Nonparametric maximum likelihood estimation of the structural mean of a sample of curves

BY DANIEL GERVINI
University of Wisconsin–Milwaukee, 3200 N. Cramer St., Room E490,
Milwaukee, WI 53211, U.S.A.
gervini@uwm.edu

AND THEO GASSER
University of Zürich, Sumatrastrasse 30,
CH-8006 Zürich, Switzerland
tgasser@ifspm.unizh.ch

April 1, 2005

Summary

A random sample of curves can be usually thought of as noisy realisations of a compound stochastic process $X(t) = Z\{W(t)\}$, where $Z(t)$ produces random amplitude variation and $W(t)$ produces random dynamic or phase variation. In most applications it is more important to estimate the so-called structural mean $\mu(t) = E\{Z(t)\}$ than the cross-sectional mean $E\{X(t)\}$, but this estimation problem is difficult because the process $Z(t)$ is not directly observable. In this article we propose a nonparametric maximum likelihood estimator of $\mu(t)$. This estimator is shown to be $\sqrt{n}$-consistent and asymptotically normal under the model assumed and robust to model misspecification. Simulations and a real-data example show that the proposed estimator is competitive with landmark registration, often considered the benchmark, and has the advantage of avoiding time-consuming and often infeasible individual landmark identification.

Some key words: Curve registration; Functional data; Longitudinal data; Phase variation; Time warping.

1 Introduction

Multivariate datasets often consist of discrete observations of continuous curves. A classical example is the longitudinal analysis of growth data, where height or other variables are measured at different ages for a sample of individuals. Although the data consist of vectors, classical multivariate techniques that do not take into account the underlying smoothness of the curves are either very inefficient or must resort to strong model assumptions that are dubious in many applications. Ramsay & Silverman (1997, 2002) make a strong case for the nonparametric approach to the statistical analysis of samples of curves.

A characteristic feature of samples of curves, as opposed to samples of arbitrary vectors, is the presence of time variability. Figure 1(a) illustrates this problem well, showing four representative leg growth velocity curves for boys (vertically shifted for better visualisation). These are raw velocity curves, obtained directly from leg length data without smoothing. The peaks of maximal pubertal growth occur approximately at 14 years of age, but the exact timing and amplitude varies from person to person. The cross-sectional mean (Fig. 1(c)) is a poor estimate of the average growth velocity, not so much because of its roughness (it can be smoothed out with any of the well-known univeriate smoothing methods) but mainly because it grossly underestimates the average growth velocity at the pubertal peak, a direct consequence of time variability. Figure 1(b) shows the same sample curves smoothed and aligned using the techniques that will be introduced in this article, and Fig. 1(d) shows the resulting estimator of mean growth velocity.

Several methods to handle time variability have been proposed in recent years. The basic idea behind these methods is to align the curves so as to remove most of the time variability prior to averaging. The method generally considered the benchmark is landmark registration (Kneip & Gasser, 1992). The procedure consists of *(i)* identifying a set of salient features in all the curves, such as local extrema or zero crossings, *(ii)* monotonically transforming the curves so that the transformed landmarks of each curve coincide with the average landmarks, and *(iii)* computing the average of the aligned curves. The disadvantage of landmark registration is that, in many situations, a completely automated identification of landmarks is not possible. In such situations, the researcher must identify the landmarks curve by curve, which is infeasible for large datasets. For this reason, alternative methods have been sought. We can cite Silverman (1995), Ramsay & Li (1998), Kneip, Li, MacGibbon & Ramsay (2000), Wang & Gasser (1999), Rønn (2001) and Gervini & Gasser (2004). These methods differ greatly in terms of range of applicability, computational complexity and theoretical background.

To make the discussion more formal, let us assume that our sample curves are n independent realizations of a compound stochastic process $X(t) = Z\{W(t)\}$, where $Z(t)$ produces random amplitude variability and $W(t)$ produces random time variability, and it is therefore assumed to be monotone increasing with prob-

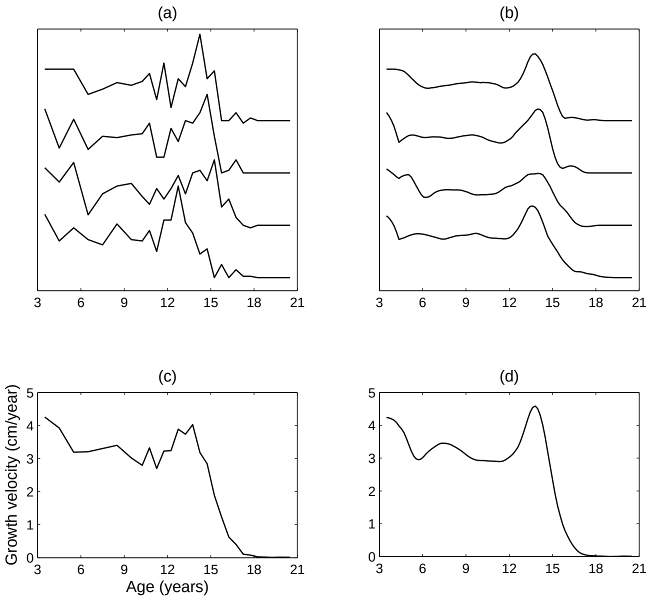

Figure 1: Leg growth velocity on boys. (a) Four representative raw sample curves and (b) the corresponding aligned and smoothed curves. (c) Cross-sectional sample mean and (d) nonparametric maximum likelihood estimator of the structural mean.

ability one. The researcher usually wants to estimate $\mu(t) = E\{Z(t)\}$, the so-called structural mean (Kneip & Gasser, 1992), rather than the cross-sectional mean $E\{X(t)\}$. Kneip & Engel (1995) showed that the mean of landmark-registered curves is a consistent estimator of μ under certain conditions. Consistency, however, holds under the assumption that the number of observations per individual, m, goes to infinity and the number of individuals n is held fixed; moreover, the rate of convergence depends on the smoothing bandwidth used to extract the individual landmarks. Consistency of the estimators of Wang & Gasser (1999) and Gervini & Gasser (2004) is also proved for m going to infinity and n

fixed. This type of asymptotics, however, is intrinsically inadequate in most applications, where individuals are random rather than fixed factors. On the other hand, the articles by Silverman (1995), Ramsay & Li (1998) and Kneip et al. (2000) do not provide any kind of consistency results for their estimators.

In contrast, the nonparametric maximum likelihood estimator of Rønn (2001) is $\sqrt{n}$-consistent and asymptotically normal as n goes to infinity and m is fixed (this has been proved by B. Rønn and I. Skovgaard in an unpublished technical report of the Royal Veterinary and Agricultural University, Frederiksberg). This is the relevant type of asymptotics when the number of random curves is large but the number of observations per curve is fixed and relatively small. The human growth data shown in Fig. 1 and analysed later in §7 are a typical example of this. Rønn's method was derived under the assumption that the warping process is a scalar random shift, which is too simplistic in most applications. The growth curves in Fig. 1(a), for example, have fixed endpoints and present two growth spurts (the midgrowth spurt about age 7 and the pubertal spurt about age 14) whose locations vary independently of each other; obviously, a single random shift for the whole curve is inadequate.

The idea of estimating μ by maximum likelihood, however, is appealing because it avoids individual identification of landmarks. Actually, it avoids estimation of individual parameters altogether, since individual random effects are integrated out rather than estimated. This is the reason why consistency of $\hat{\mu}$ as n goes to infinity is attainable. What we propose in this paper is nonparametric maximum likelihood estimation with more flexible families of warping functions. As by-products, we derive individual predictors for the warping functions $W_i(t)$ and the amplitude process $Z_i(t)$. Bootstrap methods to construct confidence bands for $\mu(t)$ are also proposed.

The article is organized as follows. The derivation of the maximum likelihood estimator and related issues are discussed in §§2 and 3. Consistency and asymptotic normality are established in §5, and bootstrap confidence bands are derived. A Monte Carlo study in §6 compares the performance of the proposed estimator with landmark registration and continuous monotone registration. An application to human growth is given in §7, and §8 briefly discusses the results and relevance of the proposed method.

2 The maximum likelihood estimator

As explained in §1, we will assume that the dataset $\{x_1, \ldots, x_n\}$, with $x_i \in \mathbb{R}^m$, consists of discrete and noisy realizations of stochastic processes $X_i : T \rightarrow \mathbb{R}$, with $T = [a, b]$, so that $x_{ij} = X_i(t_{ij}) + \varepsilon_{ij}$, where $\{\varepsilon_{ij}\}$ are independent and identically distributed random errors and $\{t_{i1}, \ldots, t_{im}\} \subset T$ is an input grid that may be different for each individual. The processes $X_1(t), \ldots, X_n(t)$ are assumed to be independent and identically distributed realizations of a stochastic process $X(t) = Z\{W(t)\}$. The warping process will be parametrically modelled as $W(t) = g(t, \theta)$, where g is a fixed, known function, monotone increasing in t for every θ. The parameter $\theta \in \mathbb{R}^p$ will be considered an unobservable random effect. Possible families of warping functions g and distributions for θ will be discussed in §3.

