Warped Functional Analysis of Variance

Daniel Gervini

Department of Mathematical Sciences

University of Wisconsin–Milwaukee

PO Box 413, Milwaukee, WI 53201

and Patrick A. Carter

School of Biological Sciences

Washington State University

PO Box 644236, Pullman, WA 99164

February 28, 2014

Abstract

This article presents an Analysis of Variance model for functional data that explicitly incorporates phase variability through a time-warping component, allowing for a unified approach to estimation and inference in presence of amplitude and time variability. The focus is on single-random-factor models but the approach can be easily generalized to more complex ANOVA models. The behavior of the estimators is studied by simulation, and an application to the analysis of growth curves of flour beetles is presented. Although the model assumes a smooth latent process behind the observed trajectories, smoothness of the observed data is not required; the method can be applied to irregular time grids, which are common in longitudinal studies.

Key words: Karhunen–Loève decomposition; longitudinal data; phase variability; quantitative genetics; random-effect models.

1 Introduction

The main motivation for the present paper is the study of functional traits in evolutionary biology and quantitative genetics. Evolutionary biology investigates the change of physical traits (phenotypes) across generations. Some traits are univariate or multivariate, but others are functional, like growth curves or thermal performance curves (Kirkpatrick and Heckman, 1989; Heckman, 2003; Kingsolver et al., 2002; Meyer and Kirkpatrick, 2005; Ragland and Carter, 2004). Understanding the modes of variability of these curves is important in order to understand the biological processes behind the trait, and in particular the genetic aspects of it.

Consider for example the flour-beetle growth curves shown in Figure 1(a) (see Irwin and Carter, 2013, for details about these data). They are mass measurements of larvae from hatching to pupation. The dataset consists of 122 half-siblings sired by 29 fathers and different mothers. A distinct characteristic of these curves is an inflection point around day 15; this is the time when larvae stop eating and begin searching for a place to pupate. This process is triggered by hormonal mechanisms whose timing varies from individual to individual; determining what proportion of the time variability can be attributed to genetic factors and what proportion can be attributed to environmental factors is important for understanding the evolution of development and growth. Similarly, in the study of thermal performance curves (which are functions of temperature, not time), the optimal temperature varies from individual to individual and characterizing the sources of this variability is important for understanding thermal adaptations (Huey and Kingsolver, 1989; Izem and Kingsolver, 2005).

We can see, then, that functional samples usually present two types of variability: what we can denominate “horizontal” or “phase” variability (e.g. variability in the location of the mass peaks in Figure 1(a)) and “vertical” or “amplitude” variability (e.g. variability in

the mass magnitude at the peak in Figure 1(a)). It is important to point out that for a given data set there is often some ambiguity about what constitutes amplitude variability and what constitutes phase variability (this will be discussed in more depth in Section 2). The problem of decomposing functional variability into amplitude and phase variability has been addressed by many authors (Kneip and Engel, 1995; Ramsay and Li, 1998; Wang and Gasser, 1999; Kneip et al., 2000; Gervini and Gasser, 2004, 2005; Kneip and Ramsay, 2008; Tang and Müller, 2008; Telesca and Inoue, 2008; Bigot and Gadat, 2010; Claeskens et al., 2010). All of these papers, however, have focused on independent and identically distributed samples of curves, but for the type of applications we have in mind the curves are not independent. For example, the growth curves in Figure 1(a) are correlated for individuals with the same father. This type of design is common in evolutionary biology and quantitative genetics for the following reason. The variability observed in physical traits has two sources: genetic and environmental. Because environmental factors generally are not passed from one generation to the next, the evolution of phenotypes is driven largely by genetic variability (but see Skinner et al., 2010 and Manikkam et al., 2012 for a discussion of epigenetic effects). Examining samples of genetically related individuals, like siblings or half-siblings, makes the genetic and environmental sources of variability mathematically identifiable and therefore estimable, allowing biologists to predict the evolution of traits in response to selection (Gomulkiewicz and Beder, 1996; Kingsolver at al., 2002).

Therefore, it is important to possess statistical tools for the study of amplitude and phase variability of non-independent functional data. Some existing functional-data methods handle non-independent or non-identically distributed curves, such as mixed-effects ANOVA models (Guo, 2002; Morris and Carroll, 2006; Di et al., 2009; Chen and Wang, 2011), but they do not specifically address phase variability. To date, the problem of amplitude/phase variability of functional traits has been addressed mostly in an ad-hoc way, by first aligning the curves with respect to some trait, and then studying amplitude vari-

ability of the aligned curves. (This process of aligning curves is variously known as “time warping” or “curve registration” in the Functional Data literature.) But evolutionary biologists frequently must make decisions about how to align or register individual curves from a population of individuals. For example, when studying growth curves in a population of animals that undergo metamorphosis from one life history state to another (usually from a non-reproductive larval form to a reproductive adult form), it is not necessarily clear how to align the individual curves. The default choice for most biologists is to align the curves at the date of birth or hatching, but an equally valid choice might be the date of metamorphosis or the peak body mass prior to metamorphosis. For example, Ragland and Carter (2004) chose to align the growth curves of larval salamanders by date of metamorphosis and then reset the growth period to a fractional scale. Although this approach was effective, it was unsophisticated and ad hoc; more rigorous methods would be beneficial.

In this paper we propose a functional ANOVA approach that explicitly models time variability. For simplicity, we consider only the one-way random factor model, but the ideas can be easily extended to more complex ANOVA models. We follow a likelihood-based approach that uses the raw data directly, without pre-smoothing. Therefore the method can be applied to irregularly sampled trajectories, with possibly different starting points and endpoints. The fact that pre-smoothing is not necessary makes the method applicable to longitudinal data, where a smooth latent process is assumed but the observed data themselves are not smooth (Rice, 2004; Müller, 2008). The paper is organized as follows: a brief background on random processes is given in Section 2; the warped ANOVA model is presented in Section 3; the asymptotic distribution of the main parameter estimators is derived in Section 4; the small sample behavior of the estimators is studied by simulation in Section 5; finally, the beetle growth data is analyzed in detail in Section 6.

2 Brief background on random processes

Before we present the warped ANOVA model, it is useful to review some basic properties of stochastic processes. Let $x : I \to \mathbb{R}$ be a random function defined on a finite interval $I \subset \mathbb{R}$. Suppose $x(t)$ is square-integrable with probability one, and has finite variance. Let $\mu(t) = \mathrm{E}\{x(t)\}$ and $\rho(s,t) = \mathrm{cov}\{x(s), x(t)\}$. Then $x(t)$ admits the decomposition

$$x(t) = \mu(t) + \sum_{k=1}^{\infty} Z_k \phi_k(t), \tag{1}$$

which is known as Karhunen–Loève decomposition (Ash and Gardner, 1975), where the Z_ks are uncorrelated random variables with $\mathrm{E}(Z_k) = 0$ and $\mathrm{var}(Z_k) = \lambda_k$ (without loss of generality we can assume $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$). The ϕ_ks form an orthonormal system in $\mathscr{L}^2(I)$ and are eigenfunctions of the covariance function ρ with eigenvalues λ_k; that is, $\int \rho(s,t)\phi_k(s)ds = \lambda_k \phi_k(t)$, which implies

$$\rho(s,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s)\phi_k(t). \tag{2}$$

If the covariance function ρ is continuous then (1) and (2) converge pointwise; otherwise the convergence is only in the sense of the $\mathscr{L}^2(I)$ norm (Gohberg *et al.*, 2003). In either case, $\sum_{k=1}^{\infty} \lambda_k < \infty$, so the sequence of eigenvalues converges to zero. The Karhunen–Loève decomposition is the functional equivalent of the multivariate principal-component decomposition.

