_i \mu(t)\right)^2 w_i'(t) dt,$$

the integrand (and then F) is minimized by

$$\hat{\mu}(t) = \frac{\sum_{i=1}^{n} \hat{a}_i \hat{w}_i'(t) \hat{x}_i^\star(t)}{\sum_{i=1}^{n} \hat{a}_i^2 \hat{w}_i'(t)}, \tag{5}$$

where $\hat{x}_i^\star(t) = x_i(\hat{w}(t_i))$. The unconstrained minimizer of (4) with respect to a_i is

$$\tilde{a}_i = \frac{\sum_{j=1}^{m} x_i(t_j) \hat{\mu}(\hat{v}_i(t_j))(t_j^\star - t_{j-1}^\star)}{\sum_{j=1}^{m} \hat{\mu}^2(\hat{v}_i(t_j))(t_j^\star - t_{j-1}^\star)}. \tag{6}$$

We need that $\bar{a} = 1$ for identifiability, so that a good approximation to the constrained minimizer $\hat{a}_i$ of (4), although not precisely the one obtained from the Lagrangian, is

$$\hat{a}_i = 1 + \tilde{a}_i - \bar{\tilde{a}}. \tag{7}$$

Minimization of (4) with respect to C and $\mathbf{s}_i$ for given μ and $\{a_i\}$ requires an iterative algorithm.

We suggest the following two-stage algorithm for minimization of (4):

1. Initialization.

   (a) Select "promising" block delimiters $K_1, \ldots, K_{q+1}$ for the coefficient matrix (see comments below). To satisfy the restrictions of $\mathcal{C}$, use the

reparameterization

$$(c_{j,K_j}, \ldots, c_{j,K_{j+1}-1}) = (1, \exp(\mathbf{y}_j))/\{1 + \|\exp(\mathbf{y}_j)\|^2\}^{\frac{1}{2}},$$

where each $\mathbf{y}_j \in \mathbb{R}^{K_{j+1}-K_j}$ is an unconstrained vector.

(b) Set $\hat{\mathbf{y}}_j = 0$, $\hat{\mathbf{s}}_i = 0$, $\hat{a}_i = 1$ and $\hat{\mu}(t) = \bar{x}(t)$.

2. Iterations.

(a) **Update warping functions.**

i. Update the coefficient matrix $\hat{C}$ using a Newton-Raphson step for each $\mathbf{y}_j$.

ii. Update each score $\hat{\mathbf{s}}_i$ using a Newton-Raphson step, making sure that the resulting warping function $\hat{w}_i$ is strictly increasing (halve the Newton-Raphson step until this is satisfied).

(b) **Update mean and scaling factors.**

i. Update $\hat{\mu}$ using (5).

ii. Update $\hat{a}_i$ using (7), computing $\hat{v}_i(t_j)$ and $x_i(\hat{w}_i(t_j))$ by linear interpolation.

iii. Update objective function (4).

(c) **Stopping rule.** Exit if there is no significant improvement of the objective function at stage 2(b)iii. Otherwise go back to (a).

Selecting adequate block delimiters at the initialization stage of the algorithm is crucial. Since there are $\binom{p-3}{q-1}$ possible sets of delimiters, an exploration of all of them is not feasible in most cases. A workable alternative is to generate a random subsample of the possible delimiters, do 2 or 3 steps of the algorithm for each of them, keep the delimiters that yield the lowest objective function (or the 3 lowest, say), and iterate with these until convergence. Normally it is not necessary to use a very large number of random starts because the shape of the components do not change much when delimiters are shifted one unit to the right or to the left. There are other little details of implementation that need some attention but are too technical to describe here. A Matlab routine implementing this algorithm is available on the first author's webpage.

3.3 Choosing the number of components

The algorithm described above estimates the parameters of models (1) and (2) for given p and q, which in turn need to be chosen in some way. We propose to select p and q that minimize the expected prediction error $\sigma_e^2 = \mathbb{E}\{\int (x - a\mu(v))^2\}$. This quantity needs to be estimated, which presents some difficulties. A simple estimator would be $F^\star(\hat{\mu}, \{\hat{a}_i\}, \hat{C}, \{\hat{\mathbf{s}}_i\})/n$, but this underestimates σ_e^2 because the same data is used for estimation of model parameters and prediction errors. What we propose instead is a simplified cross-validation procedure. Cross-validation avoids underestimation of σ_e^2 by using different subsamples for parameter estimation ("training set") and for prediction ("test set"). A general treatment of cross-validation can be found in Hastie el al. (2001).

For our estimation problem, a cross-validation procedure presents two difficulties. First, the computing time; it would take too long to actually re-estimate all the parameters for each training set. Second, the quantities a_i and $v_i(t)$ are not observable predictors; they must be estimated from $x_i(t)$, so that the estimated error $\hat{e}_i = \int (x_i - \hat{a}_i \hat{\mu}(\hat{v}_i))^2$ would be using x_i both for estimation and for prediction, violating the fundamental principle of cross-validation. These problems can be avoided by estimating the a_i's and the v_i's on a first step and treating them as fixed predictors of x_i at the cross-validation steps. Then only μ needs to be re-estimated for each training set, which is done using the non-iterative weighted average (5).