For the amplitude component a reasonable model would be

$$Z(t) = \mu(t) + \sum_{k=1}^{q} \xi_k \phi_k(t), \tag{1}$$

where μ and $\{\phi_k\}$ are fixed unknown functions, $\int_a^b \phi_k(t)\phi_l(t)\, dt = \delta_{kl}$, and $\{\xi_k\}$ are independent zero-mean random variables with finite variances. This is a truncation of the Karhunen–Loève decomposition $Z(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_k \phi_k(t)$, which holds for any square-integrable stochastic process. Unfortunately, simultaneous maximum likelihood estimation of μ and the components $\{\phi_k\}$ is very complicated. For simplicity, we will derive the maximum likelihood estimator of μ for a model without variance components. It is shown later, by simulations and example, that this maximum likelihood estimator provides good estimates of μ even under the general variance-component model (1).

As working model, then, let us assume the mean-plus-error model

$$x_{ij} = \mu\{g(t_{ij}, \theta_i)\} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \;\; \theta_i \text{ and } \varepsilon_{ij} \text{ independent.} \tag{2}$$

Under this model, $x_i|\theta_i \sim \mathcal{N}[\mu\{g(t_i^*, \theta_i)\}, \sigma^2 I_{m_i}]$, where $t_i^* = (t_{i1}, \dots, t_{im})^\top$ (here and in the rest of the paper, evaluation of a univariate function at a vector is understood in a componentwise sense). The log-likelihood function is

$$L(\mu, \sigma^2) = \sum_{i=1}^n \log \int f(x_i|\theta; \mu, \sigma^2) f(\theta)\, d\theta,$$

where $f(x_i|\theta; \mu, \sigma^2)$ denotes the conditional density of x_i given $\theta_i = \theta$ and $f(\theta)$ is the density of θ.

The estimating equation for $\hat{\sigma}^2$ is easy to derive. Since

$$\frac{\partial f(x_i|\theta_i; \mu, \sigma^2)}{\partial(\sigma^2)} = \Big[-\frac{m_i}{2\sigma^2} + \frac{\|x_i - \mu\{g(t_i^*, \theta_i)\}\|^2}{2(\sigma^2)^2} \Big] f(x_i|\theta_i; \mu, \sigma^2),$$

we have

$$\frac{\partial L(\mu, \sigma^2)}{\partial(\sigma^2)}$$
$$= \sum_{i=1}^n \frac{1}{f(x_i; \mu, \sigma^2)} \int \Big[-\frac{m_i}{2\sigma^2} + \frac{\|x_i - \mu\{g(t_i^*, \theta)\}\|^2}{2(\sigma^2)^2} \Big] f(x_i|\theta; \mu, \sigma^2) f(\theta)\, d\theta,$$

where $f(x_i; \mu, \sigma^2)$ is the marginal density of x_i. Since $\partial L(\hat{\mu}, \hat{\sigma}^2)/\partial(\sigma^2) = 0$, we obtain the following fixed-point expression:

$$\hat{\sigma}^2 = \frac{1}{\sum_{i=1}^n m_i} \sum_{i=1}^n \int \|x_i - \hat{\mu}\{g(t_i^*, \theta)\}\|^2 f(\theta|x_i; \hat{\mu}, \hat{\sigma}^2)\, d\theta. \tag{3}$$

An estimating equation for $\hat{\mu}$ can be derived using Gateaux differentials. Let $\mathcal{M}$ be the parametric space of μ, which is a linear subspace of $\mathcal{L}^\infty(T)$, the space of bounded measurable functions $T \to \mathbb{R}$. Since $\hat{\mu}$ maximizes $L(\mu, \hat{\sigma}^2)$ for all $\mu \in$

$\mathcal{M}$, the directional log-likelihood $L(\hat{\mu}+th,\hat{\sigma}^2)$ is maximized at $t=0$ for every $h\in\mathcal{M}$. Then

$$\frac{d}{dt}\,L(\hat{\mu}+th,\hat{\sigma}^2)\Big|_{t=0}=0,\ \text{for all } h\in\mathcal{M}. \tag{4}$$

After straightforward algebra we obtain

$$\begin{aligned}&\frac{d}{dt}\,L(\hat{\mu}+th,\hat{\sigma}^2)\Big|_{t=0}\\ =\;&\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}\int[x_i-\hat{\mu}\{g(t_i^*,\theta)\}]^\top h\{g(t_i^*,\theta)\}\,f(\theta|x_i;\hat{\mu},\hat{\sigma}^2)\,d\theta.\end{aligned}$$

Let $w_{ij}(\cdot|x_i;\mu,\sigma^2)$ be the conditional density of $g(t_{ij},\theta)$ given x_i. Then

$$\begin{aligned}&\int[x_i-\hat{\mu}\{g(t_i^*,\theta)\}]^\top h\{g(t_i^*,\theta)\}\,f(\theta|x_i;\hat{\mu},\hat{\sigma}^2)\,d\theta\\ =\;&\sum_{j=1}^{m_i}\int\{x_{ij}-\hat{\mu}(s)\}\,w_{ij}(s|x_i;\hat{\mu},\hat{\sigma}^2)h(s)ds,\end{aligned}$$

so (4) is equivalent to

$$\begin{aligned}\int k_n(s;\hat{\mu},\hat{\sigma}^2)h(s)\,ds\;&=\;0\ \text{ for all } h\in\mathcal{M}, \qquad (5)\\ \text{where }\ k_n(s;\mu,\sigma^2)\;&=\;\frac{1}{\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{x_{ij}-\mu(s)\}\,w_{ij}(s|x_i;\mu,\sigma^2).\end{aligned}$$

Estimating equations (3) and (5) are asymptotically unbiased, since for the true parameters (μ,σ^2) we have

$$\begin{aligned}&E[\{x_{ij}-\mu(s)\}w_{ij}(s|x_i;\mu,\sigma^2)]\\ =\;&\int\{x_j-\mu(s)\}w_{ij}(s|x;\mu,\sigma^2)f(x;\mu,\sigma^2)dx\\ =\;&f_{g(t_{ij},\theta)}(s)\int\{x_j-\mu(s)\}f_{x|g(t_{ij},\theta)}(x|s;\mu,\sigma^2)dx\\ =\;&0\text{ for all } s\in T,\end{aligned}$$

and

$$E(E[\|x_i - \mu\{g(t_i^*, \theta)\}\|^2 | x_i]) = E(E[\|x_i - \mu\{g(t_i^*, \theta)\}\|^2 | \theta]) = m_i \sigma^2.$$

The estimating equation (5) for $\hat{\mu}$ is not very useful as it is. But for certain parametric spaces it is possible to derive more explicit equations. For example, if $\mathcal{M}$ is the space of continuous functions $= \mathcal{C}(T)$, then (5) holds if and only if $k_n(s; \hat{\mu}, \hat{\sigma}^2) = 0$ almost everywhere in T, which implies that

$$\hat{\mu}(s) = \frac{\sum_{i=1}^n \sum_{j=1}^{m_i} x_{ij} w_{ij}(s|x_i; \hat{\mu}, \hat{\sigma}^2)}{\sum_{i=1}^n \sum_{j=1}^{m_i} w_{ij}(s|x_i; \hat{\mu}, \hat{\sigma}^2)} \text{ almost everywhere in } T. \tag{6}$$

The space $\mathcal{C}(T)$, however, is too large to provide reasonably smooth estimates of μ. Moreover, there is no guarantee that $L(\mu, \hat{\sigma}^2)$ attains a maximum in $\mathcal{C}(T)$. For practical and theoretical reasons, discussed in §5 below and in the technical report by Rønn and Skovgaard mentioned earlier, $\mathcal{M}$ must be restricted to be a compact subspace of $\mathcal{C}(T)$. Since $L(\mu, \hat{\sigma}^2)$ is continuous in μ, it attains a maximum in any compact space $\mathcal{M}$.

Equation (6) has an interesting intuitive interpretation. At each s, $\hat{\mu}(t)$ is a weighted average of $\{x_{ij}\}$, where $w_{ij}(s|x_i; \hat{\mu}, \hat{\sigma}^2)$ puts more weight on those x_{ij}s for which $g(t_{ij}, \theta)$ is expected to be close to s and x_{ij} is not far from its current expected value $\hat{\mu}\{g(t_{ij}, \theta)\}$. Thus, (6) provides a sort of automatic curve alingment and smoothing, since the weights are smooth functions of s. Note that, following this intuition, we can use

$$\hat{Z}_i(t) = \frac{\sum_{j=1}^{m_i} x_{ij} w_{ij}(t|x_i; \hat{\mu}, \hat{\sigma}^2)}{\sum_{j=1}^{m_i} w_{ij}(t|x_i; \hat{\mu}, \hat{\sigma}^2)}. \tag{7}$$

as estimates of the registered curves $Z_i(t) = X_i\{W_i^{-1}(t)\}$.

Estimating equations similar to (5) can be derived for the mean and the components of the general variance-component model (1). However, explicit estimating equations like (6) or even (5) cannot be obtained, so the estimators have to be numerically computed using, for instance, a rather cumbersome functional Newton-Raphson method (Luenberger, 1969, ch. 10). A more practical compromise is to estimate μ with the maximum likelihood estimator of the mean-plus-error model derived above, and then estimate the factors $\{\phi_k\}$ using the principal components of the registered functions $\{\hat{Z}_i\}$, as in chapter 6 of Ramsay & Silverman (1997).