Although from a mathematical point of view decomposition (1) always holds, from a statistical point of view it is not always the most parsimonious model. It is often the case that the sample curves present a few distinct peaks and valleys that systematically repeat themselves across curves, albeit at somewhat different locations. It may take a lot of terms in (1) to explain this kind of variability, but Kneip and Ramsay (2008, Proposition 1) show

that if the process $x(t)$ has at most K peaks and valleys and its derivative $x'(t)$ has at most K zeros, then $x(t)$ admits the decomposition

$$x(t) = \sum_{j=1}^{p} C_j \xi_j\{v(t)\} \tag{3}$$

for some $p \leq K + 2$, where the ξ_js are non-random basis functions, the C_js are random coefficients, and $v : I \to I$ is a monotone increasing stochastic process such that $\mathrm{E}\{v(t)\} = t$ (or alternatively $\mathrm{E}\{w(s)\} = s$, where $w(t)$ is the inverse function of $v(t)$.) We can re-express (3) as

$$x\{w(s)\} = \mu^*(s) + \sum_{k=1}^{p} Z_k^* \phi_k^*(s), \tag{4}$$

which is just the Karhunen–Loève decomposition of the warped process $z = x \circ w$. The process w is called the warping process, and it explains the "horizontal" variability in the location of the peaks and valleys of x.

It is important to point out that the Karhunen–Loève decomposition (1) is essentially unique (up to the usual indeterminacy of eigenfunctions for multiple eigenvalues), and so is (4) for a given warping process w; but the warping process w itself is not unique. For a given $x(t)$, different warping processes $w(t)$ can be chosen that will give rise to different decompositions (4). In general, it is not possible to uniquely define what constitutes amplitude variability and what constitutes phase variability for a given process $x(t)$. The approach usually followed in the literature is to specify a warping family $\mathscr{W}$ where $w(t)$ is constrained to live, and then define as phase variability whatever is accounted for by the family $\mathscr{W}$ and as amplitude variability whatever is accounted for by the residual decomposition (4). This may sound vague, but in fact it is possible to give simple conditions for model (4) to be identifiable given a family $\mathscr{W}$; see the discussion in Web Appendix E. Some authors choose very rigid warping families $\mathscr{W}$, like linear warping functions (San-

galli *et al.*, 2010), while others use extremely flexible nonparametric families (Telesca and Inoue, 2008; Ramsay and Li, 1998). We will follow an intermediate approach, using the semiparametric family of interpolating monotone Hermite splines (Fritsch and Carlson, 1980), although the proposed method can be implemented with any other warping family.

Monotone interpolating Hermite splines are defined as follows (more details are given in Web Appendix C). For a subject i, let $\boldsymbol{\tau}_i \in \mathbb{R}^r$ be a vector of "landmarks" in $I = [a, b]$, with $a < \tau_{i1} < \cdots < \tau_{ir} < b$; the τs can be, for example, the locations of the peaks and valleys of the observed curve. Let $\boldsymbol{\tau}_0 \in \mathbb{R}^r$ be a knot vector, usually taken as the mean of the $\boldsymbol{\tau}_i$s. For given values $s_{i0}, \ldots, s_{i,r+1}$, there exists a unique piecewise cubic function $w_i(t)$ such that $w_i(a) = a$, $w_i(b) = b$, $w_i(\tau_{0j}) = \tau_{ij}$ for all j, $w_i'(a) = s_{i0}$, $w_i'(b) = s_{i,r+1}$, and $w_i'(\tau_{0j}) = s_{ij}$ for all j. This function $w_i(t)$ aligns the individual features $\boldsymbol{\tau}_i$ with the average features $\boldsymbol{\tau}_0$ in a smooth way, so it is useful for "landmark registration" (Bookstein, 1997). For $w_i(t)$ to be strictly monotone increasing the derivatives s_{ij}s must satisfy certain conditions, given in Fritsch and Carlson (1980). But for curve-alignment purposes only the $\boldsymbol{\tau}_i$s are specified; in that case Fritsch and Carlson (1980) provide an algorithm that produces a vector of derivatives $\mathbf{s}_i$ that satisfy the sufficient conditions for $w_i(t)$ to be monotone increasing. Since the algorithm is deterministic, $\mathbf{s}_i$ is a function of $\boldsymbol{\tau}_i$ and $\boldsymbol{\tau}_0$, therefore $w_i(t)$ is entirely parameterized by $\boldsymbol{\tau}_i$ and $\boldsymbol{\tau}_0$. In this paper, instead of specifying $\boldsymbol{\tau}_i$ for each curve and taking $\boldsymbol{\tau}_0 = \bar{\boldsymbol{\tau}}$, we will specify $\boldsymbol{\tau}_0$ and treat the $\boldsymbol{\tau}_i$s as unobserved random effects. Our family of warping functions $\mathscr{W}_{\boldsymbol{\tau}_0}$, then, is an r-dimensional space (r will usually be small). In general, it is not problematic to specify a reasonable $\boldsymbol{\tau}_0$ for a given data set; for example, for the curves in Figure 1(a) a single knot at $\tau_0 = 15$ will provide reasonable warping flexibility, and the rest of the variation will be considered amplitude variability. For other warping families, such as monotone B-splines (Telesca and Inoue, 2008) or smooth monotone transformations (Ramsay and Li, 1998), the number and placement of the knots are harder to specify because they are not directly associated

with curve features.