Specifically, the cross-validation procedure we propose goes as follows. Let Ω be a set of tentative parameters (p, q). Then:

1. Split the sample into N roughly equal-sized subsamples ("test sets"). Let I_k be the indices of the observations in the k-th subsample.

2. For each $(p, q) \in \Omega$:

   (a) Find estimates $\{\hat{a}_i\}$ and $\{\hat{w}_i(t)\}$ (and consequently $\{\hat{v}_i(t)\}$ and $\{\hat{x}_i^\star(t)\}$) using the whole dataset.

   (b) For each $k = 1, \ldots, N$:

      i. Estimate μ using data not included in the test set:

$$\hat{\mu}^{(-k)}(t) = \frac{\sum_{i \notin I_k} \hat{a}_i \hat{x}_i^\star(t) \hat{w}_i'(t)}{\sum_{i \notin I_k} \hat{a}_i^2 \hat{w}_i'(t)}.$$

ii. For each $i \in I_k$ estimate the prediction error:

$$\hat{e}_i = \sum_{j=1}^{m} \{x_i(t_j) - \hat{a}_i \hat{\mu}^{(-k)}(\hat{v}_i(t_j))\}^2 (t_j^\star - t_{j-1}^\star).$$

(c) $\mathrm{CVE}(p,q) = \sum_{i=1}^{n} \hat{e}_i / n$ is the cross-validation estimate of σ_e^2 for a model with parameters (p,q).

3. Select $(p,q) \in \Omega$ that minimizes $\mathrm{CVE}(p,q)$.

Possible values of N are $N = n$ (the familiar leave-one-out cross-validation), $N = 5$ and $N = 10$. As [3] explain, the leave-one-out cross-validation estimator is nearly unbiased but very variable. They recommend five-fold or ten-fold cross-validation, which tend to overestimate σ_e^2 but are less variable.

The size of Ω is normally not too large. In practice, the number of salient features of the curves (such as peaks and valleys) will give a hint about the number of components q that may provide a good fit. On the other hand, the number of observations per person imposes a bound on q (the rule of thumb says that $q < \sqrt{m}$). Therefore, it will be hardly necessary to explore too many values of q. The number of basis functions p will in turn depend on q: large values of p tend to produce components with irregular shapes, so we suggest using $p = 3q$ or $p = 4q$ basis functions, which provide just enough warping flexibility.

4 Simulations

4.1 Assessing model fit

We carried out some simulations to evaluate the performance of self-modeling registration as compared to landmark registration. Samples $\{x_i(t_j)\}$, $i = 1, \ldots, n$, $j = 1, \ldots, m$, were generated following model (1) with $\mu(t) = \sin(2\pi t)$, $T = [0,1]$, and $t_j = (j-1)h$, where $h = 1/(m-1)$. The scaled sinus function completes a cycle between 0 and 1, with a peak at $\tau_{01} = .25$ and a valley at $\tau_{02} = .75$, providing two natural landmarks. The data was simulated as follows:

1. Two random landmarks per curve were generated: $\tau_{ik} = \tau_{0k} + \zeta_{ik}$, $k = 1, 2$, with $\zeta_{ik} = \min(\max(Z_{ik}/12, -.24), .24)$ and Z_{ik} independent $N(0,1)$ random variables (the truncation guarantees that $0 < \tau_{i1} < \tau_{i2} < 1$).

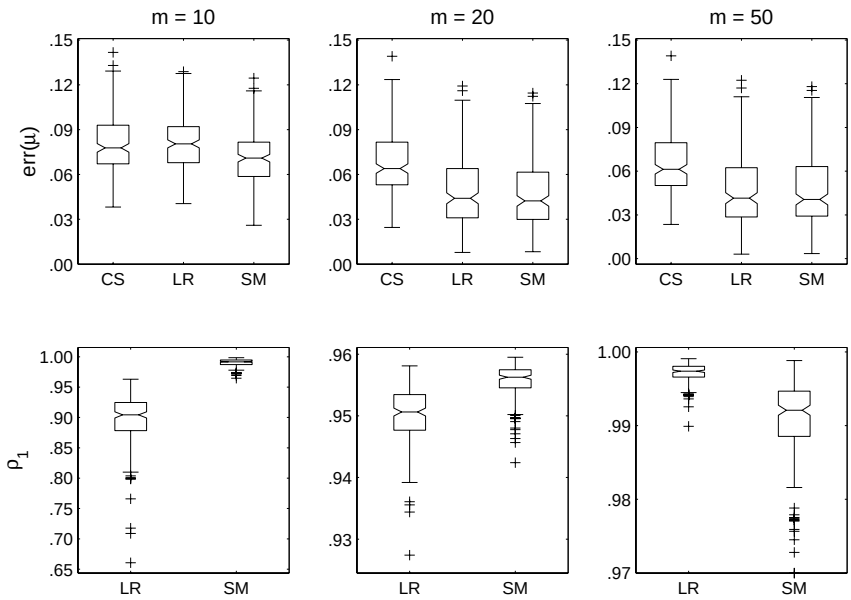

Figure 2: *Simulated* $\mathrm{err}(\hat{\mu})$ *(first row) and* ρ_1 *(second row) for cross-sectional mean (CS), landmark registration (LR) and self-modeling registration (SM). Sample size* $n = 20$*, model without error term.*