Estimation of the individual effects $\{\theta_i\}$ may also be of interest in some situations, as we will see in §7. This can be done with the conditional expectation $E(\theta|x_i; \hat{\mu}, \hat{\sigma}^2)$ or with the conditional mode $\arg\max f(\theta|x_i; \hat{\mu}, \hat{\sigma}^2)$. The conditional mode estimator has an interesting interpretation: since

$$\arg\max f(\theta|x_i; \hat{\mu}, \hat{\sigma}^2) = \arg\min \left[\frac{1}{2\hat{\sigma}^2} \sum_{j=1}^{m_i} \{x_{ij} - \hat{\mu}(g(t_{ij}, \theta))\}^2 - \log\{f(\theta)\} \right], \tag{8}$$

this is a penalized least squares estimator, with a penalty term $-\log\{f(\theta)\}$ that shrinks $\hat{\theta}_i$ towards the mode of $f(\theta)$. Except for the penalty term, this is similar to the Procrustes registration method of Silverman (1995), which minimizes the sum of squares with respect to *both* parameters μ and θ. Procrustes registration has a tendency to “overwarp” the data, producing deformed estimates of μ (see example on p. 113 of Ramsay & Silverman, 2002). To some extent this is a problem of all registration methods that estimate μ simultaneously with the individual effects. Our method avoids this problem by estimating μ independently of the θ_is. In (5) the random effect θ is integrated out, rather than estimated.

3 Warping model

In §2 we derived the maximum likelihood $\hat{\mu}$ for a generic warping function $g(t, \theta)$ and a generic distribution $f(\theta)$ of the random effect. For a successful practical implementation of the maximum likelihood estimator, it is important to specify a warping model that is versatile enough but does not have too many parameters (note that equations (3) and (5) involve multidimensional integrals in θ).

On possibility is to take $g(t,\theta)$ as a linear combination of I-splines with fixed knots (Ramsay, 1988), where θ is the vector of basis coefficients. Another possibility is to take $g(t,\theta) = a + (b-a)c(t,\theta)/c(b,\theta)$, where $c(t,\theta) = \int_a^t e^{w(s,\theta)}ds$, $w(s,\theta)$ is a linear combination of B-splines with fixed knots and θ is again the vector of basis coefficients (Ramsay, 1998). The problem with both of these models is that it is unclear what a reasonable distribution for θ might be. Moreover, it may be necessary to use a large number of basis functions to obtain enough model flexibility, which complicates the computation of $\hat{\mu}$.

What we propose is to take θ as a vector of knots, rather than basis coefficients. Intuitively, we may think of θ as a vector of "hidden landmarks". With this interpretation, a reasonable family of distributions for θ is the truncated normal

$$f(\theta) \propto \prod_{k=1}^{p} \frac{1}{\tau_k} \varphi\left(\frac{\theta_k - \theta_{0k}}{\tau_k}\right) \mathbb{I}\{a < \theta_1 < \ldots < \theta_p < b\}, \tag{9}$$

with $\theta_{01} < \ldots < \theta_{0p}$. The θ_{0k}s can be associated with salient features of μ, such as peaks and troughs, which in many practical situations will provide a good fit even with a small dimension p. For example, for the growth curves in Fig. 1 we use a two-dimensional θ, where each coordinate is associated with a growth spurt. The actual form of $g(t,\theta)$ is not so important as long as the monotonicity in t is ensured and the following identifiability conditions are satisfied: $g(a,\theta) = a$, $g(b,\theta) = b$, and $g(\theta_k,\theta) = \theta_{0k}$ for $k = 1,\ldots,p$. In §§6 and 7 we use shape-preserving cubic polynomial interpolation (Fritsch & Carlson, 1980), as implemented in the Matlab function `pchip`.

The unknown parameters θ_0 and τ could, in principle, be incorporated in the likelihood function and estimated together with μ and σ^2. But in practice this is very time consuming. A workable alternative is to try out with a few values of θ_0 and τ suggested by visual inspection of the data, and keep those with largest likelihood. In our experience, this approach works well in practice because the maximum likelihood estimator is robust to misspecification of $f(\theta)$ (see §§6 and 7).

A more refined way to determine p, θ_0 and τ is by means of the "structural intensity" method of Gasser & Kneip (1995). This method consists in computing a nonparametric density estimator of the number of local maxima; the modes of the density reveal the most important landmarks and their approximate distribution. This method only requires identification of the *set* of local maxima for each curve, which can be done in a fully automated way, as opposed to landmark registration, that requires *specific* identification of local maxima and usually cannot be carried out without human interaction. For example, a given growth curve may have more than just two local maxima, due to undersmoothing or otherwise; while landmark registration requires precise identification of the growth spurts among these local maxima, the structural intensity method only requires identification of all the local maxima.

4 COMPUTATIONAL ASPECTS

The estimators $\hat{\mu}$ and $\hat{\sigma}^2$ can be iteratively computed using the fixed-point expressions (3) and (6). As initial estimators, the simplest choices are $\hat{\sigma}^{2(0)} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} (x_{ij} - \bar{x})^2 / nm$ and $\hat{\mu}^{(0)}(t) = \sum_{i=1}^{n} \tilde{x}_i(t)/n$, where $\tilde{x}_i(t)$ is obtained from $x_{i1}, \ldots, x_{im_i}$ by interpolation (we use piecewise cubic interpolation). A potential problem of using equation (6) to update the estimate of μ is that the algorithm may not converge; remember that estimating equation (5) is always satisfied by $\hat{\mu}$, but this is not necessarily the case with equation (6). There are algorithms with guaranteed convergence, such as the steepest ascent algorithm (Luenberger 1969, ch. 10), that defines $\hat{\mu}^{(k)}(t) := \hat{\mu}^{(k-1)}(t) + \alpha_k k_n(t; \hat{\mu}^{(k-1)}, \hat{\sigma}^{2(k-1)})$ where the step α_k is chosen to maximise the likelihood function in the direction of $k_n(s; \hat{\mu}^{(k-1)}, \hat{\sigma}^{2(k-1)})$. Finding the optimal step α_k, however, involves many recomputations of the likelihood function and is very time consuming. We think it is more practical to use the reweighting algorithm suggested by equation (6). When this algorithm converges, it finds a solution of (5) and thus a stationary point of the log-likelihood function. We have used this algorithm for all simulations and data analyses in this paper and have not found any convergence problems.

The hardest part to implement efficiently is the computation of the p-dimensional integrals involved in equations (3) and (6). We use Monte Carlo integration: a random sample $\{\theta^{(1)}, \ldots, \theta^{(N)}\}$ is generated from $f(\theta)$ and, for instance, $f(x_i; \hat{\mu}, \hat{\sigma}^2) = \int f(x_i|\theta; \hat{\mu}, \hat{\sigma}^2) f(\theta) d\theta$ is approximated by $\hat{f}^{(N)}(x_i) := \sum_{l=1}^{N} f(x_i|\theta^{(l)}; \hat{\mu}, \hat{\sigma}^2)/K$. The other integrals are estimated in a similar way.

The computation of $w_{ij}(s|x_i; \hat{\mu}, \hat{\sigma}^2)$, on the other hand, requires a more careful approach. Since

$$\begin{aligned} w_{ij}(s|x_i; \hat{\mu}, \hat{\sigma}^2) &= \frac{d}{ds} \int \mathbb{I}\{g(t_{ij}, \theta) \leq s\} f(\theta|x_i; \hat{\mu}, \hat{\sigma}^2) d\theta \\ &= \frac{d}{ds} \int \mathbb{I}\{g(t_{ij}, \theta) \leq s\} \, \frac{f(x_i|\theta; \hat{\mu}, \hat{\sigma}^2)}{f(x_i; \hat{\mu}, \hat{\sigma}^2)} \, f(\theta) \, d\theta, \end{aligned}$$

we use a kernel-smoothed Monte Carlo integral:

$$\hat{w}_{ij}^{(N,\lambda)}(s|x_i) := \frac{1}{N} \sum_{l=1}^{N} \frac{1}{\lambda} \, K\left(\frac{g(t_{ij}, \theta^{(l)}) - s}{\lambda} \right) \frac{f(x_i|\theta^{(l)}; \hat{\mu}, \hat{\sigma}^2)}{\hat{f}^{(N)}(x_i)}. \tag{10}$$

As $K(t)$ we take the Epanechnikov kernel $K(t) = .75(1 - t^2)\mathbb{I}\{|t| \leq 1\}$, and as tentative λ we take the average oversmoothing bandwidth (Wand & Jones, 1995, p. 61),

$$\lambda = \left\{ \frac{243 c_1}{35 c_2^2 N} \right\}^{1/5} \frac{1}{\sum_{i=1}^{n} m_i} \sum_{i=1}^{n} \sum_{j=1}^{m_i} s_{ij}, \tag{11}$$

where s_{ij} is the sample standard deviation of $\{g(t_{ij}, \theta^{(1)}), \ldots, g(t_{ij}, \theta^{(N)})\}$, $c_1 = \int K^2(x) dx$ and $c_2 = \int x^2 K(x) dx$. In particular, for the Epanechnikov kernel we have $c_1 = 3/5$ and $c_2 = 1/5$.