3 The warped ANOVA model

Let us go back now to the original problem of a one-factor design, where the sample of n individuals can be separated into I groups, with group i containing J_i individuals. For subject j in group i we observe certain variable (e.g. mass) at time points $t_{ij1}, \ldots, t_{ij\nu_{ij}}$, obtaining observations $y_{ij1}, \ldots, y_{ij\nu_{ij}}$. The number of observations ν_{ij} as well as the time points may change from individual to individual. We assume

$$y_{ijk} = x_{ij}(t_{ijk}) + \varepsilon_{ijk}, \tag{5}$$

where $\{x_{ij}(t)\}$ are underlying smooth curves, not directly observable, and $\{\varepsilon_{ijk}\}$ are i.i.d. $N(0, \sigma^2)$ random errors independent of the underlying $x_{ij}(t)$s. Observational model (5), which treats the smooth curves $\{x_{ij}(t)\}$ as latent variables, is the usual way to bridge functional data analysis and longitudinal data analysis (Müller, 2008). As discussed in Section 2, we can write $x_{ij}(t) = z_{ij}\{w_{ij}^{-1}(t)\}$ for a warped process $z_{ij}(t)$ and a warping function $w_{ij}(t)$. These will inherit the dependence structure of the x_{ij}s, so we can assume

$$z_{ij}(t) = \mu(t) + \alpha_i(t) + \beta_{ij}(t), \;\; j = 1, \ldots, J_i, \;\; i = 1, \ldots, I, \tag{6}$$

with $\{\alpha_i(t)\}$ and $\{\beta_{ij}(t)\}$ zero-mean random factors independent of each other and among themselves. For the main factor $\alpha(t)$ and the residual term $\beta(t)$ we assume expansions analogous to (4):

$$\alpha(t) = \sum_{k=1}^{p} U_k \phi_k(t), \tag{7}$$

$$\beta(t) = \sum_{k=1}^{q} V_k \psi_k(t), \tag{8}$$

where $\{\phi_k(t)\}$ and $\{\psi_k(t)\}$ are orthonormal functions in $\mathscr{L}^2(I)$, the U_ks are uncorrelated with $\mathrm{E}(U_k) = 0$ and $\mathrm{var}(U_k) = \gamma_k$, and the V_ks are uncorrelated with $\mathrm{E}(V_k) = 0$ and $\mathrm{var}(V_k) = \lambda_k$. Without loss of generality we assume $\gamma_1 \geq \cdots \geq \gamma_p > 0$ and $\lambda_1 \geq \cdots \geq \lambda_q > 0$.

From (6), (7) and (8) it follows that the total variance of $z_{ij}(t)$, defined as $\mathrm{E}(\|z_{ij} - \mu\|^2)$ with $\|\cdot\|$ the usual $\mathscr{L}^2$-norm, can be decomposed as $\mathrm{E}(\|\alpha\|^2) + \mathrm{E}(\|\beta\|^2)$, where $\mathrm{E}(\|\alpha\|^2) = \sum_{k=1}^{p} \gamma_k$ is the main-factor variance and $\mathrm{E}(\|\beta\|^2) = \sum_{k=1}^{q} \lambda_k$ is the residual-factor variance. The ratio

$$h_z = \frac{\sum_{k=1}^{p} \gamma_k}{\sum_{k=1}^{p} \gamma_k + \sum_{k=1}^{q} \lambda_k} \tag{9}$$

is then the proportion of amplitude variability explained by the main factor. In Section 4 we will derive asymptotic confidence intervals for h_z.

The mean function $\mu(t)$ and the Karhunen–Loève components $\{\phi_k(t)\}$ and $\{\psi_k(t)\}$ are functional parameters that must be estimated from the data, using for instance semiparametric spline models. Let $\mathbf{b}(t) = (b_1(t), \ldots, b_s(t))^T$ be a spline basis in $\mathscr{L}^2(I)$ (for simplicity we will use the same spline basis for all functional parameters, but this is not strictly necessary); then we assume $\mu(t) = \mathbf{b}(t)^T\mathbf{m}$, $\phi_k(t) = \mathbf{b}(t)^T\mathbf{c}_k$, and $\psi_k(t) = \mathbf{b}(t)^T\mathbf{d}_k$, for parameters $\mathbf{m}$, $\mathbf{c}_k$ and $\mathbf{d}_k$ in $\mathbb{R}^s$. Let $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_p] \in \mathbb{R}^{s\times p}$, $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_q] \in \mathbb{R}^{s\times q}$ and $\mathbf{J} = \int_a^b \mathbf{b}(t)\mathbf{b}(t)^T dt \in \mathbb{R}^{s\times s}$. The orthogonality conditions on the ϕ_ks and the ψ_ks translate into the conditions $\mathbf{C}^T\mathbf{J}\mathbf{C} = \mathbf{I}_p$ and $\mathbf{D}^T\mathbf{J}\mathbf{D} = \mathbf{I}_q$ for $\mathbf{C}$ and $\mathbf{D}$. Regarding the U_ks and V_ks in (7) and (8), we assume that $\mathbf{U} = (U_1, \ldots, U_p)^T$ follows a multivariate $N(\mathbf{0}, \mathbf{\Gamma})$ distribution with $\mathbf{\Gamma} = \mathrm{diag}(\gamma_1, \ldots, \gamma_p)$ and that $\mathbf{V} = (V_1, \ldots, V_q)^T$ follows a multivariate $N(\mathbf{0}, \mathbf{\Lambda})$ distribution with $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$. To summarize, the warped process (6) is parameterized by $\mathbf{m}$, $\mathbf{C}$, $\mathbf{D}$, $\mathbf{\Gamma}$ and $\mathbf{\Lambda}$.

For the warping functions $w_{ij}(t)$ we cannot simply assume an additive model like (6)

with Gaussian factors, because there are no monotone Gaussian processes. Therefore, a more indirect approach is needed. We will assume the w_{ij}s belong to the family of interpolating Hermite cubic splines $\mathscr{W}_{\boldsymbol{\tau}_0}$ for a specified knot vector $\boldsymbol{\tau}_0$. We have seen in Section 2 that a $w_{ij} \in \mathscr{W}_{\boldsymbol{\tau}_0}$ is parameterized by a vector $\boldsymbol{\tau}_{ij}$ that can be treated as a random effect. However, due to the restriction $a < \tau_{ij1} < \cdots < \tau_{ijr} < b$ we cannot assume $\boldsymbol{\tau}_{ij}$ is Normal. So we follow the approach of Brumback and Lindstrom (2004) and use the Jupp (1978) transform $\boldsymbol{\theta}_{ij} = \mathscr{J}(\boldsymbol{\tau}_{ij})$, defined as $\theta_{ijk} = \log\{(\tau_{ij,k+1} - \tau_{ijk})/(\tau_{ijk} - \tau_{ij,k-1})\}$ for $k = 1, \ldots, r$, which is an invertible transformation that maps vectors $\boldsymbol{\tau}_{ij}$ with increasing coordinates into unconstrained vectors $\boldsymbol{\theta}_{ij}$. For the unconstrained vector $\boldsymbol{\theta}_{ij}$ we can assume a multivariate Normal distribution and an additive ANOVA model:

$$\boldsymbol{\theta}_{ij} = \boldsymbol{\theta}_0 + \boldsymbol{\eta}_i + \boldsymbol{\xi}_{ij}, \tag{10}$$

with $\boldsymbol{\eta}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\xi}_{ij} \sim N(\mathbf{0}, \boldsymbol{\Omega})$ independent of each other and among themselves. We will also assume the $\boldsymbol{\theta}_{ij}$s are independent of the amplitude factors $\alpha_i(t)$ and $\beta_{ij}(t)$, although a model with correlations between amplitude and warping factors can be set up (see below). We take $\boldsymbol{\theta}_0 = \mathscr{J}(\boldsymbol{\tau}_0)$; the covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ will be estimated from the data. In analogy with (9) we define

$$h_w = \frac{\mathrm{tr}(\boldsymbol{\Sigma})}{\mathrm{tr}(\boldsymbol{\Sigma} + \boldsymbol{\Omega})}, \tag{11}$$

which is the proportion of the warping variability explained by the main factor.

