Self-modelling warping functions

Daniel Gervini[1] and Theo Gasser
Department of Biostatistics, University of Zürich
Sumatrastrasse 30, 8006 Zürich, Switzerland

May 13, 2004

[1]Supported by Swiss National Science Foundation grant BE 20.63579.00.

Abstract

This article introduces a semiparametric model for functional data. The warping functions are assumed to be linear combinations of q common components, which are estimated from the data (hence the name "self-modelling"). Even small values of q provide remarkable model flexibility, comparable to nonparametric methods. At the same time, this approach avoids overfitting because the common components are estimated combining data across individuals. As a convenient by-product, component scores are often interpretable and can be used for statistical inference (an example of classification based on scores is given).

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

1 Introduction

Data sets consisting of samples of functions are increasingly common in statistics. Many examples of functional data, as well as an excellent introduction to the methodology, are given in the books of Ramsay and Silverman (1997, 2002). The aims of functional data analysis are similar to those of more traditional areas of statistics, but these kinds of data 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 mean $\bar{x}(t)$, does not always produce sensible results. This is illustrated in Fig. 3, where some growth velocity curves are plotted together with the cross-sectional means. From a biomedical point of view, it is important to estimate the average amplitude of the pubertal spurt (the peak occurring at age 12 for girls and age 14 for boys). But as Fig. 3 shows, the cross-sectional mean underestimates the amplitude of the pubertal spurt. The reason is that the peaks vary from person to person not only in intensity but also in timing. Thus, 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, and consequently $\bar{x}(\tau_0) < \overline{x_\cdot(\tau_\cdot)}$. The cross-sectional mean, then, always underestimates the amplitude of local maxima and overestimates the amplitude of local minima. This is a serious problem, because local extrema are often important features of the physical or biological process under study, and their amplitude should be estimated accurately.

As explained above, the bias of the sample mean at local extrema is due to time variability. We can think of a sample curve as a realization of a composed stochastic process: $x(t) = \eta(v(t))$, $t \in T$, where $E\{\eta(t)\} = \mu(t)$ and $v : T \to T$ is strictly increasing with probability one. The η-component is the source of amplitude variability about the structural mean $\mu(t)$. The v-component generates time variability about μ, shifting the location of important features of the curves (like the pubertal peaks of Fig. 3). In general $E\{x(t)\} \neq \mu(t)$, so that $\bar{x}(t)$ is not a consistent estimator of $\mu(t)$.

There are a number of proposals that deal with the time-variability problem, providing estimates for $\{v_i(t)\}$ and $\mu(t)$: landmark registration (Gasser et al., 1990, and Kneip and Gasser, 1992), shape-invariant modelling (Kneip and Engel, 1995), continuous monotone registration (Ramsay and Li, 1998), dynamic time warping (Wang and Gasser, 1999), local regression (Kneip et al., 2000) and maximum likelihood registration (Rønn, 2001), the latter restricted to random shifts. The idea behind these methods is to estimate the processes $v_i(t)$ (or their inverses $w_i(t) = v_i^{-1}(t)$, the so-

called warping functions) and then compute aligned functions $x_i^\star(t) = x_i(w_i(t))$. The structural mean is then estimated by $\hat{\mu}(t) = \overline{x^\star}(t)$. We note that, with the exception of Rønn (2001), all these methods (as well as the method introduced in this paper) treat the functions $\{v_i(t)\}$ as fixed parameters to be estimated, rather than realizations of a stochastic process. A consequence of this is that the asymptotic theory of the estimators when n goes to infinity becomes intractable. Nevertheless, simulations and examples have repeatedly shown that these methods provide good estimates of $\mu(t)$.

Landmark registration, when feasible, is probably the best of the existing methods. It consists in identifying for each individual the timings of salient features (landmarks), such as local extrema and zero crossings, and then aligning the curves so that individual landmarks coincide with mean landmarks. Unfortunately, individual identification of landmarks may be very difficult and time-consuming. For instance, Gasser et al. (1991) used eight landmarks in their analysis of growth data, which had to be manually determined for each of the 232 sample curves.

Continuous monotone registration, on the other hand, is a nonparametric method that does not require landmark identification. It models $\log\{w'(t)\}$ as a linear combination of B-spline basis functions, allowing great flexibility for the warping functions. However, the method is susceptible to "time-variability overfitting" when too many basis functions are used (see Ramsay and Li, 1998, for an illustration of this problem). This problem can be ameliorated by a roughness penalty approach, at the cost of increased computational time. But we think that a more efficient way to avoid overparameterisation is to combine data across individuals, as we propose in this paper.

In this article we introduce a semiparametric method of registration that does not involve landmark identification, offers considerable warping flexibility and avoids to a large extent the problem of overfitting. The idea is to model the warping functions as linear combinations of a small number of components: $w_i(t) = t + \sum_{j=1}^{q} s_{ij}\phi_j(t)$. Component functions $\{\phi_j\}$ are estimated from the data, hence the name "self-modelling" warping functions. Each of these components is a linear combination of B-spline basis functions. Since the components are common to all subjects, their spline coefficients are estimated combining data across curves. Only the scores $s_{i1}, \ldots, s_{iq}$ are individually estimated, but q is usually small (for the growth example $q = 4$ suffices, as we show in Section 4). The parsimony of individual parameters is what makes this method less prone to overfitting than continuous monotone registration.