2. The inverse warping functions $v_i = w_i^{-1}$ where defined as $v_i(0) = 0$, $v_i(\tau_{i1}) = \tau_{01}$, $v_i(\tau_{i2}) = \tau_{02}$, $v_i(1) = 1$, and linearly interpolated at the grid points $\{t_j\}$.

3. The scaling factors $\{a_i\}$ were i.i.d. $N(1, \sigma_a^2)$ random variables with $\sigma_a = .10$.

4. The observations were defined as $x_i(t_j) = a_i\mu(v_i(t_j)) + \varepsilon_i(t_j)$, where $\{\varepsilon_i(t_j)\}$ were i.i.d $N(0, \sigma_e^2)$ random variables. Two models were considered: $\sigma_e^2 = 0$ (no error term) and $\sigma_e^2 = 1/10\sqrt{2}$ (signal-to-noise error $\sigma_e^2/(\sigma_a^2 \int \mu^2 + \sigma_e^2) = .5$).

5. Sample sizes $n = 10, 20, 50$ and grid sizes $m = 10, 20, 50$ were used. The 18 sampling situations were replicated 200 times each.

To estimate μ, we used three methods:

- No registration: cross-sectional mean.

- Landmark registration: for each curve, the landmarks were estimated as $\hat{\tau}_{i1} = \arg\max_{t_j}\{x_i(t_j)\}$ and $\hat{\tau}_{i2} = \arg\min_{t_j}\{x_i(t_j)\}$.

Table 1: *Simulation results for model (1) without error term (CS:cross-sectional mean; LR: landmark registration; SM: self-modeling registration).*

		$m = 10$		$m = 20$		$m = 50$	
n	Estim	$\mathrm{e}(\hat{\mu})\dagger$	ρ_1	$\mathrm{e}(\hat{\mu})$	ρ_1	$\mathrm{e}(\hat{\mu})$	$\rho_1\ddagger$
10	CS	.89	—	.79	—	.77	—
	LR	.90	.89	.66	.98	.66	1.0
	SM	.81	.99	.65	.99	.66	.99
20	CS	.80	—	.68	—	.65	—
	LR	.80	.90	.49	.98	.47	1.0
	SM	.71	.99	.47	.99	.47	.99
50	CS	.72	—	.57	—	.54	—
	LR	.73	.90	.31	.98	.28	1.0
	SM	.64	.99	.30	.99	.28	.99

$\dagger \mathrm{e}(\hat{\mu}) = \mathrm{err}(\hat{\mu}) \times 10$.
‡Values appearing as 1.0 have been rounded up.

- Self-modeling registration, with $q = 2$ components and $p = 6$ B-spline basis functions of order 3 with equispaced knots. All possible sets of delimiters, $\binom{p-3}{q-1} = 3$ in this case, were taken as random starts.

As a measure of estimation error we used $\mathrm{err}(\hat{\mu}) = \{\sum_{j=1}^{m}(\hat{\mu}(t_j) - \mu(t_j))^2 \cdot h\}^{\frac{1}{2}}$. In addition to estimating μ, it is of interest to investigate the correlations between self-modeling scores and landmarks, to confirm our claim that self-modeling components associate themselves with the underlying landmarks. For this purpose we will compare the correlations $\rho_1^{\mathrm{SM}} = \mathrm{corr}(\{\hat{s}_{i1}\}, \{\tau_{i1}\})$ and $\rho_1^{\mathrm{LR}} = \mathrm{corr}(\{\hat{\tau}_{i1}\}, \{\tau_{i1}\})$ (we computed similar correlations for the second landmark but the results were almost identical, so they are not given).

Tables 1 and 2 give mean simulated values of $\mathrm{err}(\hat{\mu}) \times 10$ and ρ_1. Monte Carlo standard errors are not given, but a clear picture of the significance of the results emerges from Fig. 2, which exemplifies the typical behavior of the simulated quantities and their variability. In general, we see that the registered means are significantly better than the cross-sectional mean for $m = 20$ and $m = 50$. For these values of m both registration methods are comparable. For $m = 10$ self-modeling registration estimates μ better than the cross-sectional mean, but

Table 2: *Simulation results for model (1) with error term (CS:cross-sectional mean; LR: landmark registration; SM: self-modeling registration).*