Putting together the models for $z_{ij}(t)$, $w_{ij}(t)$ and the observational model (5), we can derive the likelihood function for the observed data vectors $\mathbf{y}_{ij} = (y_{ij1}, \ldots, y_{ij\nu_{ij}})$. Given a realization of the random effect $\boldsymbol{\theta}_{ij}$, which is determined by realizations of $\boldsymbol{\eta}_i$ and $\boldsymbol{\xi}_{ij}$, the corresponding warped time grids are $t^*_{ijk}(\boldsymbol{\theta}_{ij}) = w_{ij}^{-1}(t_{ijk})$, $k = 1, \ldots, \nu_{ij}$, and the corresponding warped B-spline matrices $\mathbf{B}^*_{ij}(\boldsymbol{\theta}_{ij}) \in \mathbb{R}^{\nu_{ij} \times s}$ are given by $[\mathbf{B}^*_{ij}(\boldsymbol{\theta}_{ij})]_{kl} =$

$b_l\{t^*_{ijk}(\boldsymbol{\theta}_{ij})\}$. Then

$$\mathbf{y}_{ij}|(\mathbf{u}_i, \mathbf{v}_{ij}, \boldsymbol{\eta}_i, \boldsymbol{\xi}_{ij}) \sim N\left\{\mathbf{B}^*_{ij}(\boldsymbol{\theta}_{ij})\mathbf{m} + \mathbf{B}^*_{ij}(\boldsymbol{\theta}_{ij})\mathbf{C}\mathbf{u}_i + \mathbf{B}^*_{ij}(\boldsymbol{\theta}_{ij})\mathbf{D}\mathbf{v}_{ij}, \sigma^2\mathbf{I}_{\nu_{ij}}\right\},$$

and the $\mathbf{y}_{ij}$s are conditionally independent given $(\mathbf{u}_i, \mathbf{v}_{ij}, \boldsymbol{\eta}_i, \boldsymbol{\xi}_{ij})$. If $\mathbf{y}_{i\cdot} = (\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iJ_i})$, we have

$$f(\mathbf{y}_{i\cdot}) = \iint g(\mathbf{u}_i, \boldsymbol{\eta}_i) f(\mathbf{u}_i) f(\boldsymbol{\eta}_i) \mathrm{d}\mathbf{u}_i \mathrm{d}\boldsymbol{\eta}_i \tag{12}$$

with $g(\mathbf{u}_i, \boldsymbol{\eta}_i) = \prod_{j=1}^{J_i} \iint f(\mathbf{y}_{ij}|\mathbf{u}_i, \mathbf{v}_{ij}, \boldsymbol{\eta}_i, \boldsymbol{\xi}_{ij}) f(\mathbf{v}_{ij}) f(\boldsymbol{\xi}_{ij}) \mathrm{d}\mathbf{v}_{ij} \mathrm{d}\boldsymbol{\xi}_{ij}$, and the log-likelihood function is $\ell = \sum_{i=1}^{I} \log f(\mathbf{y}_{i\cdot})$. The maximum likelihood estimators are $(\hat{\mathbf{m}}, \hat{\mathbf{C}}, \hat{\mathbf{D}}, \hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\Gamma}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\Omega}}, \hat{\sigma}^2) = \arg\max \ell$. We compute them via the EM algorithm. The implementation of the EM algorithm presents certain complications arising from the orthogonality restrictions on $\mathbf{C}$ and $\mathbf{D}$, which are discussed in detail in Web Appendix A.

In the rest of the paper we will use the estimators as presented above, but to conclude the section we discuss a few possible generalizations. First, it is possible to use other families of warping functions, such as B-splines with monotone increasing coefficients (Brumback and Lindstrom, 2004; Telesca and Inoue, 2008) or smooth monotone transformations (Ramsay and Li, 1998). The problem is that the spline coefficients for these families cannot be directly related to features of the sample curves in the way interpolating Hermite spline coefficients can; therefore, one may have to use a relatively large number of knots placed at somewhat arbitrary locations (equally spaced, for example). This may result in a warping family that is too flexible and lead to overwarping (i.e., produce warping functions with flat parts that are close to singular). To prevent this, the warping variability must somehow be penalized. This can be done by adding a penalty term to the log-likelihood function and minimize $\ell_\lambda = \sum_{i=1}^{I} \log f(\mathbf{y}_{i\cdot}) - \lambda \mathrm{tr}(\boldsymbol{\Sigma} + \boldsymbol{\Omega})$, where $\lambda \geq 0$ is a penalty parameter chosen by the user.

Second, it is possible to incorporate correlations between the warping process $w_{ij}(t)$

and the amplitude process $z_{ij}(t)$. This can be done by assuming that $(\mathbf{u}_i, \boldsymbol{\eta}_i)$ and $(\mathbf{v}_{ij}, \boldsymbol{\xi}_{ij})$ have joint Normal distributions, for instance. The only change in (12) would be that $f(\mathbf{u}_i)f(\boldsymbol{\eta}_i)$ is replaced by the joint density $f(\mathbf{u}_i, \boldsymbol{\eta}_i)$ and $f(\mathbf{v}_{ij})f(\boldsymbol{\xi}_{ij})$ by $f(\mathbf{v}_{ij}, \boldsymbol{\xi}_{ij})$. From the computational point of view this does not have a big impact, because the EM algorithm can be easily modified to accommodate this (a Matlab implementation is available as supplementary material). But from a statistical point of view the results may be harder to interpret, and the extra $rp + rq$ covariance parameters that need to be estimated may affect the precision of the rest of the estimators if the sample size is not very large.

Finally, we note that the assumption of normality of the random effects is mostly a working assumption to derive estimators. It is usually the case that properties like consistency and asymptotic normality of maximum likelihood estimators hold for broader families of distributions than the one they were derived for. But such a thorough asymptotic analysis is beyond the scope of this paper. Instead, in Section 5 we will study by simulation the robustness of the estimators to at least some mild departures from normality. On the other hand, if robustness to outliers is desired, this may be attained by substituting the Normal distributions by multivariate t distributions, as in Gervini (2009); the Normal EM algorithm is easy to adapt for multivariate t distributions.

4 Asymptotics and inference

It is usually of interest in applications to determine if the main-factor variance is significantly different from zero or not. To this end, we derive in this section the asymptotic distributions of the maximum likelihood estimators and the variance ratios (9) and (11), which can then be used to construct asymptotic confidence intervals and tests for h_z and h_w. For simplicity, we assume that *(i)* the true functional parameters $\mu(t)$, $\{\phi_k(t)\}$ and $\{\psi_k(t)\}$ belong to the spline space used for estimation, which is fixed, and *(ii)* the $\mathbf{y}_i$.s are

identically distributed, so $J_i = J$ for all i and the time grid $(t_1, \ldots, t_\nu)$ is the same for all individuals. The asymptotic distribution of the estimators will be derived for $I \to \infty$ and J fixed, or in practical terms, for "large I and small J"; this is the usual situation in random-effect one-way ANOVA models.