2 Self-modelling registration

2.1 The model

We will assume that the sample curves $x_1(t), \ldots, x_n(t)$ follow the model

$$x_i(t) = a_i \mu(v_i(t)) + \varepsilon_i(t), \quad t \in T \subset \mathbb{R}, \quad i = 1, \ldots, n, \tag{1}$$

where $\mu : T \to \mathbb{R}$ is the structural mean, $\{v_i : T \to T\}$ are monotone increasing functions and $\{\varepsilon_i\}$ are random errors. For identifiability (which is proved in Appendix A.1), we assume that $E\{\varepsilon_i(t)\} = 0$ for all $t \in T$, $a_i \neq 0$, $\bar{a} = 1$, and $\bar{w}(t) = t$ for all $t \in T$ (where $w_i(t) = v_i^{-1}(t)$). It is true that the type of amplitude variability allowed in model (1) is rather limited, but it is a good approximation to many real-life datasets that we have analyzed (in particular, for the growth example of Section 4). We use (1) essentially as a working model, since we are mostly concerned with estimation of the warping functions. The ideas presented below can be extended to more complex models.

As mentioned in the Introduction, for the warping functions we propose the model

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

where the score vectors $\mathbf{s}_i = (s_{i1}, \ldots, s_{iq})^\top$ satisfy $\bar{\mathbf{s}} = \mathbf{0}$. The components are modeled as $\phi_j(t) = \mathbf{c}_j^\top \boldsymbol{\beta}(t)$, where $\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_p(t))^\top$ is a vector of B-spline basis functions. The kind of ϕ-functions we have in mind are localized, nonnegative bell-shaped functions as in Fig. 4. These functions are easy to interpret, each of them accounting for time variability at different segments of T. We can think of each ϕ_j as a component associated with a "hidden landmark" (Section 2.2 gives the rationale for this).

To obtain ϕ-functions of the form described above, and for identifiability reasons, the spline coefficients must satisfy the following conditions:

1. $c_{jk} \geq 0$ for $k = 1, \ldots, p$ and $\|\mathbf{c}_j\| = 1$, for all $j = 1, \ldots, q$.

2. The coefficient matrix $C = (c_{jk}) \in \mathbb{R}^{q \times p}$ has the following 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 $k \geq K_{j+1}$.

3. $c_{j1} = c_{jp} = 0$ 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 guarantees that the components have connected and localized supports. This rules out, for example, the case of "one" component actually consisting of two disjoint bell-shaped curves. Restriction 3 guarantees that the warping functions are "tied down" at the extremes of T; that is, $w_i(a) = a$ and $w_i(b) = b$ when $T = [a, b]$. The identifiability of model (2) is proved in Appendix A.2.

2.2 Self-modelling and landmark registration

We are not proposing model (2) simply because it is flexible and convenient; this model is motivated by landmark registration. Consider, for instance, the case of two landmarks per curve, τ_{i1} and τ_{i2}. Let $\tau_{01} = \bar{\tau}_{\cdot 1}$ and $\tau_{02} = \bar{\tau}_{\cdot 2}$ be the average landmarks. The registered curves $\hat{x}_i^\star(t) = x_i(w_i(t))$ are aligned so that $\hat{x}_i^\star(\tau_{0j}) = x_i(\tau_{ij})$, $j = 1, 2$, for all i. Therefore, the warping functions must be constructed so that $w_i(a) = a$, $w_i(\tau_{01}) = \tau_{i1}$, $w_i(\tau_{02}) = \tau_{i2}$ and $w_i(b) = b$. Using piecewise linear functions (the simplest interpolation method) we have that

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

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

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

These functions are triangles with peaks at τ_{01} and τ_{02}, respectively, which can be expressed as combinations of linear B-splines with knots $\{a, \tau_{01}, \tau_{02}, b\}$. Hence model (2) holds with $q = 2$; each component is associated with an underlying landmark. The component scores are just the deviations of the individual landmarks from the average landmarks. Roughly speaking, we can say that self-modelling registration is an attempt to back-engineer landmark registration: instead of estimating the individual landmarks, we estimate the associated components.

2.3 Estimation

The parameters of models (1) and (2) can be estimated by minimizing the average integrated squared error,

$$\begin{aligned} \mathrm{AISE}_n &= \frac{1}{n}\sum_{i=1}^{n}\int_a^b \{x_i(t) - a_i\mu(v_i(t))\}^2\,dt \qquad (3)\\ &= \frac{1}{n}\sum_{i=1}^{n}\int_a^b \{x_i(w_i(t)) - a_i\mu(t)\}^2 w_i'(t)\,dt. \end{aligned}$$

Clearly, we have

$$\hat{\mu}(t) = \frac{\sum_{i=1}^n \hat{a}_i\hat{w}_i'(t)\hat{x}_i^\star(t)}{\sum_{i=1}^n \hat{a}_i^2\hat{w}_i'(t)}, \qquad (4)$$

where $\hat{x}_i^\star(t) = x_i(\hat{w}_i(t))$. For the $\hat{a}_i$'s there are explicit estimating equations too, but not for $\hat{C}$ or the $\hat{\mathbf{s}}_i$'s.