At this point it is important to remark that the maximum likelihood estimator itself does not depend on any bandwidths. If a closed expression for $w_{ij}(s|x_i; \mu, \sigma^2)$ existed, it would not be necessary to use the smoother $\hat{w}_{ij}^{(N,\lambda)}(s|x_i)$. It turns out, however, that the conditional density w_{ij} never has a closed expression in practice, and must be estimated. The choice of λ will determine the smoothness of

$\hat{w}_{ij}^{(N,\lambda)}(s|x_i)$ and this, in turn, will determine the smoothness of $\hat{\mu}$. In our experience, (11) provides a reasonable bandwidth or at least a good initial guess; a plot of $\hat{\mu}$ will clearly tell the user when a smaller or a larger bandwidth is advisable.

5 Asymptotics and inference

In this section we prove that the maximum likelihood estimator of μ is consistent and asymptotically normal as the number of curves n goes to infinity. Let us assume that $\{x_1, \ldots, x_n\}$ are independent and identically distributed, so $m_i = m$ and the input grids $\{t_{i1}, \ldots, t_{im}\}$ are the same for all i. For simplicity of notation, we will also assume that the error variance σ^2 is known, but it is clear that Theorem 1 can be extended to simultaneous estimation of μ and σ^2 in a straightforward manner.

Given $x \in \mathbb{R}^m$, let $\ell_x(\mu) = \log f(x;\mu)$. The maximum likelihood estimator $\hat{\mu}$ maximises $L_n(\mu) := E_n\{\ell_x(\mu)\}$, where E_n denotes expectation with respect to the empirical measure. Then $\hat{\mu}$ always exists if $\mathcal{M}$ is compact, because $L_n(\mu)$ is continuous (it is Fréchet differentiable, as shown in Theorem 2 in the Appendix, and Fréchet differentiability implies continuity; see Luenberger, 1969, p. 173). Let $L_0(\mu) := E_0\{\ell_x(\mu)\}$ be the asymptotic log-likelihood function, where E_0 is the expectation under the mean-plus-error model (2) with parameter μ_0. If model (2) is identifiable, then μ_0 is the unique maximiser of $L_0(\mu)$ (see the proof of Theorem 1(i)). The assumption that model (2) is identifiable is clearly necessary for consistency. A simple modification of the identifiability proof of Gervini and Gasser (2004) shows that model (2) is identifiable provided μ is piecewise strictly monotone (i.e., without “flat” parts) and the warping model $g(t,\theta)$ is identifiable (the cubic spline model proposed in §3 is). For a precise definition of the supremum norm, tensor product and covariance functional used in the following theorem, see the Appendix.

Theorem 1 *If $\mathcal{M}$ is compact and model (2) is identifiable, then:*

(i) *(Strong consistency).* $P\{\lim_{n\to\infty} \|\hat{\mu} - \mu_0\| = 0\} = 1.$

(ii) *(Asymptotic normality). Let* $\mathcal{I} := E_0\{\mathsf{D}\ell_x(\mu_0) \otimes \mathsf{D}\ell_x(\mu_0)\}$*, which coincides with* $-E_0\{\mathsf{D}^2\ell_x(\mu_0)\}$ *under model (2). Then* $\sqrt{n}(\hat{\mu} - \mu_0)$ *converges in distribution to a Gaussian process with mean zero and covariance functional* $\mathcal{I}^{-1}$ *in the space* $(\mathcal{M}, \|\cdot\|)$*.*

The strong consistency of $\hat{\mu}$ in the supremum norm implies that $\hat{\mu}(t) \to \mu_0(t)$ almost surely for all $t \in T$. The asymptotic normality of $\sqrt{n}(\hat{\mu} - \mu_0)$ as a stochastic element of $(\mathcal{M}, \|\cdot\|)$ implies that the finite-dimensional projections are asymptotically Normal in the classical multivariate sense. That is: given an arbitrary vector $t^* = (t_1, \ldots, t_M)$, $\sqrt{n}\{\hat{\mu}(t^*) - \mu_0(t^*)\}$ converges in distribution to an M-variate Normal distribution with covariance matrix explicitly computable from the covariance functional $\mathcal{I}$. In principle, asymptotic confidence bands for μ_0 could be derived from this result, but we have found that such bands tend to be too narrow in practice, having finite-sample coverage levels much smaller than the nominal asymptotic levels. For that reason, we see Theorem 1 mainly as a qualitative result that shows that the nonparametric maximum likelihood estimator is able to attain the parametric consistency rate $n^{-1/2}$.

Although Theorem 1, as given above, applies only to the maximum likelihood estimator for the mean-plus-error model (2), a similar result can be obtained for the maximum likelihood estimator of the general variance-component model (1). However, for the reasons indicated in §2, we think that the latter estimator is impractical. The simulations and the example in §§6 and 7 show that the maximum likelihood estimator of the mean-plus-error model does not have a much larger bias under the general variance-component model, so that bootstrap methods based on this estimator can be used for inference under the more general model.

To construct confidence bands for μ we propose two bootstrap procedures. The simplest one is based on the so-called wild bootstrap: take B bootstrap samples $\{x_1^*, \ldots, x_n^*\}$ from the sample $\{x_1, \ldots, x_n\}$, find the corresponding maximum likelihood estimators $\{\hat{\mu}_1^*(t), \ldots, \hat{\mu}_B^*(t)\}$, and construct confidence bands for μ using the empirical percentiles of this sample.

The second method, that we call model-based bootstrap, is as follows:

Step 1. Find the maximum likelihood estimators $\hat{\mu}$ and $\hat{\sigma}^2$, individual predictors $\{\hat{\theta}_i\}$ and registered curves $\{\hat{Z}_i(t)\}$ (as defined in (7)).

Step 2. Using the spectral decomposition of the covariance matrix of $\{\hat{Z}_i(t)\}$, find estimates of the components $\{\hat{\phi}_k\}$ and the individual scores $\{\hat{\xi}_{ik}\}$ of the variance-component model (1), and choose the number of components q. Define the residuals

$$\hat{\varepsilon}_{ij} = x_{ij} - \hat{\mu}\{g(t_j, \hat{\theta}_i)\} - \sum_{k=1}^{q} \hat{\xi}_{ik}\hat{\phi}_k\{g(t_j, \hat{\theta}_i)\}.$$

Step 3. Repeat B times:

a. Take independent bootstrap samples $\{\hat{\theta}_i^*\}$ from $\{\hat{\theta}_i\}$, $\{(\hat{\xi}_{i1}^*, \ldots, \hat{\xi}_{iq}^*)\}$ from $\{(\hat{\xi}_{i1}, \ldots, \hat{\xi}_{iq})\}$, and $\{\hat{\varepsilon}_{ij}^*\}$ from $\{\hat{\varepsilon}_{ij}\}$. Define the pseudo-observations

$$x_{ij}^* = \hat{\mu}\{g(t_j, \hat{\theta}_i^*)\} + \sum_{k=1}^{q} \hat{\xi}_{ik}^*\hat{\phi}_k\{g(t_j, \hat{\theta}_i^*)\} + \hat{\varepsilon}_{ij}^*.$$

b. Find the maximum likelihood estimator $\hat{\mu}^*$ for the bootstrapped dataset $\{x_1^*, \ldots, x_n^*\}$.

Step 4. Construct confidence bands for μ using the empirical percentiles of $\{\hat{\mu}_1^*(t), \ldots, \hat{\mu}_B^*(t)\}$.

6 SIMULATIONS

In this section we study by simulation the finite-sample performance of the maximum likelihood estimator and of the bootstrap confidence bands. In particular, we compare the performance of maximum likelihood registration with the

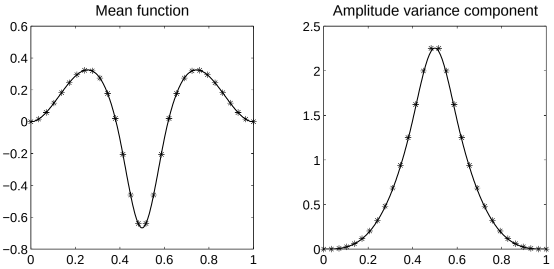

Figure 2: Mean function $\mu(t)$ and amplitude variance component $\phi_1(t)$ of simulated models. Asterisks indicate function values at input grid points.

two most commonly used methods, landmark registration and continuous monotone registration (Ramsay & Li, 1998), under different models of amplitude and time variability.

The variance-component model (1) is very general and offers a huge number of interesting sampling situations to consider. But our goal here is not to exhaust all possible scenarios, but to focus on a few non-trivial situations where dealing with time variability is problematic and the advantages or disadvantages of different methods are easy to see.