		$m = 10$		$m = 20$		$m = 50$	
n	Estim	$\mathrm{e}(\hat{\mu})\dagger$	ρ_1	$\mathrm{e}(\hat{\mu})$	ρ_1	$\mathrm{e}(\hat{\mu})$	$\rho_1\ddagger$
10	CS	.93	—	.84	—	.82	—
	LR	.93	.89	.70	.98	.69	1.0
	SM	.88	.97	.69	.97	.69	.97
20	CS	.78	—	.66	—	.63	—
	LR	.80	.90	.48	.98	.47	1.0
	SM	.72	.98	.47	.98	.47	.98
50	CS	.72	—	.57	—	.54	—
	LR	.75	.90	.34	.98	.31	1.0
	SM	.66	.98	.33	.98	.31	.98

$\dagger \mathrm{e}(\hat{\mu}) = \mathrm{err}(\hat{\mu}) \times 10$.
‡Values appearing as 1.0 have been rounded up.

landmark registration does not. Remarkably, self-modeling registration is not outperformed in any of the 18 scenarios.

Regarding the correlations ρ_1^{SM}, they turn out to be surprisingly high and stable for all m and n. For landmark registration the correlations increase with m, as expected, since the accuracy of landmark estimation depends solely on m. Self-modeling registration, on the other hand, borrows strength from the whole dataset to estimate the components, so that for $m = 10$ the scores are more strongly correlated with the underlying landmarks than the landmark estimators themselves.

To summarize, this small simulation study has shown that self-modeling registration is able to provide estimates of the structural mean comparable to those of landmark registration, and component scores that are highly correlated with the underlying landmarks. The scope of these simulations is rather limited, but the real example analyzed in section 5 confirms these conclusions.

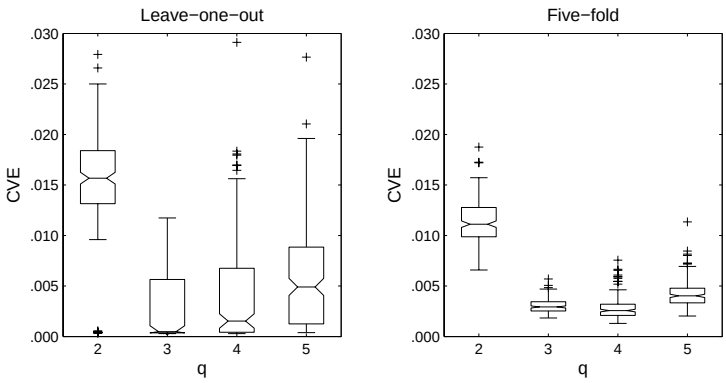

Figure 3: *Simulated cross-validation estimates of prediction errors of self-modeling registration for four different models (actual model: $q = 3$).*

4.2 Assessing model choice

We ran another simulation study to evaluate the cross-validation method proposed in section 3.3. The model from which the data was generated was similar to that of the preceding subsection, only that we took $T = [0, 1.5]$ and added a third landmark at $t = 1.25$ (the second peak of μ). Thus, the "true" number of components is $q = 3$. We took $n = 50$ and $m = 30$, and a model without error term (roughly equivalent to using smoothed data). Each sampling situation was replicated 200 times, and 50 random starts for the self-modeling algorithm were used in all cases. The set of tentative parameters considered was $\Omega = \{(p, q) : p = 3q,\ q = 2, 3, 4, 5\}$.

Fig. 3 shows the simulated cross-validation error estimators for each q. As expected, leave-one-out cross-validation turns out to be less biased but much more variable than five-fold cross-validation. The former undoubtedly picks out the right model, on average. The latter, on the other hand, is minimized on average by the slightly overparameterized model $q = 4$, although the difference with $q = 3$ is very small. This is due to the tendency of five-fold cross-validation to overestimate the prediction error and select overparameterized models (see comments on p. 215 of Hastie et al., 2001).

A more detailed analysis reveals the following: leave-one-out cross-validation selects the right parameter $q = 3$ for 111 of the 200 samples (55.5%), $q = 4$ for 65 samples (32.5%), $q = 5$ for 18 samples (9%) and $q = 2$ for 6 samples (3%). On the other hand, five-fold cross-validation selects $q = 4$ for 69% of the samples, $q = 3$ for 26%, $q = 5$ for 5%, and never selects $q = 2$. It is clear that five-fold

cross-validation is more stable than the leave-one-out rule, although it selects the overparameterized model $q = 4$ in the majority of cases. Combining the figures of the 3-component model and the second best 4-component model, we have that the leave-one-out cross-validation selects either of these models 88% of the time, while five-fold cross-validation does so 95% of the time. All things considered, five-fold cross-validation appears to be a good guide for model selection, provided one keeps in mind that it tends to select slightly overparameterized models.