To minimize (3) we have implemented the following two-stage algorithm (a Matlab routine is available on the first author's webpage):

1. INITIALIZATION.

   (a) Select "promising" block delimiters $\mathbf{K} = (K_1, \ldots, K_{q+1})$ for the coefficient matrix C (see comments below). Reparameterise

   $$(c_{j,K_j}, \ldots, c_{j,K_{j+1}-1}) = (1, \exp(\mathbf{y}_j))/\{1 + \|\exp(\mathbf{y}_j)\|^2\}^{\frac{1}{2}}, \qquad (5)$$

   where $\mathbf{y}_j \in \mathbb{R}^{K_{j+1}-K_j}$ are unconstrained vectors.

   (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. For each $j = 1, \ldots, q$: update $\mathbf{y}_j$ using a Newton–Raphson step.

      ii. Recenter current $\hat{\mathbf{s}}_i$'s so that $\bar{\hat{\mathbf{s}}} = 0$. For each $i = 1, \ldots, n$: update $\hat{\mathbf{s}}_i$ using a Newton–Raphson step (make sure that the resulting $\hat{w}_i(t)$ is strictly increasing by reducing the step size, if necessary).

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

i. Update $\hat{\mu}$ using (4), computing $x_i(\hat{w}_i(t))$ by linear interpolation.

ii. Update $\hat{a}_i$.

iii. Update objective function. Exit if there is no significant improvement; otherwise go back to (a).

Selecting adequate block delimiters at the initialization stage of the algorithm is crucial. There are $\binom{p-3}{q-1}$ different vectors of delimiters, so that an exploration of all possibilities is infeasible in many situations. A workable alternative is to (i) generate random delimiters (50, say), (ii) make 2 or 3 iterations of the algorithm for each of them, keeping the delimiters that yield the lowest objective function (or the three lowest, say), and (iii) iterate the latter until convergence. This procedure does not guarantee that the optimal $\mathbf{K}$ will be found, but since similar delimiters produce similar components, an approximate good solution is likely to be found.

So far we have assumed that the sample curves $x_i(t)$ are observed at every $t \in T$. This is an idealized situation, though. In practice, the raw dataset will consist of vectors $\mathbf{x}_i \in \mathbb{R}^m$, where

$$x_{ij} = a_i \mu(v_i(t_j)) + \varepsilon_{ij} \tag{6}$$

and $\{t_1, \ldots, t_m\} \subset T$ is a finite input grid (which is assumed, for simplicity, to be the same for all curves). There are two ways to proceed. One way is to pre-smooth the data and then minimize (3) using the smoothed curves. The other alternative is to use the raw data and minimize a discretized version of (3), namely

$$\text{AISE}_{nm} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \{x_{ij} - a_i \mu(v_i(t_j))\}^2 (t_j^\star - t_{j-1}^\star), \tag{7}$$

where $t_0^\star = t_1$, $t_j^\star = (t_j + t_{j-1})/2$ for $j = 2, \ldots, m-1$, and $t_m^\star = t_m$. The minimizer $\hat{\mu}(t)$ of (7), properly speaking, is not well-defined for all $t \in T$, but (4) is a practical approximation. Alternatively, one can use a spline estimator for μ: assuming that $\mu(t) = \mathbf{d}^\top \tilde{\boldsymbol{\beta}}(t)$, where $\tilde{\boldsymbol{\beta}}(t)$ is a vector of $\tilde{p}$ spline basis functions, we have

$$\begin{aligned} \hat{\mathbf{d}} \;=\; & \{\sum_{i=1}^{n} \sum_{j=1}^{m} \hat{a}_i \tilde{\boldsymbol{\beta}}(\hat{v}_i(t_j)) \tilde{\boldsymbol{\beta}}^\top (\hat{v}_i(t_j)) (t_j^\star - t_{j-1}^\star)\}^{-1} \times \\ & \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} \tilde{\boldsymbol{\beta}}(\hat{v}_i(t_j)) (t_j^\star - t_{j-1}^\star). \end{aligned} \tag{8}$$

Using raw data, as opposed to individually smoothing the curves, has the advantage of avoiding smoothing bias to a large extent, because (8) makes use of all nm observations and this allows for a larger number of basis functions $\tilde{p}$ than would be feasible for individual curves.

2.4 Choosing the number of components

Choosing the number of components and basis functions, q and p, is an essential step of self-modelling registration. Usually, this kind of model choice is carried out by minimizing a measure of prediction error (Hastie, Tibshirani and Friedman 2001, ch. 7). In our context, the natural measure of prediction error is the mean integrated squared error, $\mathrm{MISE} = n^{-1}\sum_{i=1}^{n} E\int\{x(t) - a_i\mu(v_i(t))\}^2\,dt$. It is well known that $\mathrm{AISE}_n(\hat{\mu}, \{\hat{a}_i\}, \{\hat{v}_i\})$ underestimates MISE, so that an alternative method such as cross-validation has to be used.

We propose the following cross-validation algorithm. Let Ω be a set of tentative parameters (p, q). Then:

1. Split the sample into N subsamples ("test sets") of roughly the same size. Let $\mathcal{I}_k = \{i : x_i \text{ is in the } k\text{th test set}\}$, $k = 1, \ldots, N$.

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

   (a) Find $\{\hat{a}_i\}$, $\{\hat{w}_i(t)\}$ and $\{\hat{x}_i^\star(t)\}$ (using all observations).

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

      i. Estimate μ as $\hat{\mu}^{(-k)}(t) = \sum_{i\notin\mathcal{I}_k} \hat{a}_i\hat{x}_i^\star(t)\hat{w}_i'(t) / \sum_{i\notin\mathcal{I}_k} \hat{a}_i^2\hat{w}_i'(t)$.

      ii. For each $i \in \mathcal{I}_k$, estimate the squared prediction error as:

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

   (c) $\widehat{\mathrm{CMISE}}(p, q) = \sum_{i=1}^{n}\hat{e}_i^2/n$.