Under these conditions the standard maximum likelihood asymptotic theory applies: if $\boldsymbol{\omega} = (\gamma_1, \ldots, \gamma_p, \lambda_1, \ldots, \lambda_q)$, then $\sqrt{I}(\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}) \xrightarrow{D} N(\mathbf{0}, \mathbf{F}^{-1})$, where $\mathbf{F} = \mathrm{E}[\{\frac{\partial}{\partial \boldsymbol{\omega}} \log f(\mathbf{y}_{i\cdot})\} \{\frac{\partial}{\partial \boldsymbol{\omega}} \log f(\mathbf{y}_{i\cdot})\}^T]$ is the Fisher Information Matrix for the parameter $\boldsymbol{\omega}$. Straightforward differentiation of (12), which is carried out in detail in Web Appendix B, gives

$$\frac{\partial}{\partial \gamma_k} \log f(\mathbf{y}_{i\cdot}) = -\frac{1}{2\gamma_k} + \frac{\mathrm{E}(u_{ik}^2 | \mathbf{y}_{i\cdot})}{2\gamma_k^2}, \quad k = 1, \ldots, p,$$

and

$$\frac{\partial}{\partial \lambda_k} \log f(\mathbf{y}_{i\cdot}) = -\frac{J}{2\lambda_k} + \frac{1}{2\lambda_k^2} \sum_{j=1}^{J} \mathrm{E}(v_{ijk}^2 | \mathbf{y}_{i\cdot}), \quad k = 1, \ldots, q.$$

Let $\widehat{u_{ik}^2} = \mathrm{E}(u_{ik}^2 | \mathbf{y}_{i\cdot})$ and $\widehat{v_{ijk}^2} = \mathrm{E}(v_{ijk}^2 | \mathbf{y}_{i\cdot})$. Since $\mathrm{E}(\widehat{u_{ik}^2}) = \mathrm{E}(u_{ik}^2) = \gamma_k$ and $\mathrm{E}(\widehat{v_{ijk}^2}) = \mathrm{E}(v_{ijk}^2) = \lambda_k$, we obtain the following expressions:

$$F_{kl} = -\frac{1}{4\gamma_k \gamma_l} + \frac{\mathrm{E}(\widehat{u_{ik}^2} \widehat{u_{il}^2})}{4\gamma_k^2 \gamma_l^2}, \quad \text{for } k = 1, \ldots, p \text{ and } l = 1, \ldots, p,$$

$$F_{k,p+l} = -\frac{J}{4\gamma_k \lambda_l} + \frac{\mathrm{E}(\widehat{u_{ik}^2} \sum_{j=1}^{J} \widehat{v_{ijl}^2})}{4\gamma_k^2 \lambda_l^2}, \quad \text{for } k = 1, \ldots, p \text{ and } l = 1, \ldots, q,$$

and

$$F_{p+k,p+l} = -\frac{J^2}{4\lambda_k \lambda_l} + \frac{\mathrm{E}(\sum_{j=1}^{J} \widehat{v_{ijk}^2} \sum_{j=1}^{J} \widehat{v_{ijl}^2})}{4\lambda_k^2 \lambda_l^2}, \quad \text{for } k = 1, \ldots q \text{ and } l = 1, \ldots, q.$$

The estimator $\hat{\mathbf{F}}$ is obtained replacing expectations by averages over $i = 1, \ldots, I$.

The asymptotic distribution of (9) is derived via the Delta Method: since h_z is a differ-

entiable function of $\boldsymbol{\omega}$, $\sqrt{I}(\hat{h}_z - h_z) \xrightarrow{D} N\left(0, \{\partial h_z/\partial\boldsymbol{\omega}\}^T \mathbf{F}^{-1} \{\partial h_z/\partial\boldsymbol{\omega}\}\right)$ with

$$\frac{\partial h_z}{\partial \gamma_k} = \frac{\sum_{k=1}^q \lambda_k}{\left(\sum_{k=1}^p \gamma_k + \sum_{k=1}^q \lambda_k\right)^2}, \quad k = 1, \ldots, p,$$

and

$$\frac{\partial h_z}{\partial \lambda_k} = -\frac{\sum_{k=1}^p \gamma_k}{\left(\sum_{k=1}^p \gamma_k + \sum_{k=1}^q \lambda_k\right)^2}, \quad k = 1, \ldots, q.$$

The asymptotic variance of $\hat{h}_z$ is then given by

$$\begin{aligned}
\operatorname{avar}(\hat{h}_z) \;=\; & \frac{\left(\sum_{k=1}^p \lambda_k\right)^2}{\left(\sum_{k=1}^p \gamma_k + \sum_{k=1}^q \lambda_k\right)^4} \cdot \sum_{k=1}^p \sum_{l=1}^p \left(\mathbf{F}^{-1}\right)_{kl} \\
& - \frac{2\left(\sum_{k=1}^p \gamma_k\right)\left(\sum_{k=1}^q \lambda_k\right)}{\left(\sum_{k=1}^p \gamma_k + \sum_{k=1}^q \lambda_k\right)^4} \cdot \sum_{k=1}^p \sum_{l=1}^q \left(\mathbf{F}^{-1}\right)_{k,p+l} \\
& + \frac{\left(\sum_{k=1}^q \gamma_k\right)^2}{\left(\sum_{k=1}^p \gamma_k + \sum_{k=1}^q \lambda_k\right)^4} \cdot \sum_{k=1}^q \sum_{l=1}^q \left(\mathbf{F}^{-1}\right)_{p+k,p+l}.
\end{aligned}$$

The asymptotic distribution of (11) is derived in a similar way. If $\boldsymbol{\zeta} = (\operatorname{diag}(\boldsymbol{\Sigma}), \operatorname{diag}(\boldsymbol{\Omega}))$, then $\sqrt{I}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) \xrightarrow{D} N(\mathbf{0}, \mathbf{G}^{-1})$ with $\mathbf{G} = \mathrm{E}[\{\frac{\partial}{\partial\boldsymbol{\zeta}} \log f(\mathbf{y}_{i\cdot})\}\{\frac{\partial}{\partial\boldsymbol{\zeta}} \log f(\mathbf{y}_{i\cdot})\}^T]$, and $\sqrt{I}(\hat{h}_w - h_w) \xrightarrow{D} N(0, \operatorname{avar}(\hat{h}_w))$ with $\operatorname{avar}(\hat{h}_w) = (\partial h_w/\partial\boldsymbol{\zeta})^T \mathbf{G}^{-1} (\partial h_w/\partial\boldsymbol{\zeta})$, where

$$\frac{\partial h_w}{\partial \Sigma_{kk}} = \frac{\operatorname{tr}(\boldsymbol{\Omega})}{\{\operatorname{tr}(\boldsymbol{\Sigma}) + \operatorname{tr}(\boldsymbol{\Omega})\}^2}, \quad \text{for } k = 1, \ldots, r,$$

and

$$\frac{\partial h_w}{\partial \Omega_{kk}} = -\frac{\operatorname{tr}(\boldsymbol{\Sigma})}{\{\operatorname{tr}(\boldsymbol{\Sigma}) + \operatorname{tr}(\boldsymbol{\Omega})\}^2}, \quad \text{for } k = 1, \ldots, r.$$