Therefore, as structural mean $\mu(t)$ we took a function with three prominent landmarks: two peaks and a trough (Fig. 2). Specifically, $\mu(t) = \beta_3(t) - \beta_4(t) + \beta_5(t)$, where $\beta_1(t), \ldots, \beta_7(t)$ are the cubic B-spline basis functions in $[0, 1]$ with knots $\{0.4, 0.5, 0.6\}$. Samples were generated from the mean-plus-error model (2) and also from the variance-component model (1) with $q = 1$ and $\phi_1(t) = \beta_4(t)$, shown in Fig. 2. As input grid we took $m = 30$ equispaced points in $[0, 1]$. The component ϕ_1 was standardized so that $\sum_{j=1}^{m} \phi_1^2(t_j)/m = 1$. The component scores $\{\xi_{1i}\}$ followed a $\mathcal{N}(0, \lambda_1)$ distribution with $\lambda_1 = 0.75 \times 0.10^2$. The errors

$\{\varepsilon_{ij}\}$ had a $\mathcal{N}(0, \sigma^2)$ distribution with $\sigma = 0.10$ for the mean-plus-error model and $\sigma = 0.10\sqrt{0.25}$ for the variance-component model. Note that both models have the same overall amplitude variance $E\{\sum_{j=1}^{m} Z^2(t_j)\} = 0.10^2 m$, only that differently split between systematic amplitude variability and random noise. For the variance-component model, 75% of the amplitude variance is associated with ϕ_1. As warping model we took a piecewise cubic monotone function $g(t, \theta)$ with θ following a truncated normal distribution, with parameters $\theta_0 = (0.25, 0.50, 0.75)$ (corresponding to the peaks and the trough of μ) and $\tau = 0.05 \times (1, 1, 1)$.

The maximum likelihood estimator was computed with the algorithm described in §4, which implicitly assumes that the mean-plus-error model is the correct one. For $g(t, \theta)$ we used the correct model and also two misspecified models: $\theta_0 = (0.25, 0.75)$ and $\theta_0 = 0.50$ (the scale parameters were always equal to 0.05). Explicitly, we considered the following situation:

S1. The data is generated from the mean-plus-error model, so the maximum likelihood estimator assumes the right amplitude variability and time warping models.

S2. The data is generated from the variance-component model. The maximum likelihood estimator assumes the right warping model but the amplitude variability model is misspecified.

S3. The data is generated from the mean-plus error model. The maximum likelihood estimator assumes the right amplitude variability model but a misspecified two-landmark model (corresponding to the peaks) is used.

S4. The data is generated from the mean-plus error model. The maximum likelihood estimator assumes the right amplitude variability model but a misspecified one-landmark model (corresponding to the trough) is used.

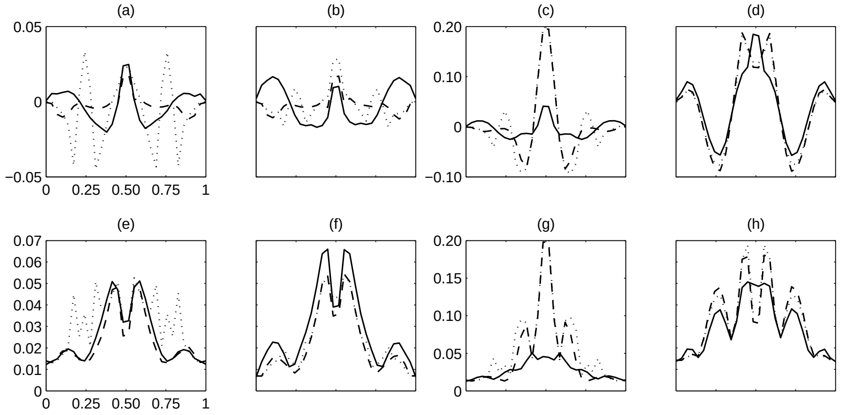

Figure 3: Bias (a-d) and root mean square error (e-h) of maximum likelihood estimation (solid line), oracle landmark registration (dashed line) and raw landmark registration (dotted line) for sampling situations described in the text (S1-S4, from left to right). Note the different scales for the vertical axes of (c) and (g).

6.1 Comparison with landmark registration

Proper landmark registration is impossible to simulate in this setup, since it requires individual smoothing of the curves and careful identification of the landmarks, which cannot be done in a fully automated way. Therefore, we had to consider two simplified procedures: one that we denominate “oracle” landmark registration and uses the actual realizations of θ as landmarks, and one that we denominate “raw” landmark registration and uses the two local maxima and the minimum of the LOWESS smoother of the curves as landmarks. A properly implemented landmark registration will show an intermediate behavior between these two simplified methods. For sampling situations S3 and S4 we also considered two misspecified warping models, one that takes only the peaks as landmarks and one that takes only the trough as landmark, respectively.

Each sampling situation was replicated 1000 times. As sample sizes we took $n = 50$ and $n = 100$, but the results were qualitatively similar, so we only report results for $n = 50$. Simulated biases and root mean squared errors (as functions of t) are shown in Fig. 3. The performance of the maximum likelihood estimator when both amplitude and warping components are well specified (sampling situation S1) is comparable to that of oracle landmark registration. Misspecifying amplitude variability (sampling situation S2) increases the variance of the maximum likelihood estimator but not the bias; in fact, the bias at the trough is smaller here than for sampling situation S1.

The robustness of the maximum likelihood estimator to underspecification of landmarks is remarkable. As Fig. 3(c) shows, the bias of landmark registration at the trough is four times as large as the bias of the maximum likelihood estimator; it is practically as large as the bias of the cross-sectional mean (not shown). This behavior has a simple explanation: the maximum likelihood estimator minimises a lack-of-fit criterion that, although not optimal for situations S2 to S4, still penalizes large shape deviations from the structural mean. On the other hand, landmark registration relies solely on the specified landmarks; since no lack-of-fit criterion is minimised, the method cannot “make up” for missing landmarks, even when it is plain to see that the shape of the resulting estimator is not representative of the sample curves, as in situation S3.

6.2 Comparison with continuous monotone registration

Continuous monotone registration was proposed by Ramsay & Li (1998) as a fully nonparametric alternative to landmark registration. This method does not require identification (or even existence) of landmarks. To compare maximum likelihood estimation with continuous monotone registration, then, we simulated the data using a warping model that is not associated with any landmarks. We took $g(t, \theta)$ such that

$$\frac{\partial^2 \log\{g^{-1}(t,\theta)\}}{\partial t^2} = \sum_{k=1}^{5} c_k \beta_k(t),$$

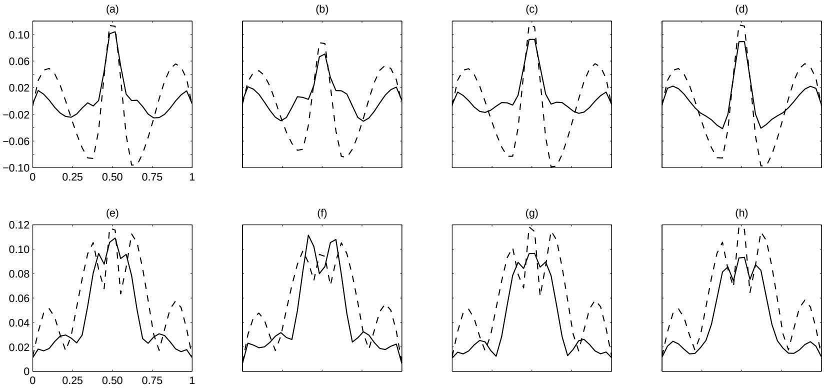

Figure 4: Bias (a-d) and root mean square error (e-h) of maximum likelihood estimation (solid line) and continuous monotone registration (dashed line) for sampling situations described in the text (S1-S4, from left to right).

with $\{\beta_k(t)\}$ cubic B-spline basis functions with equispaced knots in $[0,1]$ and $\{c_k\}$ independent and identically distributed coefficients following a $\mathcal{U}(-1,1)$ distribution. The same B-spline basis was used for registration, to avoid roughness penalization of the warping functions and the consequent problem of choosing the smoothing parameter, which is excessively time consuming for this method. We used the software provided by James Ramsay in his website.

The raw data was generated from the same mean-plus error model and a variance-component model as in §6.1. But since continuous monotone registration cannot be applied to raw data, we pre-smoothed the observations using penalized regression splines with 20 cubic B-spline basis functions with equispaced knots, choosing the smoothing parameter by generalized cross-validation. The maximum likelihood estimator was computed on discretizations of these smooth curves on a grid of $m = 30$ equispaced points. As regards the warping models assumed for maximum likelihood registration, we considered again the four situations S1-S4 as in §6.1. Each sampling situation was replicated 1000 times, with $n = 50$ curves per sample.