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

Usual values of N are 5, 10 and n (the latter being the familiar leave-one-out cross-validation). Hastie, Tibshirani and Friedman (2001, ch. 7.10) note that the leave-one-out cross-validation estimator is nearly unbiased but very variable, so they recommend

five-fold or ten-fold cross-validation, even when they tend to overestimate the prediction error.

In practice Ω will not be too large, making cross-validation feasible. The number of salient features of the curves (such as peaks or valleys) will suggest plausible values of q. If raw data are used, the number of observations per curve will also impose a bound on q, since q scores have to be estimated for each curve ($q < \sqrt{m}$ is the rule of thumb). Regarding the number of basis functions p, we have found that $p = 3q$ or $p = 4q$ work well for many datasets, while larger values of p tend to produce components with irregular shapes and higher prediction errors. In the growth example of Section 4, all these considerations reduce Ω to only six pairs of parameters.

2.5 Consistency and asymptotic normality

Two types of consistency results are of interest. One corresponds to the continuous-time model (1) with the number of curves n going to infinity, where $\{a_i\}$ and $\{\mathbf{s}_i\}$ play the role of nuisance parameters and the focus is on the asymptotic behavior of $\hat{\mu}$ and $\hat{\phi}_1, \ldots, \hat{\phi}_q$. The other type of result corresponds to the discrete-time (raw-measurement) model (6) with a fixed number of individuals n and the number of observations per individual m going to infinity. Admittedly, the first situation is of more interest in the present context, but the number of estimated nuisance parameters grows with n and the problem becomes intractable, so we will study the second type of asymptotics in this section.

Consider, then, model (6) with i.i.d. errors $\{\varepsilon_{ij}\}$ such that $E(\varepsilon_{ij}) = 0$, $V(\varepsilon_{ij}) = \sigma^2$ and $E(\varepsilon_{ij}^4) < \infty$. Assume that $\mu_0(t) = \mathbf{d}_0^\top \tilde{\boldsymbol{\beta}}(t)$ (in what follows, true parameters are indicated by a nought subscript). Then

$$\begin{aligned} E(\mathrm{AISE}_{nm}) &= \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n}\{a_{0i}\mu_0(v_{0i}(t_j)) - a_i\mu(v_i(t_j))\}^2(t_j^\star - t_{j-1}^\star) \\ &\quad +\sigma^2(b-a), \\ V(\mathrm{AISE}_{nm}) &= \frac{1}{n^2}\sum_{j=1}^{m}\sum_{i=1}^{n}\kappa_{ij}(t_j^\star - t_{j-1}^\star)^2, \end{aligned}$$

where $\kappa_{ij} = V\{(\varepsilon_{ij} + a_{0i}\mu_0(v_{0i}(t_j)) - a_i\mu(v_i(t_j)))^2\}$. Assuming that the time grid is regularly spaced in the sense that $\max_j(t_j^\star - t_{j-1}^\star) = O(m^{-1})$, we have $V(\mathrm{AISE}_{nm}) \to$

0 as $m \to \infty$ and then

$$\mathrm{AISE}_{nm} \xrightarrow{P} \frac{1}{n}\sum_{i=1}^{n}\int \{a_{0i}\mu_0(v_{0i}(t)) - a_i\mu(v_i(t))\}^2\,dt + \sigma^2(b-a).$$

The right-hand side of this equation is minimized by $a_i = a_{0i}$, $\mu = \mu_0$ and $v_i = v_{0i}$, and these minimizers are unique by identifiability. Under appropriate regularity conditions, Theorem 5.7 of van der Vaart (1998) implies that $\hat{a}_i$, $\hat{\mathbf{s}}_i$, $\hat{\mathbf{c}}_j$ and $\hat{\mathbf{d}}$ are consistent as m goes to infinity.

The asymptotic normality of these estimators is not difficult to establish, assuming that the warping model (involving p, q and the delimiters $\mathbf{K}$) is correctly specified. Let $\boldsymbol{\theta} = (\mathbf{d}, \mathbf{y}, a_1, \ldots, a_n, \mathbf{s}_1, \ldots, \mathbf{s}_n)$, where $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_q)$ is the reparameterisation (5) of $(\mathbf{c}_1, \ldots, \mathbf{c}_q)$, and let $l_{ij}(\boldsymbol{\theta}) = \{x_{ij} - a_i\mu(v_i(t_j))\}^2/2\sigma^2$. Then $\hat{\boldsymbol{\theta}}$ solves the estimating equation

$$\sum_{i=1}^{n}\sum_{j=1}^{m} \mathsf{D}l_{ij}(\hat{\boldsymbol{\theta}})(t_j^\star - t_{j-1}^\star) = \mathbf{0},$$

where D denotes the differential. Under appropriate conditions (see Theorem 5.21 of van der Vaart 1998) we have that $\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{D} \mathcal{N}(0, A_n^{-1}B_nA_n^{-1})$ as m goes to infinity, with

$$\begin{aligned} A_n &= \lim_{m\to\infty}\sum_{j=1}^{m}\sum_{i=1}^{n} E\{\mathsf{D}^2 l_{ij}(\boldsymbol{\theta}_0)\}(t_j^\star - t_{j-1}^\star), \\ B_n &= \lim_{m\to\infty}\sum_{j=1}^{m} E[\{\sum_{i=1}^{n}\mathsf{D}l_{ij}(\boldsymbol{\theta}_0)\}\{\sum_{i=1}^{n}\mathsf{D}l_{ij}(\boldsymbol{\theta}_0)\}^\top](t_j^\star - t_{j-1}^\star). \end{aligned}$$

The explicit computation of A_n and B_n is tedious but straightforward. Some simplifications are possible when $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$: in that case, $A_n = B_n$ and the asymptotic covariance of $\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is just A_n^{-1}.