Then

$$\begin{aligned}\mathrm{avar}(\hat{h}_w) &= \frac{\{\mathrm{tr}(\boldsymbol{\Omega})\}^2}{\{\mathrm{tr}(\boldsymbol{\Sigma})+\mathrm{tr}(\boldsymbol{\Omega})\}^4}\cdot\sum_{k=1}^{r}\sum_{l=1}^{r}\left(\mathbf{G}^{-1}\right)_{kl}\\ &\quad -\frac{2\mathrm{tr}(\boldsymbol{\Sigma})\mathrm{tr}(\boldsymbol{\Omega})}{\{\mathrm{tr}(\boldsymbol{\Sigma})+\mathrm{tr}(\boldsymbol{\Omega})\}^4}\cdot\sum_{k=1}^{r}\sum_{l=1}^{r}\left(\mathbf{G}^{-1}\right)_{k,r+l}\\ &\quad +\frac{\{\mathrm{tr}(\boldsymbol{\Sigma})\}^2}{\{\mathrm{tr}(\boldsymbol{\Sigma})+\mathrm{tr}(\boldsymbol{\Omega})\}^4}\cdot\sum_{k=1}^{r}\sum_{l=1}^{r}\left(\mathbf{G}^{-1}\right)_{r+k,r+l}.\end{aligned}$$

As shown in Web Appendix B, differentiation of (12) gives

$$\frac{\partial}{\partial\Sigma_{kk}}\log f(\mathbf{y}_{i\cdot})=-\frac{1}{2}\left(\boldsymbol{\Sigma}^{-1}\right)_{kk}+\frac{1}{2}\left(\boldsymbol{\Sigma}^{-1}\right)_{\cdot k}^{T}\mathrm{E}\left(\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T|\mathbf{y}_{i\cdot}\right)\left(\boldsymbol{\Sigma}^{-1}\right)_{\cdot k}$$

and

$$\frac{\partial}{\partial\Omega_{kk}}\log f(\mathbf{y}_{i\cdot})=-\frac{J}{2}\left(\boldsymbol{\Omega}^{-1}\right)_{kk}+\frac{1}{2}\left(\boldsymbol{\Omega}^{-1}\right)_{\cdot k}^{T}\sum_{j=1}^{J}\mathrm{E}\left(\boldsymbol{\xi}_{ij}\boldsymbol{\xi}_{ij}^T|\mathbf{y}_{i\cdot}\right)\left(\boldsymbol{\Omega}^{-1}\right)_{\cdot k},$$

where $(\boldsymbol{\Sigma}^{-1})_{\cdot k}$ and $(\boldsymbol{\Omega}^{-1})_{\cdot k}$ denote the kth columns of $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Omega}^{-1}$, respectively. Then, if we define $\widehat{\boldsymbol{\eta}_i^{\otimes 2}}=\mathrm{E}\left(\boldsymbol{\eta}_i\otimes\boldsymbol{\eta}_i|\mathbf{y}_{i\cdot}\right)$ and $\widehat{\boldsymbol{\xi}_{ij}^{\otimes 2}}=\mathrm{E}\left(\boldsymbol{\xi}_{ij}\otimes\boldsymbol{\xi}_{ij}|\mathbf{y}_{i\cdot}\right)$, after some algebra we obtain:

$$\begin{aligned}G_{kl} &= -\frac{1}{4}\left(\boldsymbol{\Sigma}^{-1}\right)_{kk}\left(\boldsymbol{\Sigma}^{-1}\right)_{ll}\\ &\quad +\frac{1}{4}\left\{\left(\boldsymbol{\Sigma}^{-1}\right)_{\cdot k}^{T}\otimes\left(\boldsymbol{\Sigma}^{-1}\right)_{\cdot k}^{T}\right\}\mathrm{E}\left(\widehat{\boldsymbol{\eta}_i^{\otimes 2}}\widehat{\boldsymbol{\eta}_i^{\otimes 2}}^T\right)\left\{\left(\boldsymbol{\Sigma}^{-1}\right)_{\cdot l}\otimes\left(\boldsymbol{\Sigma}^{-1}\right)_{\cdot l}\right\},\end{aligned}$$

$$\begin{aligned}G_{k,r+l} &= -\frac{J}{4}\left(\boldsymbol{\Sigma}^{-1}\right)_{kk}\left(\boldsymbol{\Omega}^{-1}\right)_{ll}\\ &\quad +\frac{1}{4}\left\{\left(\boldsymbol{\Sigma}^{-1}\right)_{\cdot k}^{T}\otimes\left(\boldsymbol{\Sigma}^{-1}\right)_{\cdot k}^{T}\right\}\mathrm{E}\left(\widehat{\boldsymbol{\eta}_i^{\otimes 2}}\sum_{j=1}^{J}\widehat{\boldsymbol{\xi}_{ij}^{\otimes 2}}^T\right)\left\{\left(\boldsymbol{\Omega}^{-1}\right)_{\cdot l}\otimes\left(\boldsymbol{\Omega}^{-1}\right)_{\cdot l}\right\},\end{aligned}$$

and

$$
\begin{aligned}
G_{r+k,r+l} &= -\frac{J^2}{4}\left(\boldsymbol{\Omega}^{-1}\right)_{kk}\left(\boldsymbol{\Omega}^{-1}\right)_{ll} \\
&\quad +\frac{1}{4}\left\{\left(\boldsymbol{\Omega}^{-1}\right)_{\cdot k}^{T}\otimes\left(\boldsymbol{\Omega}^{-1}\right)_{\cdot k}^{T}\right\}\mathrm{E}\Big(\sum_{j=1}^{J}\widehat{\boldsymbol{\xi}_{ij}^{\otimes 2}}\sum_{j=1}^{J}\widehat{\boldsymbol{\xi}_{ij}^{\otimes 2}}^{T}\Big)\left\{\left(\boldsymbol{\Omega}^{-1}\right)_{\cdot l}\otimes\left(\boldsymbol{\Omega}^{-1}\right)_{\cdot l}\right\},
\end{aligned}
$$

for $k = 1, \ldots, r$ and $l = 1, \ldots, r$. As before, $\hat{\mathbf{G}}$ is obtained replacing expectations by averages. Since the random-effect estimators $\widehat{u_{ik}^2}$, $\widehat{v_{ijk}^2}$, $\widehat{\boldsymbol{\eta}_i^{\otimes 2}}$ and $\widehat{\boldsymbol{\xi}_{ij}^{\otimes 2}}$ are by-products of the EM algorithm, no extra computational costs are incurred in computing $\hat{\mathbf{F}}$ and $\hat{\mathbf{G}}$.