The results are summarized in Fig. 4. Biases and mean squared errors were

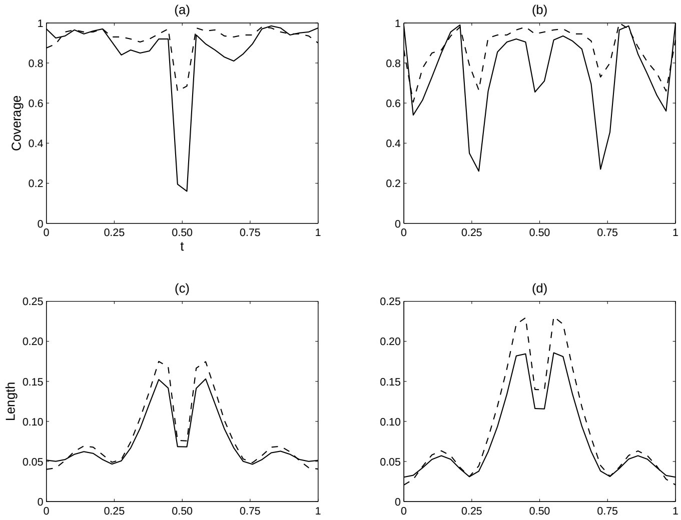

Figure 5: Pointwise coverage level and mean length of model-based (solid line) and wild (dashed line) bootstrapped confidence bands for mean-plus-error model (a,c) and variance-component model (b,d).

computed with 10% trimmed means because a few samples produced very outlying estimates for the continuous monotone registration method. We see that it makes little difference which warping model is used for the maximum likelihood estimator. This method outperforms continuous monotone registration in all situations. In particular, maximum likelihood provides much more acurate estimation than continuous monotone registration at the peaks. The explanation for this behavior is that continuous monotone registration minimizes a criterion that penalizes misalignment at the trough much more strongly than misalignment at the peaks. In contrast, maximum likelihood estimation explicitly penalizes misalignment at the peaks in sampling situations S1 to S3, thus providing better estimates even when the assumed warping models were not the true ones.

6.3 Confidence band coverage

Finally, we ran some simulations to evaluate the finite-sample accuracy of the bootstrapped confidence intervals proposed in §5. Two hundred datasets were generated for sampling scenarios S1 and S2, with $n = 50$ and $m = 30$. Five hundred bootstrap samples were taken for each dataset and confidence bands of nominal level 90% were constructed. Fig. 5 shows the simulated coverage levels and average lengths. As expected, the coverage level deteriorates at the peaks and the trough, as it usually happens with nonparametric estimators. Wild bootstrap intervals show a more stable coverage level around the nominal value, and while they tend to be wider, in the present situation they seem to be preferable over model-based bootstrap bands. In models with more variance components, however, model-based bootstrap may produce wider confidence bands and have better coverage, as in the example shown in §7.

7 Application: analysis of human growth data

The First Zurich Growth Longitudinal Study produced a large number of datasets, consisting of measurements of different parts of the body taken from birth to adulthood. One of the goals of the researchers was to estimate the mean growth velocity curve, in order to characterize the growth spurts. Gasser et al. (1991) estimated individual velocity and acceleration curves using Gasser–Müller kernel smoothers and computed landmark registration means using eight landmarks (the four zero crossings and the four local extrema observable in typical acceleration curves).

Here we will analyze leg growth velocity from 3 to 21 years of age. We chose leg measurements because these curves have prominent mid-growth spurts, in addition to the well-known pubertal spurt. For girls, both spurts occur in close succession and are roughly of the same size, complicating the registration process. The observed data consisted of leg length measurements taken annually from 3 to 9 years and biannually from then on. From these measurements we computed raw

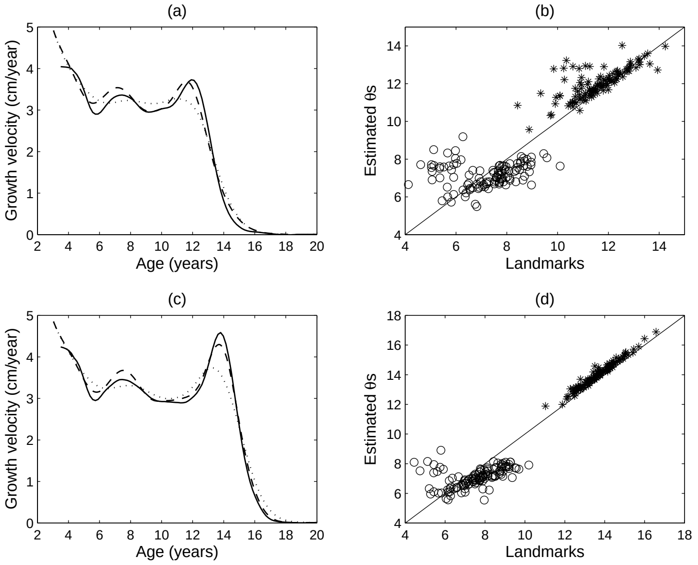

Figure 6: Leg growth velocity on children. Estimators of mean growth velocity for girls (a) and boys (c) obtained by nonparametric maximum likelihood (solid line), landmark registration (dashed line) and cross-sectional mean (dotted line). Scatter plots of midgrowth-spurt location versus $\hat{\theta}_1$ (circles) and pubertal-spurt location versus $\hat{\theta}_2$ (asterisks) for girls (b) and boys (d).

velocities by finite differenciation, taking the midpoints of age intervals as input grid. This yields a total of $m = 29$ observations per person, for 112 girls and 120 boys.

The maximum likelihood estimator was computed using a two-dimensional warping model. Implicitly, we are interpreting θ as the location of the growth spurts. We tried several values of θ_0 and τ and chose those that maximised the log-likelihood function: for girls, $\theta_0 = (7, 12)$ and $\tau = (1, 1)$; for boys, $\theta_0 = (7, 14)$ and $\tau = (1, 1)$. The maximum likelihood estimator was computed on an output grid of 100 equispaced points between 3.5 and 20.5 years.

For comparison, we also computed landmark registration means of smooth velocities, using the growth spurts as landmarks. The comparison is somewhat unfair with the maximum likelihood estimator, which is computed on much noisier

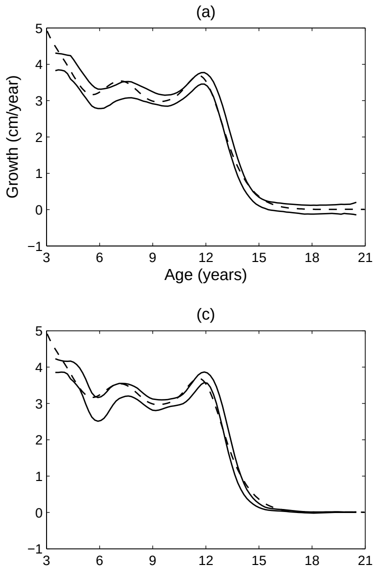

Figure 7: Confidence bands of level 90% (solid lines) and landmark registration means (dashed lines) of leg growth velocities for girls (a,c) and boys (b,d). Confidence bands obtained by model-based bootstrap (a,b) and wild bootstrap (c,d).

raw velocities. Nevertheless, we can see in Fig. 6 that the maximum likelihood estimator is very close to the landmark registration mean for both sexes and, in particular, the midgrowth spurts are well determined.

Fig. 6 also plots individual predictors $\{\hat{\theta}_{i1}\}$ and $\{\hat{\theta}_{i2}\}$ against growth spurt locations. The strong association observed supports our interpretation of the random effects θ_i as "hidden landmarks".

We also obtained 90% confidence bands for the means, based on 1000 bootstrap samples for each method and each sex. As explained in §5, to apply model-based bootstrap the number of variance components in model has to be determined. For each sex we estimated the 50 leading components and their variances, obtaining that the first six components explain, respectively, 23, 16, 15, 14, 9 and 7 percent of the total variance for girls, and 24, 15, 13, 11, 8 and 6 percent of the total variance for boys. We chose $q = 4$ for both sexes, discarding those components that explain less than 10% of the amplitude variance. The resulting confidence bands are shown in Fig. 7 together with landmark registration means, which can be seen as the "true means" in this example. We observe that model-based bootstrap produces somewhat wider confidence bands than wild bootstrap and has better coverage. Among other things, a useful inference that can be drawn from the confidence bands is that the midgrowth spurt is a real structural feature of the growth process and not just an artifact of undersmoothing. Since most classical parametric models miss the growth spurt, its actual existence was debated in the early 80's, when it was first detected and characterized by nonparametric methods.

8 Discussion

The registration method proposed in this paper has a number of advantages over existing methodology. Compared with landmark registration, maximum likelihood does not require tiresome and error-prone individual landmark identification, and it is more robust to underspecification of the number of landmarks. Compared with continuous monotone registration, it does not require presmoothing of the data and from a computational point of view it is considerably less time consuming and easy to implement. As we see it, maximum likelihood registration combines appealing properties of the other two methods: like continuous monotone registration, it minimizes a lack-of-fit criterion and is thus robust to misspecification of the warping model; like landmark registration, it explicitly models time variability at the salient features of the curves, which makes the warping model flexible and parsimonious at the same time.

As far as we can extrapolate from the simulations and the example analysed in this paper, the proposed method is competitive with landmark registration and better than continuous monotone registration in many non-trivial situations. From a theoretical point of view, it is one of the few methods with proved $\sqrt{n}$-consistency and asymptotic normality as the number of curves n goes to infinity, at least when the model is well specified.