3 Simulations

3.1 Evaluating model fit

We carried out some simulations to evaluate the comparative performance of self-modelling registration with landmark registration. We wanted to assess (i) the perfor-

mance of the estimator of the structural mean and (ii) the association between component scores and curve landmarks. As structural mean we chose the scaled sinus function $\mu(t) = \sin(2\pi t)$, for $t \in [0, 1]$, which has a peak at $\tau_{01} = .25$ and a valley at $\tau_{02} = .75$, providing two natural landmarks. As input grid we used equidistant points $t_j = (j-1)/(m-1)$, $j = 1, \ldots, m$. The data were simulated as follows:

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

2. The inverse warping functions $v_i(t) = w_i^{-1}(t)$ were piecewise linear with $v_i(0) = 0$, $v_i(\tau_{i1}) = \tau_{01}$, $v_i(\tau_{i2}) = \tau_{02}$ and $v_i(1) = 1$.

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

4. The observations followed model (1) with $\mathcal{N}(0, \sigma_e^2)$ random errors. Two error variances were considered: $\sigma_e^2 = 0$ (no error term) and $\sigma_e^2 = 1/10\sqrt{2}$ (error-to-model ratio: $\sigma_e^2/(\sigma_a^2 \int \mu^2 + \sigma_e^2) = .5$). Note that a model without error term is roughly equivalent to using smoothed data.

5. Sample sizes $n = 10, 20, 50$ and grid sizes $m = 10, 20, 50$ were used, with 200 replications for each combination.

We compared three estimators of μ:

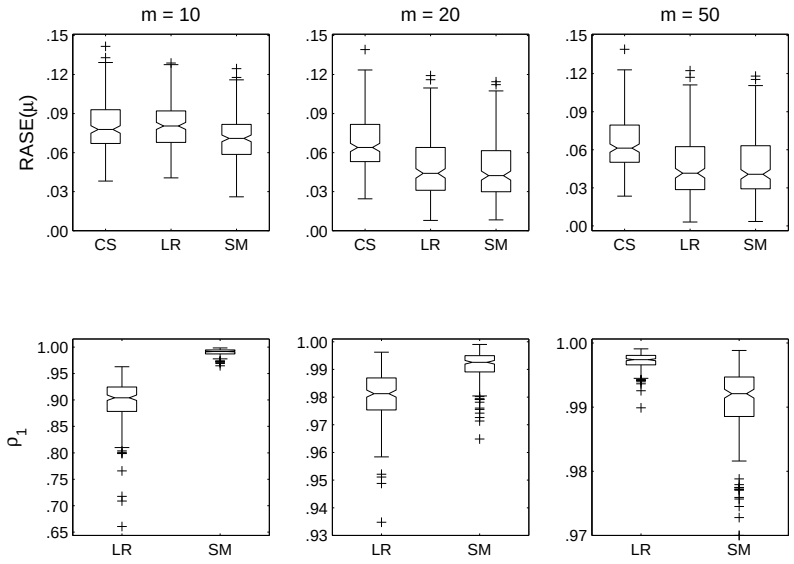
- The cross-sectional mean.

- Landmark registration, with landmarks estimated from the raw data: $\hat{\tau}_{i1} = \arg\max_{t_j}\{x_i(t_j)\}$ and $\hat{\tau}_{i2} = \arg\min_{t_j}\{x_i(t_j)\}$.

- Self-modelling registration, with $q = 2$ components and $p = 6$ basis functions (B-splines of order 3 with equidistant knots), computed with the algorithm of Section 2.3.

As measure of estimation error we used the root average squared error $\mathrm{RASE}(\hat{\mu}) = \{\sum_{j=1}^{m}(\hat{\mu}(t_j) - \mu(t_j))^2/m\}^{\frac{1}{2}}$. To measure the association between scores and landmarks we used $\rho_k^{\mathrm{SM}} = \mathrm{corr}(\{\hat{s}_{ik}\}, \{\tau_{ik}\})$, $k = 1, 2$. For comparison, we also computed $\rho_k^{\mathrm{LR}} = \mathrm{corr}(\{\hat{\tau}_{ik}\}, \{\tau_{ik}\})$, $k = 1, 2$.

Figure 1: Simulated root average squared errors of $\hat{\mu}$ (upper panels) and landmark correlations (lower panels) for cross-sectional mean (CS), landmark registration (LR) and self-modelling registration (SM). Sample size $n = 20$, model without error term.