5 Example: leg growth velocity

The First Zurich Longitudinal Growth Study has been analyzed many times in the literature, so it offers an excellent test set for the new registration method. In Gasser et al. (1991) the authors find mean velocity curves using landmark registration. They use eight landmarks: the four zero-crossings and the four local extrema of a typical acceleration curve. Data smoothing is necessary to find the landmarks, but the registration itself and subsequent analysis is performed on the raw velocity curves (i.e., first derivatives computed with finite differences). Using raw velocities is not recommendable if the sampling rate is high and the signal-to-noise ratio is small, but this is not the problem in this growth study.

Here we analyze leg growth velocity because these curves present a clear-cut midgrowth spurt (the peak around age 7) besides the well-known pubertal spurt. The dataset consists of distance measurements for 120 boys and 112 girls, from birth to 21 years of age. We will focus on the range 2 to 21 years, for which there are 31 measurements available. Then raw velocities will consist of 30 observations per individual. The data is regularly spaced but not equispaced (observations at puberty are more frequent).

Three registration methods were applied to the data:

- Landmark registration, as in Gasser et al. (1991).
- Self-modeling registration with $p = 3q$ B-spline basis functions of order 3 and equispaced knots. Fifty random starts were used (although 10 would have been sufficient for these data). The number of components q was selected by cross-validation.

Table 3: *Estimated prediction errors for leg growth example, using leave-one-out (CV-1) and five-fold (CV-5) cross-validation.*

	Males		Females		Combined	
q	CV-1	CV-5	CV-1	CV-5	CV-1	CV-5
2	17.3	16.4	18.1	15.6	18.1	16.4
3	10.9	15.6	17.6	14.8	9.6	15.2
4	15.6	16.8	11.7	14.6	19.1	17.5
5	8.8	16.6	13.5	14.8	18.6	18.0

- Continuous monotone registration, using the software provided by Jim Ramsay on his webpage. This method demands data smoothing, for which we have used a B-spline basis of order 4, with knots at 2, 4, 6, ..., 18 and 21 years, totaling 13 basis functions. The same basis was used for the warping functions.

Table 3 gives self-modeling cross-validation prediction errors for boys, girls and the combined sample of both sexes (to be used later). Both cross-validation procedures indicate a model with $q = 3$ components for boys and for the combined sample, and $q = 4$ components for girls. So these are the numbers we use. Fig. 4 shows the three registered means for boys and girls. We see that self-modeling registration captures the pubertal spurt both in timing and amplitude practically as well as landmark registration. At the midgrowth spurt the performance of self-modeling registration is not so remarkable, although for girls it is clearly better than monotone registration. Monotone registration produces a pubertal peak with low amplitude and shifted to the left. This time shift is mostly due to smoothing bias (caused by a highly asymmetric peak), but in part may also be due to the registration procedure itself.

To better discriminate the effects of registration from the effects of data smoothing, we have applied self-modeling registration to the same smooth velocity curves to which monotone registration was applied, evaluated at the same 30 points as the raw velocities (although using an equispaced grid would have improved the results). The result is shown in Fig. 5, where the structural means obtained by landmark registration were included in the plots again to facilitate comparisons. We observe that self-modeling registration captures the midgrowth spurt better now than with the raw velocities, specially for males. The pubertal