We foresee a number of extensions and modifications of this method that can be better suited for some particular situations. For instance, when the number of observations per curve is large and the data very noisy, it may be worth considering spline models for μ, rather than the full nonparametric approach of this paper. This will reduce the estimation problem to a more manageable finite dimensional optimization, and simultaneous estimation of the mean and the variance components may be less cumbersome. Of course, this would also introduce the problem of knot placement and selection, or roughness penalization and selection of smoothing parameters, so more reasearch is needed before we can make claims about the relative merits of each approach.

We also think that this method can be extended to fields of applications that require more complex warping models, such as image alignment, more easily than other registration methods. This is currently being investigated by the authors.

Acknowledgment

This work was supported in part by grants of the Swiss National Science Foundation.

Appendix

Tecnical details on asymptotic results

In this section we introduce some basic concepts on differentiation in functional spaces; a more detailed treatment, with applications to optimization of functionals, is given in ch. 7 of Luenberger (1969). Let $(\mathcal{S}_1, \|\cdot\|_1)$ and $(\mathcal{S}_2, \|\cdot\|_2)$ be normed linear spaces and $F : \mathcal{S}_1 \rightarrow \mathcal{S}_2$. F is said to be Fréchet differentiable at $a \in \mathcal{S}_1$ if there is a linear functional $\mathsf{D}F(a) : \mathcal{S}_1 \rightarrow \mathcal{S}_2$ such that $\|F(a+b) - F(a) - \mathsf{D}F(a)b\|_2 = o(\|b\|_1)$. When $\mathsf{D}F(a)$ is itself differentiable as a function of a in the norm $\|\mathsf{D}F(a)\| = \sup\{\|\mathsf{D}F(a)b\|_2 : \|b\|_1 \leq 1\}$, F is said to be twice Fréchet differentiable and the second differential is denoted by $\mathsf{D}^2F(a)$. These definitions will be applied to $\mathcal{S}_1 = \mathcal{M} \subseteq \mathcal{L}^\infty(T)$ equipped with the sup norm, $\|f\| = \sup_{t \in T} |f(t)|$, and $\mathcal{S}_2 = \mathbb{R}$ with the usual absolute value as norm. For Theorem 1 we also need to define the tensor product of functionals: for each $h \in \mathcal{S}_1$, $F_1 \otimes F_2 h$ is defined as the functional $(F_2 h)F_1$, and for a pair $(h_1, h_2) \in \mathcal{S}_1^2$, $F_1 \otimes F_2(h_1, h_2) = (F_2 h_1)(F_1 h_2)$.

Theorem 2 *Given $x \in \mathbb{R}^m$, let $\ell_x(\mu) = \log f(x;\mu)$. Then $\ell_x : \mathcal{L}^\infty(T) \rightarrow \mathbb{R}$ is twice Fréchet differentiable at every $\mu \in \mathcal{L}^\infty(T)$. The first differential is given by*

$$\mathsf{D}\ell_x(\mu)h = \int k_x(s;\mu)h(s)\,ds,$$

where

$$k_x(s;\mu) = \frac{1}{\sigma^2}\sum_{j=1}^{m}\{x_j - \mu(s)\}\,w_j(s|x;\mu)$$

and $w_j(s|x;\mu)$ *is the conditional density of* $g(t_j,\theta)$ *given* x*. The second differential is given by*

$$\begin{aligned}\mathsf{D}^2\ell_x(\mu)(h_1,h_2) &= \int\int \rho_x(s,t)h_1(s)h_2(t)\,ds\,dt + \\ &\quad \int \eta_x(s)h_1(s)h_2(s)\,ds - \mathsf{D}\ell_x(\mu)\otimes\mathsf{D}\ell_x(\mu)(h_1,h_2),\end{aligned}$$

where

$$\begin{aligned}\rho_x(s,t) &= \frac{1}{\sigma^4}\sum_{j=1}^m\sum_{\substack{k=1\\k\neq j}}^m\{x_j-\mu(s)\}\{x_k-\mu(t)\}v_{jk}(s,t|x;\mu),\\ \eta_x(s) &= \frac{1}{\sigma^4}\sum_{j=1}^m[\{x_j-\mu(s)\}^2-\sigma^2]w_j(s|x;\mu),\end{aligned}$$

and $v_{jk}(s,t|x;\mu)$ *is the joint conditional density of* $(g(t_j,\theta),g(t_k,\theta))$ *given* x*.*

Proof: We only need to show that $f(x;\mu)$ is twice differentiable as a function of μ for each x, and it will follow that $\mathsf{D}\ell_x(\mu)=\mathsf{D}f(x;\mu)/f(x;\mu)$ and $\mathsf{D}^2\ell_x(\mu)=\mathsf{D}^2f(x;\mu)/f(x;\mu)-\{\mathsf{D}f(x;\mu)\otimes\mathsf{D}f(x;\mu)\}/f^2(x;\mu)$. Given $x\in\mathbb{R}^m$, let $F_x(v)=(2\pi\sigma^2)^{-\frac{m}{2}}\exp\{-\|x-v\|^2/2\sigma^2\}$. This function is twice differentiable for every $v\in\mathbb{R}^m$, and the differentials are $\mathsf{D}F_x(v)=F_x(v)\sigma^{-2}(x-v)^\top$ and $\mathsf{D}^2F_x(v)=F_x(v)\sigma^{-4}(x-v)(x-v)^\top-F_x(v)\sigma^{-2}I$. Then the residuals $R_x^{(1)}(v,w)=F_x(v+w)-F_x(v)-\mathsf{D}F_x(v)w$ and $R_x^{(2)}(v,w)=\mathsf{D}F_x(v+w)-\mathsf{D}F_x(v)-\mathsf{D}^2F_x(v)w$ are $o(\|w\|)$ for each v.

Now, since $f(x;\mu)=\int F_x[\mu\{g(t^*,\theta)\}]f(\theta)\,d\theta$, it is not difficult to show that

$$\mathsf{D}f(x;\mu)h=\int\mathsf{D}F_x[\mu\{g(t^*,\theta)\}]h\{g(t^*,\theta)\}f(\theta)\,d\theta \tag{12}$$

and

$$\begin{aligned} & \mathsf{D}^2 f(x;\mu)(h_1,h_2) \\ = & \int h_2\{g(t^*,\theta)\}^\top \mathsf{D}^2 F_x[\mu\{g(t^*,\theta)\}] h_1\{g(t^*,\theta)\} f(\theta)\, d\theta. \end{aligned} \tag{13}$$

To prove this, note that

$$\begin{aligned} & f(x;\mu+h) - f(x;\mu) - \mathsf{D}f(x;\mu)h \\ = & \int R_x^{(1)}[\mu\{g(t^*,\theta)\}, h\{g(t^*,\theta)\}] f(\theta)\, d\theta \end{aligned}$$

and since $\|h\{g(t^*,\theta)\}\| \leq \sqrt{m}\|h\|$,

$$\begin{aligned} & \frac{|f(x;\mu+h) - f(x;\mu) - \mathsf{D}f(x;\mu)h|}{\|h\|} \\ \leq & \sqrt{m}\int \frac{|R_x^{(1)}[\mu\{g(t^*,\theta)\}, h\{g(t^*,\theta)\}]|}{\|h\{g(t^*,\theta)\}\|} f(\theta)\, d\theta. \end{aligned}$$

By dominated convergence, the right-hand side goes to zero as $\|h\|$ goes to zero and then (12) holds. For the second differential, we have that

$$\begin{aligned} & \{\mathsf{D}f(x;\mu+h_1) - \mathsf{D}f(x;\mu) - \mathsf{D}^2 f(x;\mu)h_1\}h_2 \\ = & \int h_2\{g(t^*,\theta)\}^\top R_x^{(2)}[\mu\{g(t^*,\theta)\}, h_1\{g(t^*,\theta)\}] f(\theta)\, d\theta \end{aligned}$$

and then

$$\begin{aligned} & \|\mathsf{D}f(x;\mu+h_1) - \mathsf{D}f(x;\mu) - \mathsf{D}^2 f(x;\mu)h_1\| \\ \leq & \int \|R_x^{(2)}[\mu\{g(t^*,\theta)\}, h_1\{g(t^*,\theta)\}]\| f(\theta)\, d\theta. \end{aligned}$$

Again, this implies that $\|\mathsf{D}f(x;\mu+h_1) - \mathsf{D}f(x;\mu) - \mathsf{D}^2 f(x;\mu)h_1\| = o(\|h_1\|)$ and then (13) holds. ■

The first and second differentials of $L_n(\mu)$ are $\Psi_n(\mu) := E_n \mathsf{D}\ell_x(\mu)$ and $\dot{\Psi}_n(\mu) := E_n \mathsf{D}^2\ell_x(\mu)$, respectively. The asymptotic versions of these functionals are obtained by substituting E_n with E_0, and will be respectively denoted by Ψ_0 and $\dot{\Psi}_0$. Being a maximum of L_n, $\hat{\mu}$ is zero of Ψ_n in the functional sense; that is, $\Psi_n(\hat{\mu})h = 0$ for all $h \in \mathcal{M}$. Similarly, μ_0 maximizes L_0 and then $\Psi_0(\mu_0)h = 0$ for all $h \in \mathcal{M}$ (which can be verified by direct calculation).