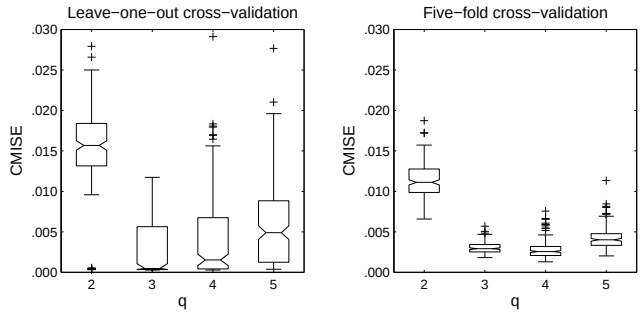


Simulation results are illustrated by Fig. 1, which shows boxplots of simulated errors and correlations for $n = 20$ (boxplots of ρ_2^{SM} and ρ_2^{LR} are almost identical to those of ρ_1^{SM} and ρ_1^{LR}). A more detailed report can be obtained from the authors. In general, we see that estimation errors of both registration methods decrease as m increases. For large m both methods are comparable, but for small m self-modelling registration is better. This shows the advantage of pooling information across individuals when the number of observations per individual is small. The correlations ρ_1^{SM} are consistently high for all n and m, indicating that the components are well-localized and associated with the underlying landmarks, as expected.

3.2 Evaluating model choice

Another simulation study was carried out to evaluate the cross-validation method proposed in Section 2.4. We compared leave-one-out and five-fold cross-validation. The data were generated as in Section 3.1, except that we took $T = [0, 1.5]$ and added a third landmark at $t = 1.25$ (the second peak of the scaled sinus function μ). Now the "true" number of components was $q = 3$. We took $n = 50$ and $m = 30$, and a model without error term. Each sampling situation was replicated 200 times, and 50 random

Figure 2: Simulated cross-validation estimators $\widehat{\mathrm{CMISE}}$ of self-modelling registration, for four different models ($q = 3$ is the true model).



starts were used for the self-modelling algorithm. The set of tentative parameters was $\Omega = \{(p, q) : p = 3q,\ q = 2, 3, 4, 5\}$.

Fig. 2 shows boxplots of simulated cross-validation estimators $\widehat{\mathrm{CMISE}}(p, q)$ for different q's. As expected, leave-one-out cross-validation is less biased but much more variable than five-fold cross-validation. On average, leave-one-out cross-validation picks out the right model, while five-fold cross-validation selects the slightly overparameterised model $q = 4$. This is due to the tendency of five-fold cross-validation to overestimate prediction error and select overparameterised 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 model 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%). Five-fold cross-validation selects the right model for 26% of the samples, $q = 4$ for 69%, $q = 5$ for 5%, and never selects $q = 2$. It is clear that five-fold cross-validation is more stable, even though it selects the overparameterised model $q = 4$ in most cases. Observe that leave-one-out cross-validation selects either one of the worse two models ($q = 2, 5$) for 12% of the samples, while five-fold cross-validation does so for only 5% of the samples. All things considered, five-fold cross-validation is a recommendable model-selection method, keeping in mind its tendency to select slightly overparameterised models.

4 Example: leg growth velocity

The datasets produced by the First Zurich Growth Study have been analyzed many times in the literature. To illustrate our method we have chosen leg growth curves among the body measurements available, because these velocity curves present a clear-cut midgrowth spurt (the peak about age 7), in addition to the well-known pubertal spurt. The raw dataset consists of leg length measurements for 120 boys and 112 girls. We will focus on growth velocity from 2 to 20 years; there are 30 measurements per person in this range. We computed smoothed velocity curves using Gasser–Müller kernel smoothers with locally optimal bandwidths (as in Gasser et al., 1991), evaluated at 100 equidistant points between 2 and 20. Some of these curves are shown in Fig. 3.

Three registration methods were applied to the data:

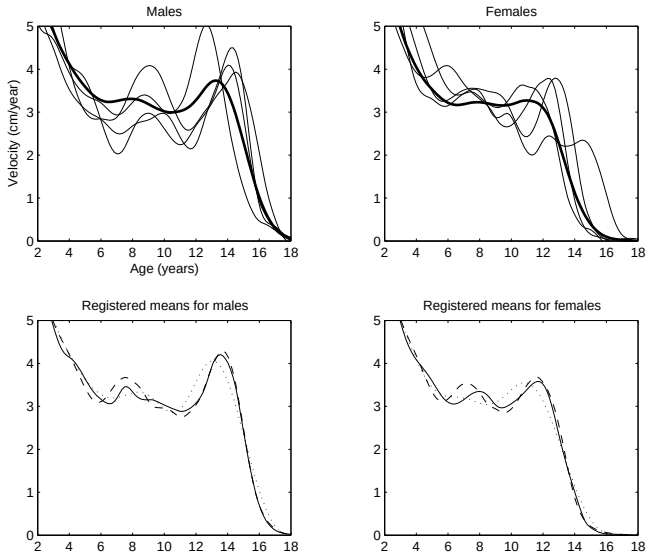
- Landmark registration with eight landmarks, as in Gasser et al. (1991). The landmarks were the four zero-crossings and the four local extrema of a typical acceleration curve.

- Self-modelling registration with 4 components and 12 B-spline functions of order 3 and equally spaced knots. These parameters were selected by five-fold cross-validation among tentative parameters $p = 3q$ and $p = 4q$ with $q = 2, 3, 4$. For the estimation algorithm we tried all $\binom{p-3}{q-1}$ possible sets of delimiters when this number was less than 50; otherwise, 50 randomly chosen delimiters were tried.