Finally, we note that since $\hat{h}_z$ and $\hat{h}_w$ live in the interval $[0, 1]$, a transformation like $\arcsin\sqrt{h}$ usually provides a better Normal approximation when $\hat{h}_z$ or $\hat{h}_w$ are close to the boundaries. The asymptotic variance of $\arcsin\sqrt{\hat{h}}$ is given by $\mathrm{avar}(\hat{h})/\{4\hat{h}(1-\hat{h})\}$. The simplest procedure to derive a confidence interval for $\hat{h}$ in that case is to construct a standard confidence interval for $\arcsin\sqrt{\hat{h}}$ and then back-transform the endpoints.

5 Simulations

In this section we study the finite-sample behavior of the new estimators by simulation. The main goals are to determine if the new method *(i)* represents a substantial improvement over common functional ANOVA in presence of time variability, *(ii)* is at least comparable to the naive approach of pre-warping the data using an existing warping method, *(iii)* is robust to mild departures from the normality assumptions, and *(iv)* does not overfit, i.e. is not worse than common functional ANOVA in absence of time variability.

To this end we generated data from ten different models, all balanced, with $I = 10$ groups and $J = 5$ observations per group. The raw data (5) was sampled on an equally-spaced time grid of $\nu = 20$ points in $[0, 1]$, and the noise variance was $\sigma^2 = .1^2$ in all cases. The mean function was $\mu(t) = .6\varphi(t, .3, .1) + .4\varphi(t, .6, .1)$ in all cases, where $\varphi(t, a, b)$

denotes the $N(a, b^2)$ density function. The models considered were the following:

1. One-component models (7) and (8) with no warping and $\phi_1(t) = \psi_1(t) = \varphi(t, .3, .1)/1.68$. The variances were $\gamma_1 = .2^2$ and $\lambda_1 = .1^2$, so $h_z = .80$.

2. One-component models (7) and (8) with no warping but with different components for $\alpha(t)$ and $\beta(t)$: $\phi_1(t)$ as in Model 1, but $\psi_1(t) = \varphi(t, .6, .1)/1.68$. The variances γ_1 and λ_1 were as in Model 1.

3. Same $\alpha(t)$ and $\beta(t)$ as in Model 1, but with a Hermite-spline warping process $w(t)$ with knot $\tau_0 = .3$ and variances $\Sigma = .2^2$ and $\Omega = .1^2$, so $h_w = .80$.

4. Same $\alpha(t)$ and $\beta(t)$ as in Model 2, with warping $w(t)$ as in Model 3.

5. Same $\alpha(t)$ and $\beta(t)$ as in Model 1, but with a warping process $w(t)$ with knots $\boldsymbol{\tau}_0 = (.3, .6)$ and covariance matrices $\boldsymbol{\Sigma} = .2^2\mathbf{I}_2$ and $\boldsymbol{\Omega} = .1^2\mathbf{I}_2$, so $h_w = .80$ as before.

6. Same $\alpha(t)$ and $\beta(t)$ as in Model 2, with warping $w(t)$ as in Model 5.

7. Same as Model 4, but the random factors U and V in (7) and (8) have Student's t distributions with 4 degrees of freedom and scale parameters $\gamma_1^{1/2} = .2$ and $\lambda_1^{1/2} = .1$ (so the variance ratio is still $h_z = .80$).

8. Same as Model 4, but the random factors U and V in (7) and (8) have contaminated Normal distributions $(1-\varepsilon)N(0, \gamma_1)+\varepsilon N(0, k\gamma_1)$ and $(1-\varepsilon)N(0, \lambda_1)+\varepsilon N(0, k\lambda_1)$ respectively, with $\varepsilon = .10$ and $k = 5$ (the variance ratio is still $h_z = .80$).

9. Two-component models (7) and (8), with $\phi_1(t) = \psi_1(t) = \varphi(t, .3, .1)/1.68$, $\phi_2(t) = \psi_2(t) = (\varphi(t, .6, .1)/1.68 - .105\phi_1(t))/.99$ (so that each pc is associated with amplitude variation at each peak), variances $\gamma_1 = .2^2$, $\gamma_2 = .1^2$, $\lambda_1 = .1^2$, $\lambda_2 = .05^2$ (so $h_z = .80$ as in previous models), and a one-knot warping process as in Model 3.

10. Same $\alpha(t)$ and $\beta(t)$ as in Model 9, with two-knot warping as in Model 5.

For each sample we computed the common (un-warped) ANOVA estimator, the warped ANOVA estimator proposed in this paper, and a naive two-step warped ANOVA estimator. The latter is computed as follows: first the curves are aligned by least-squares registration (i.e. minimizing $\sum_{i=1}^{n} \|x_i \circ w_i - \mu\|^2$ over w_is in $\mathscr{W}_{\boldsymbol{\tau}_0}$) and then the common ANOVA estimators are computed on the warped data. We used cubic B-splines with 10 equispaced knots as basis functions for the functional parameters. As warping families we used interpolating Hermite splines with $\tau_0 = .3$ for models 1–4 and 7–9, and $\tau_0 = (.3, .6)$ for models 5–6 and 10. As error measures we used the bias, the standard deviation and the root mean squared error, defined as follows: if $f_0 \in \mathscr{L}^2(I)$ and $\hat{f}$ is the estimator, then $\mathrm{bias}(\hat{f}) = [\int \{\mathrm{E}\hat{f}(t) - f_0(t)\}^2 dt]^{1/2}$, $\mathrm{sd}(\hat{f}) = [\int \mathrm{E}\{\hat{f}(t) - \mathrm{E}\hat{f}(t)\}^2 dt]^{1/2}$ and $\mathrm{rmse}(\hat{f}) = \{\mathrm{bias}^2(\hat{f}) + \mathrm{sd}^2(\hat{f})\}^{1/2}$. Some care must be taken with the principal components, because their sign is undefined: to determine the "right" sign, we multiplied $\hat{\phi}_1$ and $\hat{\psi}_1$ by $\langle \hat{\phi}_1, \phi_1 \rangle$ and $\langle \hat{\psi}_1, \psi_1 \rangle$, respectively.

The estimation errors based on 200 Monte Carlo replications for each model are shown in Table 1. The effect of warping is more clearly seen in the estimators of the principal components ϕ and ψ. The common ANOVA estimators, as expected, have the largest biases; lacking a specific mechanism to handle time variability, common ANOVA estimators $\hat{\phi}$ and $\hat{\psi}$ attempt to fit amplitude and phase variability at the same time and get severely distorted compared to the true ϕ and ψ. The two warped estimators, on the other hand, can handle phase variability well. The maximum-likelihood estimator proposed in this paper always has smaller bias than the naive two-step approach; this is to be expected, since the warping step of the two-step estimator minimizes variation about the mean μ without taking into account amplitude variability or the dependence structure in the data, whereas the maximum likelihood estimator explicitly models ϕ and ψ. The down side of the new estimators is that, as always, the bias reduction provided by the more complex model is

accompanied by a higher variance. However, looking at the total root mean squared errors, we see that the new estimators outperform the naive two-step estimators in almost all cases. This is also true for the non-normal models 7 and 8, so the warped maximum likelihood estimators are robust to mild departures from normality.