Proof of Theorem 1. (i) First, note that L_0 has a unique maximum at μ_0: since $\log x \leq 2(\sqrt{x} - 1)$ for all $x \geq 0$, we have

$$\begin{aligned} L_0(\mu) - L_0(\mu_0) &= \int \log \left\{ \frac{f(x;\mu)}{f(x;\mu_0)} \right\} f(x;\mu_0)\, dx \\ &\leq 2\Big\{ \int \sqrt{f(x;\mu)}\sqrt{f(x;\mu_0)}\, dx - 1 \Big\} \\ &= -\int \Big\{ \sqrt{f(x;\mu)} - \sqrt{f(x;\mu_0)} \Big\}^2 dx. \end{aligned}$$

Then $L_0(\mu) < L_0(\mu_0)$ whenever $\mu \neq \mu_0$, because the integral on the third line of the display is strictly negative for all $\mu \neq \mu_0$, by identifiability.

On the other hand, by Theorem 19.4 of van der Vaart (1998) we have $\sup_{\mu\in\mathcal{M}} |L_n(\mu) - L_0(\mu)| \to 0$ almost surely. This theorem applies because $|\ell_x(\mu_1) - \ell_x(\mu_2)| \leq \max_{\mu\in\mathcal{M}} \|\mathsf{D}\ell_x(\mu)\| \|\mu_1 - \mu_2\|$, then the finiteness of the bracketing numbers required by this theorem follows from the compactness of $\mathcal{M}$.

In a compact space, almost sure uniform convergence of L_n and uniqueness of the maximizer of L_0 imply strong consistency of $\hat{\mu}_n$. To see this, take a realization $\{\hat{\mu}_n^{(\omega)}\}$ such that $\|\hat{\mu}_n^{(\omega)} - \mu_0\| \nrightarrow 0$ (here ω denotes an element in the underlying probability space). By compactness of $\mathcal{M}$, there is a subsequence $\hat{\mu}_{n_k}^{(\omega)}$ that converges to certain $\mu^* \neq \mu_0$. For this subsequence we have $L_0(\hat{\mu}_{n_k}^{(\omega)}) \to L_0(\mu^*)$, and also $L_{n_k}^{(\omega)}(\mu_0) \leq L_{n_k}^{(\omega)}(\hat{\mu}_{n_k}^{(\omega)})$; if ω was such that $\|L_n^{(\omega)} - L_0\| \to 0$, this would imply that $L_0(\mu_0) \leq L_0(\mu^*)$, contradicting the uniqueness of μ_0 as maximizer of L_0. Therefore $\|\hat{\mu}_n^{(\omega)} - \mu_0\| \nrightarrow 0$ implies $\|L_n^{(\omega)} - L\| \nrightarrow 0$, hence $P(\|\hat{\mu}_n - \mu_0\| \nrightarrow 0) = 0$.

(ii) Since $\Psi_n(\hat{\mu}) = \Psi_0(\mu_0) = 0$ in the functional sense, we can write $-\sqrt{n}(\Psi_0(\hat{\mu})-\Psi_0(\mu_0)) = \sqrt{n}(\Psi_n-\Psi_0)(\mu_0)+r_n$ with $r_n = \sqrt{n}(\Psi_n-\Psi_0)(\hat{\mu}-\mu_0)$. The first step of this proof is to show that $\|r_n\| = o_P(1)$. Let $\mathbb{G}_n = \sqrt{n}(E_n - E_0)$ denote the empirical process. Then $\sqrt{n}(\Psi_n - \Psi_0)(\hat{\mu} - \mu_0)h = \mathbb{G}_n\psi_{\hat{\mu},h}$, where $\psi_{\mu,h}(x) = (\mathsf{D}\ell_x(\mu) - \mathsf{D}\ell_x(\mu_0))h$. The family $\{\psi_{\mu,h} : (\mu, h) \in \mathcal{M} \times \mathcal{M}\}$ is a Donsker class because it is Lipschitz in (μ, h), with a square-integrable Lipschitz factor, and the parametric space $\mathcal{M} \times \mathcal{M}$ is compact (van der Vaart 1998, Theorem 19.5). Then, since $\hat{\mu} \xrightarrow{P} \mu_0$, we have that $\sup_{h\in\mathcal{M}} |\mathbb{G}_n\psi_{\hat{\mu},h}| \xrightarrow{D} \sup_{h\in\mathcal{M}} |\mathbb{G}\psi_{\mu_0,h}|$, where $\mathbb{G}$ is a Gaussian element with zero mean and covariance $E\{\mathbb{G}\psi_{\mu_1,h_1}\mathbb{G}\psi_{\mu_2,h_2}\} = E_0\{\psi_{\mu_1,h_1}(x)\psi_{\mu_2,h_2}(x)\} - E_0\{\psi_{\mu_1,h_1}(x)\}E\{\psi_{\mu_2,h_2}(x)\}$. Since $\psi_{\mu_0,h} \equiv 0$ for all h, it follows that $\sup_{h\in\mathcal{M}} |\mathbb{G}_n\psi_{\hat{\mu},h}| \xrightarrow{D} 0$, which is just another way of writing $\|r_n\| = o_P(1)$.

Let us find now the limit distribution of $\sqrt{n}(\Psi_n - \Psi_0)(\mu_0)$. Again, we can write $\sqrt{n}(\Psi_n - \Psi_0)(\mu_0)h = \mathbb{G}_n\xi_h$ where $\xi_h(x) = \mathsf{D}\ell_x(\mu_0)h$. As before, $\{\xi_h : h \in \mathcal{M}\}$ is a Donsker family, so that $\mathbb{G}_n\xi_h \xrightarrow{D} \mathbb{G}\xi_h$ uniformly in h, where $\mathbb{G}$ is a zero-mean Gaussian element with covariances given by $E\{\mathbb{G}\xi_{h_1}\mathbb{G}\xi_{h_2}\} = E_0\{\xi_{h_1}(x)\xi_{h_2}(x)\} - E_0\{\xi_{h_1}(x)\}E_0\{\xi_{h_2}(x)\} = \mathcal{I}h_1h_2$. This together with $\|r_n\| = o_P(1)$ imply that $\sqrt{n}(\Psi_0(\hat{\mu}) - \Psi_0(\mu_0))$ converges to a Gaussian random element with mean zero and covariance operator $\mathcal{I}$. Since $\dot{\Psi}_0(\mu_0) = E_0\{\mathsf{D}^2\ell_x(\mu_0)\} = -\mathcal{I} \neq 0$, the functional delta method (van der Vaart 1998, Theorem 20.8) applied to Ψ_0^{-1} implies that $\sqrt{n}(\hat{\mu} - \mu_0)$ converges in distribution to a Gaussian random element with mean zero and covariance operator $\mathcal{I}^{-1}$. ■

REFERENCES

FRITSCH, F. N. & CARLSON, R. E. (1980). Monotone piecewise cubic interpolation. *SIAM J. Numer. Anal.* **17**, 238–46.

GASSER, T., KNEIP, A., BINDING, A., PRADER, A. & MOLINARI, L. (1991). The dynamics of linear growth in distance, velocity and acceleration. *Annals of Human Biology* **18**, 187–205.

GASSER, T. & KNEIP, A. (1995). Searching for structure in curve samples. *J. Amer. Statist. Assoc.* **90**, 1179–88.

GERVINI, D. & GASSER, T. (2004). Self-modeling warping functions. *J. R. Statist. Soc.* B **66**, 959–71.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* New York: Springer–Verlag.

KNEIP, A. & GASSER, T. (1992). Statistical tools to analyze data representing a sample of curves. *Ann. Statist.* **20**, 1266–305.

KNEIP, A. & ENGEL, J. (1995). Model estimation in nonlinear regression under shape invariance. *Ann. Statist.* **23**, 551–70.

KNEIP, A., LI, X., MACGIBBON, K. B. & RAMSAY, J. O. (2000). Curve registration by local regression. *Canad. J. Statist.* **28**, 19-29.

LUENBERGER, D. (1969). *Optimization by Vector Space Methods.* New York: John Wiley.

RAMSAY, J. O. (1988). Monotone regression splines in action. *Statist. Sci.* **3**, 425–41.

RAMSAY, J. O. (1998). Estimating smooth monotone functions. *J. R. Statist. Soc.* B **60**, 365–75.

RAMSAY, J. O. & SILVERMAN, B. W. (1997). *Functional Data Analysis.* New York: Springer–Verlag.

RAMSAY, J. O. & LI, X. C. (1998). Curve registration. *J. R. Statist. Soc.* B **60**, 351–63.

RAMSAY, J. O. & SILVERMAN, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies.* New York: Springer–Verlag.

RØNN, B. B. (2001). Nonparametric maximum likelihood estimation for shifted curves. *J. R. Statist. Soc.* B **63**, 243–59.

SILVERMAN, B. W. (1995). Incorporating parametric effects into functional principal components analysis. *J. R. Statist. Soc.* B **57**, 673–89.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics.* United Kingdom: Cambridge University Press.

WAND, M. P. & JONES, M. C. (1995). *Kernel Smoothing.* Boca Raton: Chapman and Hall/CRC Press.

WANG, K. & GASSER, T. (1999). Synchronizing sample curves nonparametrically. *Ann. Statist.* **27**, 439–60.