- Continuous monotone registration, using the software provided by Jim Ramsay on his webpage. We used a B-spline basis of order 4, with knots $\{2, 4, 6, \ldots, 20\}$, both for velocity curves and warping functions. (Ramsay's software requires that the curves be expressed as combinations of B-splines, but this was done with already smoothed velocities, so that little smoothing bias was introduced at this step).

Fig. 3 shows registered means for boys and girls. The estimates obtained by landmark registration may be considered the benchmark. Clearly, self-modelling registration estimates the pubertal spurt very accurately for both sexes, both in timing and amplitude, while continuous monotone registration is biased. The midgrowth spurt is more difficult to estimate (especially for girls, because it is too close to the pubertal spurt and of roughly the same amplitude), but self-modelling registration still produces

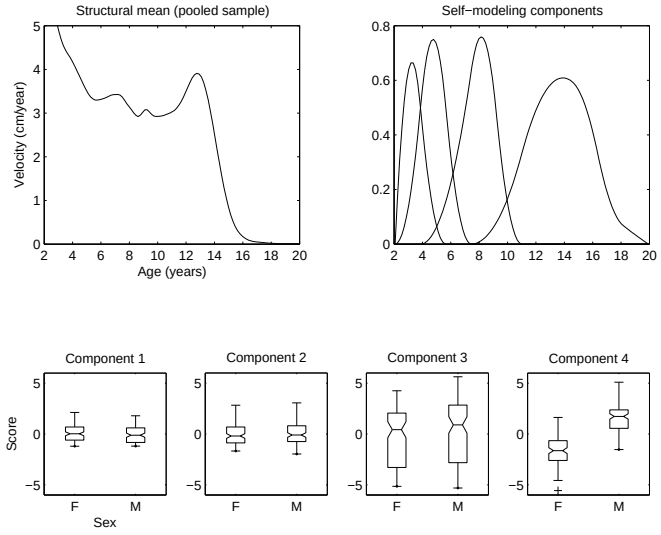
Figure 3: Upper panels: some leg growth velocity curves of boys and girls (thin lines), and cross-sectional means of the whole samples (thick lines). Lower panels: registered means (self-modelling registration, solid line; landmark registration, dashed line; continuous monotone registration, dotted line).



better estimates of it than continuous monotone registration. The ASEs (with respect to landmark registered means) of self-modelling registration are .34 for boys and .33 for girls, while those of continuous monotone registration are 1.14 and .84, respectively.

Although we are mainly concerned with estimation of $\mu(t)$, this dataset serves as a good example of classification based on self-modelling scores. We registered the sample of the 232 curves together, pooling both sexes. The results are shown in Fig. 4. The registered mean is not meaningful from a biological point of view; however, we see that the third self-modelling component is associated with the midgrowth spurt and the fourth component with the pubertal spurt. Looking at the boxplots of the scores, we see that there is no significant difference between sexes for the third component, indicating that the midgrowth spurt occurs, on average, at the same time for boys and girls. In contrast, there is a significant difference for the fourth component, reflecting the well-known fact that puberty occurs earlier for girls than for boys. If we classify as "females" those individuals with $s_{i4} < \bar{s}_{\cdot 4}$ and as "males" those with $s_{i4} > \bar{s}_{\cdot 4}$, the misclassification rates are 15% for males and 11% for females. If we use the time of maximal pubertal velocity (the seventh landmark) to classify individuals in this way,

Figure 4: Self-modelling registration of leg growth velocities (pooling both sexes). Upper panels: registered mean and estimated components. Lower panels: component scores.



the misclassification rates are 14% for males but 16% for females. This illustrates the potential use of self-modelling scores for discrimination and classification; the topic deserves further elaboration, but that exceeds the scope of this article.

5 Discussion

We have introduced a semiparametric model for warping functions of random curves. As explained in Section 2.2, this model is natural when the warping functions depend on recognizable landmarks. It is not necessary that the landmarks be visible in all sample curves; it suffices that the underlying stochastic model has recognizable landmarks, which is a plausible assumption in many situations. The advantage of our proposal over landmark registration is that individual identification of landmarks is not necessary. The advantage over continuous monotone registration is that, by exploiting the common structure of the warping functions, it makes a more efficient use of the data and avoids overfitting to a large extent. Admittedly, our simulations and examples were restricted to situations with clear underlying landmarks. Since continuous monotone registration is essentially model-free, it may be superior in other situations.

We also note that, by assuming a simple model for amplitude variability, we have

kept the focus on the structural mean. But more realistic models involve components for amplitude variability, where time warping also plays a role (see ch. 6 of Ramsay and Silverman 2002 for some examples). It does not seem difficult to extend self-modelling registration to such models, but this requires further research.

A Appendix

A.1 Identifiability of model (1)

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

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

This assumption is clearly necessary: if $\mu(t)$ were constant on some subinterval of T, the warping functions could be arbitrarily defined there and the model would not be identifiable.