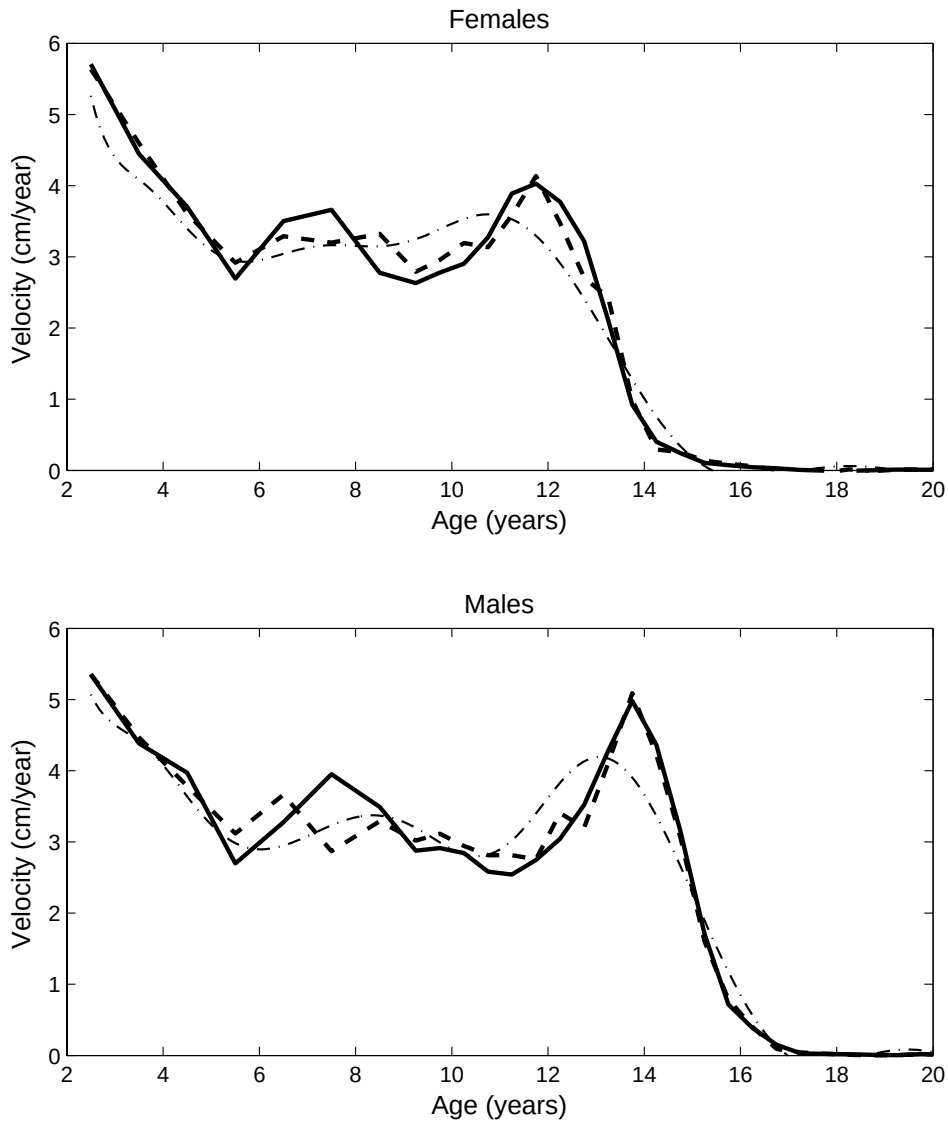

Figure 4: *Structural means of leg growth velocity curves, obtained via landmark registration (solid line), self-modeling registration (dashed line) and continuous monotone registration (dash-dot line).*

peak shows a considerable bias both in time and amplitude, although the time bias is worse for monotone registration. Data smoothing, then, improves registration in regions where the signal-to-noise ratio is low, but this comes at the price of amplitude bias in other regions.

Finally, let us exemplify the use of self-modeling scores for discrimination. We have registered the whole sample of 232 curves, combining both sexes. The results are shown in Fig. 6. The registered mean itself is not very meaningful, except that the maximal pubertal growth occurs somewhere between the normal ages for girls and boys. Since the third component is associated with the pubertal spurt, one would expect that the scores are larger for males than for females. This is exactly what happens. The boxplots indicate a significant difference, offering a clear-cut discrimination between sexes. A similar procedure can be derived from continuous monotone registration, using principal component scores of the warping functions. However, principal components are not as well localized as self-modeling components, so that the discrimination they provide is not so strong and not so neatly attributable to a specific time period (see e.g. the analysis on p. 96 of Ramsay and Silverman, 2002).

A Appendix

A.1 Identifiability of a general shape-invariant model

Consider model (1) with all the assumptions given in the text, plus the following:

1. The local extrema of μ are isolated points (i.e. μ is piecewise monotone without "flat" parts).

2. There exists a $t_0 \in T$ such that $\mu(t_0) > 0$ and $\mathbb{P}\{v(t_0) = t_0\} = 1$.

These assumptions are sufficient for identifiability, and probably necessary. The second assumption is always satisfied in practice by one or both extreme points of T (the restriction $\mu(t_0) > 0$ is just a technicality, which is always fulfilled by adding a constant to $x(t)$ if necessary). The first assumption is more restrictive but unavoidable: if $\mu(t)$ were constant on certain subintervals of T, the warping functions could be arbitrarily defined on those intervals.