The identifiability proof goes as follows. Since $E\{\varepsilon_i(t)\} = 0$, we have that $\varepsilon_i(t) = x_i(t) - E\{x_i(t)\}$, so there is no ambiguity about the error term. Now, suppose that $E\{x_i(t)\} = a_i\mu(v_i(t)) = a_i^*\mu^*(v_i^*(t))$ for all $i = 1, \ldots, n$. Then we have that $\mu(t) = (a_i^*/a_i)\mu^*(v_i^*(w_i(t)))$ for all t and for all i. Since the left-hand side does not depend on i, we have that $a_i^*/a_i = c$ for all i and some constant c, and that $v_i^*(w_i(t)) = g(t)$ for all i and some function g (a consequence of the piecewise monotonicity of μ^*). Then $w_i(t) = w_i^*(g(t))$ for all i, and since $\bar{w}(t) = \overline{w^*}(t) = t$ by assumption, we have that $g(t) = t$. Therefore $v_i = v_i^*$ for all i and the warping functions are identifiable. On the other hand, $a_i^* = ca_i$ for all i, and the assumption $\overline{a^*} = \bar{a} = 1$ implies that $c = 1$; then $a_i = a_i^*$ for all i, so that the scaling parameters are also identifiable. The identifiability of the scaling factors and the warping functions immediately imply that $\mu = \mu^*$, which completes the proof.

A.2 Identifiability of self-modelling warping functions

In this section we prove that model (2) is identifiable under the assumptions given in the text plus the following condition:

- $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ contains a subset of q linearly independent scores.

Suppose that two parametrisations of the warping functions were possible, so that $\mathbf{s}_i^\top CB(t) = \tilde{\mathbf{s}}_i^\top \tilde{C}B(t)$ for all $t \in T$ and all $i = 1, \ldots, n$. Since a B-spline basis consists of linearly independent functions, this implies that $C^\top \mathbf{s}_i = \tilde{C}^\top \tilde{\mathbf{s}}_i$ for all i. Assumptions 1 and 2 in the text imply that $CC^\top = \tilde{C}\tilde{C}^\top = I$, so that if $A \doteq C\tilde{C}^\top$, then $\mathbf{s}_i = A\tilde{\mathbf{s}}_i$ and $\tilde{\mathbf{s}}_i = A^\top \mathbf{s}_i$ for all i. Therefore $AA^\top \mathbf{s}_i = \mathbf{s}_i$ for all i and, since $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ contains q linearly independent vectors, $AA^\top = I$. Now we only have to prove that $A = I$, because this implies that $\mathbf{s}_i = \tilde{\mathbf{s}}_i$ for all i and consequently $C = \tilde{C}$, because C has full row rank.

Let us prove, then, that A must be the identity matrix. Since $a_{ij} = \mathbf{c}_i^\top \tilde{\mathbf{c}}_j$, all the elements of A are non-negative. However, if an off-diagonal element a_{ij} is strictly positive, then its mirror element a_{ji} must be zero. This is very easy to visualize with a sketch of the block structure of C and $\tilde{C}$. Formally, the proof goes as follows: if $a_{ij} = \mathbf{c}_i^\top \tilde{\mathbf{c}}_j > 0$, then either $K_i \leq \tilde{K}_j < K_{i+1}$ or $\tilde{K}_j \leq K_i < \tilde{K}_{j+1}$ because the blocks must overlap. But then the blocks $\tilde{\mathbf{c}}_i$ and $\mathbf{c}_j$ cannot overlap: if $K_i \leq \tilde{K}_j < K_{i+1}$ we have either $\tilde{K}_{i+1} < K_j$ when $i < j$ (because $\tilde{K}_i < \tilde{K}_j$ and $K_{i+1} \leq K_j$) or $K_{j+1} < \tilde{K}_i$ when $i > j$ (because $K_{j+1} \leq K_i$ and $\tilde{K}_j < \tilde{K}_i$); if $\tilde{K}_j \leq K_i < \tilde{K}_{j+1}$ we again have that either $\tilde{K}_{i+1} < K_j$ when $i < j$ or $K_{j+1} < \tilde{K}_i$ when $i > j$. Either way, the blocks $\tilde{\mathbf{c}}_i$ and $\mathbf{c}_j$ do not overlap and then $a_{ji} = 0$.

Hence A is orthogonal, with non-negative elements, and such that $a_{ij} > 0 \Rightarrow a_{ji} = 0$ for $i \neq j$. All this together imply that $A = I$. To prove this, let us proceed inductively. Write

$$A = \begin{bmatrix} A_1 & \mathbf{b} \\ \mathbf{c}^\top & d \end{bmatrix}$$

with A_1 the $(q-1) \times (q-1)$ upper-left block of A. Since $AA^\top = I$ we have that $A_1 A_1^\top + \mathbf{b}\mathbf{b}^\top = I$, $A_1\mathbf{c} + d\mathbf{b} = \mathbf{0}$ and $\mathbf{c}^\top\mathbf{c} + d^2 = 1$. But $A_1\mathbf{c}$ and $d\mathbf{b}$ have non-negative elements, then $A_1\mathbf{c} = d\mathbf{b} = \mathbf{0}$. Moreover, $d = \mathbf{c}_q^\top \tilde{\mathbf{c}}_q$ cannot be zero, so that $\mathbf{b} = \mathbf{0}$. This implies that $A_1 A_1^\top = I$, hence A_1 is full-rank and then $A_1\mathbf{c} = \mathbf{0}$ implies that $\mathbf{c} = 0$, which in turn implies that $d = 1$. In the next step A_1 plays the role of A and then, inductively, we deduce that $A = I$. 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., and Engel, J. (1995). Model estimation in nonlinear regression under shape invariance. *The Annals of Statistics,* **23**, 551-570.

[6] 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.

[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] Rønn, B. B. (2001). Nonparametric maximum likelihood estimation of shifted curves. *Journal of the Royal Statistical Society*, Series B, **63**, 243-259.

[11] Van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press, UK.

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