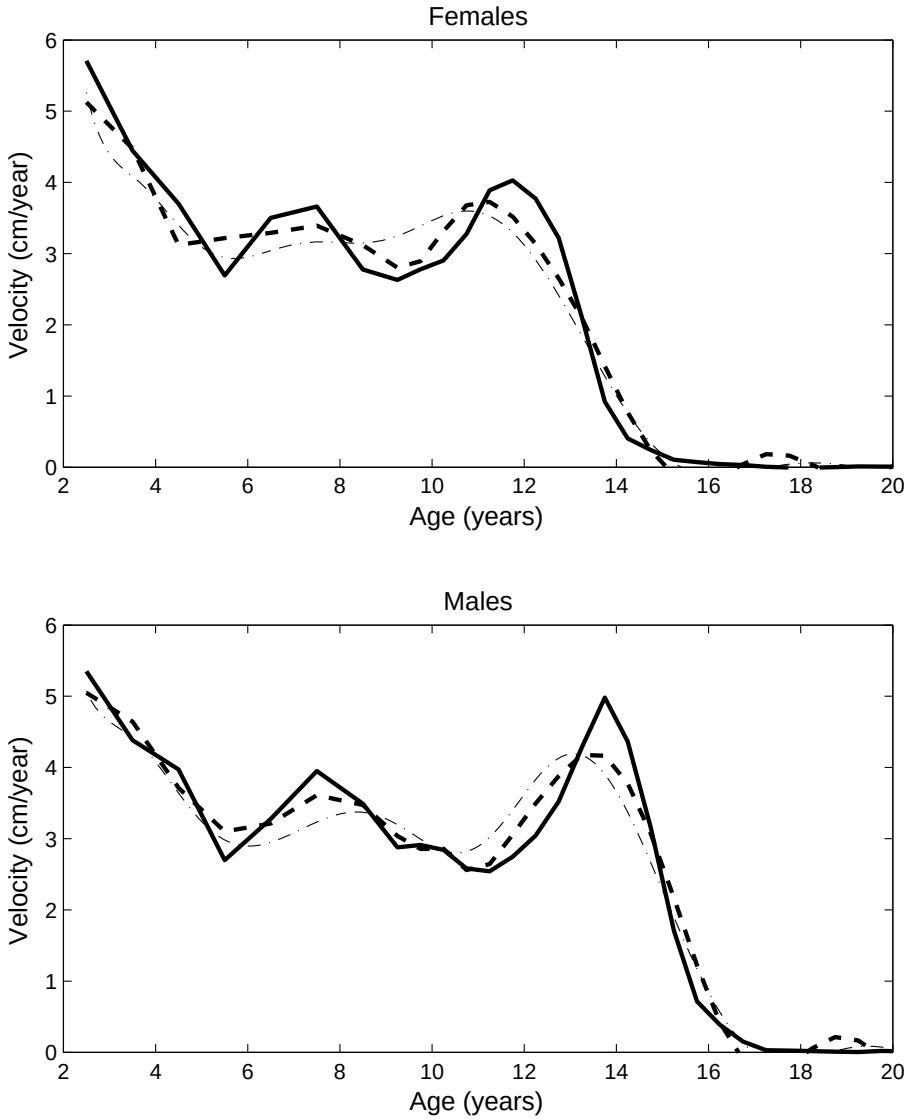

Figure 5: *Same as Fig. 4 but using smooth velocities for self-modeling registration.*

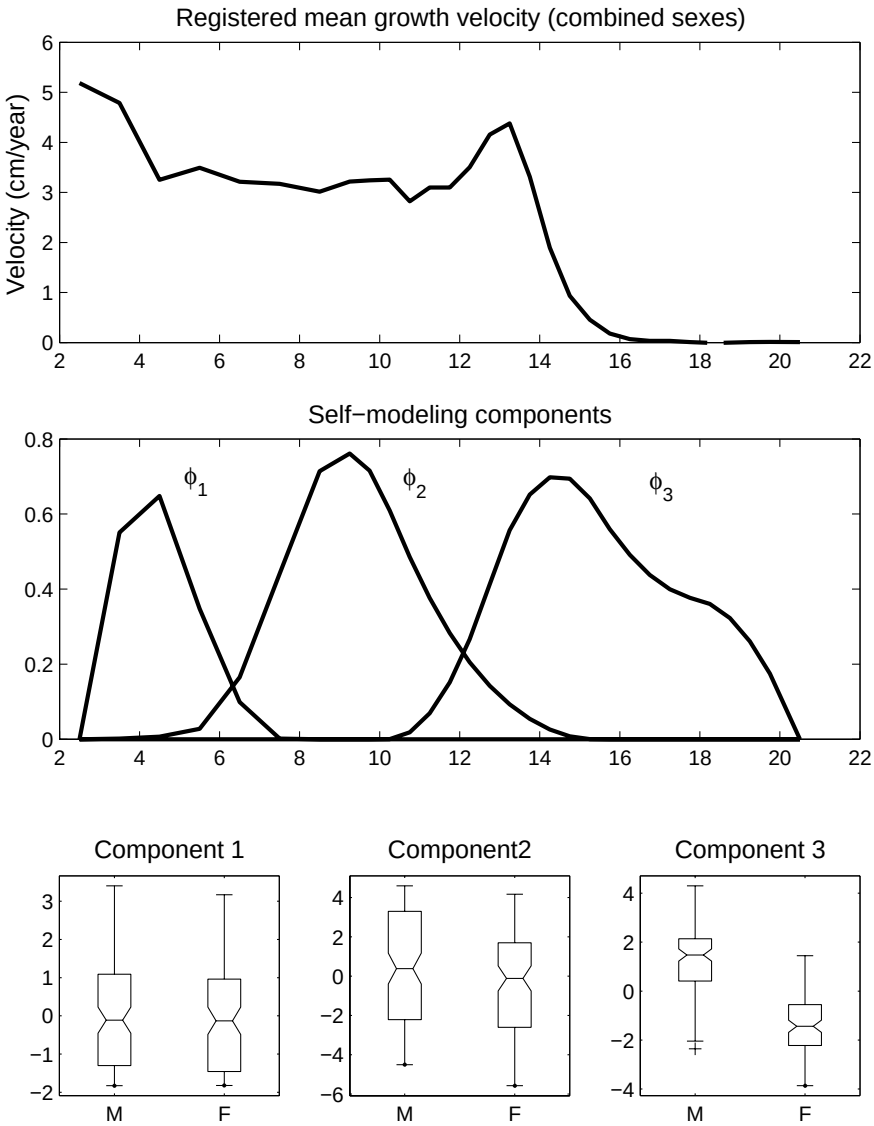

Figure 6: *Self-modeling registration of leg growth velocity curves combining both sexes.*

The proof that the model is identifiable goes as follows. Suppose that there are a, μ, v, ε and $\tilde{a}, \tilde{\mu}, \tilde{v}, \tilde{\varepsilon}$ such that $\mathbb{P}\{a \cdot \mu \circ v + \varepsilon = \tilde{a} \cdot \tilde{\mu} \circ \tilde{v} + \tilde{\varepsilon}\} = 1$. Then $\varepsilon - \tilde{\varepsilon} = \tilde{a} \cdot \tilde{\mu} \circ \tilde{v} - a \cdot \mu \circ v$ with probability one. Given the independence of the variables at both sides of the equation and the fact that $\mathbb{E}\{\varepsilon(t) - \tilde{\varepsilon}(t)\} = 0$ for all t, this implies that $\mathbb{P}\{\varepsilon = \tilde{\varepsilon}\} = 1$ and then $a \cdot \mu \circ v = \tilde{a} \cdot \tilde{\mu} \circ \tilde{v}$ with probability one. Evaluating the last equation at $t = t_0$ and taking expectations on both sides, we deduce that $\mu(t_0) = \tilde{\mu}(t_0)$, and then $\mathbb{P}\{a = \tilde{a}\} = 1$. So we have that $\mathbb{P}\{\mu \circ v = \tilde{\mu} \circ \tilde{v}\} = 1$, which by a conditional argument implies that $\mathbb{P}\{\mu \circ v \circ \tilde{v}^{-1} = \tilde{\mu}\} = 1$. Now take two independent realizations of $v \circ \tilde{v}^{-1}$, say $v_1 \circ \tilde{v}_1^{-1}$ and $v_2 \circ \tilde{v}_2^{-1}$, for which we then have $\mu(v_1 \circ \tilde{v}_1^{-1}(t)) = \mu(v_2 \circ \tilde{v}_2^{-1}(t))$ for all t. The piecewise monotonicity of μ implies that $v_1 \circ \tilde{v}_1^{-1} = v_2 \circ \tilde{v}_2^{-1}$ and, since they were independent realizations, it follows that $v \circ \tilde{v}^{-1}$ is non-stochastic. That is, $v \circ \tilde{v}^{-1} = c$ with probability one for some non-stochastic function c. Then $\tilde{v}^{-1}(t) = v^{-1}(c(t))$ for all t, and taking expectations on both sides we have that $c(t) = t$ for all t. Thus $\tilde{v}^{-1}(t) = v^{-1}(t)$ for all t with probability one, implying that $\mathbb{P}\{v = \tilde{v}\} = 1$ and consequently that $\mu = \tilde{\mu}$, which completes the proof.

A.2 Identifiability of self-modeling warping functions

Let us prove that model (2), under the assumptions given in the text, is identifiable. Suppose that $\mathbb{P}\{\mathbf{s}^\top CB(t) = \tilde{\mathbf{s}}^\top \tilde{C}B(t)\} = 1$ for all $t \in T$. Since a B-spline basis consists of linearly independent functions, this implies that $\mathbb{P}\{\mathbf{s}^\top C = \tilde{\mathbf{s}}^\top \tilde{C}\} = 1$. Since C and $\tilde{C}$ are in $\mathcal{C}$, we have that $CC^\top = \tilde{C}\tilde{C}^\top = I$ and that $D = C\tilde{C}^\top$ is diagonal. But then $\mathbf{s} = D\tilde{\mathbf{s}}$ and $\tilde{\mathbf{s}} = D\mathbf{s}$ with probability one, which implies that either $\mathbf{s} = \tilde{\mathbf{s}} = 0$ or $D = I$. This, in turn, implies that $\mathbf{s} = \tilde{\mathbf{s}}$ with probability one. Consequently, $\mathbf{s}^\top C = \mathbf{s}^\top \tilde{C}$ and then $C = \tilde{C}$, because matrices in $\mathcal{C}$ have full row rank. This completes the proof.

References

[1] Gasser, T., Kneip, A., and Ziegler, P. (1990). A method for determining the dinamycs and intensity of average growth. *Annals of Human Biology,* **17**, 459-574.

[2] Gasser, T., Kneip, A., Binding, A., Prader, A., and Molinari, L. (1991). The dynamics of linear growth in distance, velocity and acceleration. *Annals of Human Biology*, **18**, 187-205.

[3] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* Springer, NY.

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

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

[6] Möcks, J. (1986). The influence of latency jitter in principal component analysis of event-related potentials. *Psychophysiology*, **23**, 480-484.

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

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

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

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

[11] Wang, K. and Gasser, T. (1999). Synchronizing sample curves nonparametrically. *The Annals of Statistics*, **27**, 439-460.