6 Example: beetle growth data

In this section we study mass growth curves of flour beetles from birth to pupation, from Irwin and Carter (2013). A total of 122 insects are considered. This is a subset of a larger dataset that includes both siblings and half-siblings, but in order to apply the one-way ANOVA model, which assumes independence between groups, we consider only the half-siblings. (The full data set can be modeled as a nested two-way ANOVA, with the mother factor nested within the father factor.) The insects were sired by 29 different fathers, which will constitute the grouping variable. The number of insects per father varies between 2 and 5, with a median of 4.

Part of the raw data is shown in Figure 1(a); for better visualization we only plotted half of the sample curves. The mass measures were taken about every 3 days early in the growth curve, and up to once per day late in the growth curve. However, only 18 of the 122 larvae were measured for mass for the first time on the day they hatched; 76 were measured for mass for the first time on the second day, 22 on the third day, 5 on the fourth day, and one was not measured for mass until the seventh day. Therefore, the starting points of the curves are unequal. The endpoints are also irregular, because larvae reached pupation at different points between days 16 and 25. However, while the unequal starting points are due to missing data, the unequal endpoints are due to a well-defined biological landmark which is reached at different times. Therefore we rescaled the time grids so that all trajectories end at the median pupation day 19, but we did not align the starting points

at day 1. We also took logarithms to stabilize the error variance. The rescaled log-data is shown in Figure 1(b).

These curves have a noticeable inflection point around day 15. This is because in response to hormonal changes occurring prior to pupation, larvae stop eating and start wandering in search of a place to pupate, and so lose body mass. Therefore we fitted warped ANOVA models with a single warping knot at $\tau_0 = 15$. As spline basis we used cubic B-splines with 7 equispaced knots; this gives a total of 9 basis functions, providing enough flexibility without excessive irregularity. We considered several ANOVA models with equal number of components for the main factor and the residual term, ranging from 0 (mean-only model) to 3 components. The resulting parameter estimators were:

- For $p = q = 0$ (mean-only model): $\hat{\Sigma} = .013$, $\hat{\Omega} = .031$, $\hat{\sigma} = .181$.
- For $p = q = 1$: $\hat{\Sigma} = .010$, $\hat{\Omega} = .035$, $\hat{\gamma} = .323$, $\hat{\lambda} = .128$, $\hat{\sigma} = .138$.
- For $p = q = 2$: $\hat{\Sigma} = .010$, $\hat{\Omega} = .051$, $\hat{\boldsymbol{\gamma}} = (.344, .021)$, $\hat{\boldsymbol{\lambda}} = (.168, .010)$, $\hat{\sigma} = .124$.
- For $p = q = 3$: $\hat{\Sigma} = .005$, $\hat{\Omega} = .035$, $\hat{\boldsymbol{\gamma}} = (.426, .022, .005)$, $\hat{\boldsymbol{\lambda}} = (.186, .028, .012)$, $\hat{\sigma} = .121$.

Overall, it seems that a single principal component is sufficient to explain amplitude variability, so we chose the model with $p = q = 1$. The fitted curves $\hat{x}_{ij}(t)$ are shown in Figure 2(a) and we see that they provide a good approximation to the data in Figure 1(b). The estimated warping functions $\hat{w}_{ij}(t)$ are shown in Figure 2(b); the time variability around day 15, which is substantial, is captured well by these curves. The amplitude principal components $\hat{\phi}(t)$ and $\hat{\psi}(t)$ are shown in Figure 2(c); to facilitate interpretation of the components we plotted $\hat{\mu}(t)$ together with $\hat{\mu}(t) \pm \hat{\phi}(t)$ in Figure 2(d). We see that $\hat{\phi}(t)$ and $\hat{\psi}(t)$, which are very similar, explain variation in overall mass: individuals with

positive pc scores tend to have trajectories above the mean and individuals with negative pc scores tend to have trajectories below the mean.

The similarity between $\hat{\phi}(t)$ and $\hat{\psi}(t)$ has a biological explanation: the main factor of the ANOVA model represents the genetic contribution of the father, while the residual term represents the genetic contribution of the mother together with environmental factors (see e.g. Heckman, 2003, sec. 3). For a population in Hardy-Weinberg equilibrium, the genetic contribution of both parents is identical, so the $\hat{\phi}_k(t)$s and the $\hat{\psi}_k(t)$s will be similar if the environmental factors are not very strong. Supporting this result is the fact that Irwin and Carter (2013) showed that most of the phenotypic variance was explained by genetic effects in most parts of the growth curve.

The amplitude principal components reveal a very interesting biological result that was not apparent in the original analysis of the raw data in Irwin and Carter (2013): very little variation in amplitude exists at the inflection point at day 15 (Figure 2(c) and 2(d)). This indicates that the beetles have a target peak mass that is reached prior to entry into the wandering phase, which suggests that the target peak mass must be reached before pupation can begin, and that selection for that peak mass (or a related physiological trait) may occur. Interestingly the warping functions shown in Figure 2(b), as well as the original analysis in Irwin and Carter (2013) demonstrate there is substantial variation at the age at which peak mass is reached. In combination these two results provide a basis for future experiments investigating physiological mechanisms, genetic underpinnings and evolutionary implications of size and age of peak mass.

The variance ratios for the amplitude and warping components are $\hat{h}_z = .72$ and $\hat{h}_w = .23$, with respective asymptotic standard deviations $.15$ and $.13$. The bootstrap distributions of $\hat{h}_z$ and $\hat{h}_w$ are shown in Web Appendix D; the bootstrap standard deviations are $.20$ and $.16$ respectively, not far from the asymptotic values, but the Normal approximation is more accurate for the transformations $\arcsin\sqrt{\hat{h}_z}$ and $\arcsin\sqrt{\hat{h}_w}$. The 90% asymptotic

confidence intervals obtained by the back-transformation method are $(.45, .92)$ for h_z and $(.06, .47)$ for h_w. Clearly the father effect is strong on the amplitude component, but weak on the warping component. This can be cross-checked by applying the classical ANOVA F-test on the estimated random effects $\hat{\theta}_{ij}$s: it yields a p-value of 0.058 for the hypothesis of no father effect (the reasonableness of the normality assumption on the random factors is also discussed in Web Appendix D.) The reason the father effect is weak on the warping component is that we removed a lot of time variability by aligning the endpoints at the median pupation day. In fact, the ANOVA F-test on the original endpoints yields a p-value of 0.020, indicating that there is a significant father effect on the date of pupation; this is also supported by Irwin and Carter (2013) demonstrating a highly significant heritability (genetic variance ratio) for date of pupation in the full sample. But once the endpoints are aligned, the time variability that remains, although still substantial, does not have a strong father effect.

If we assume $\phi = \psi$, which is not unreasonable given Figure 2(c), then (6) comes down to $z_{ij}(t) = \mu(t) + (U_i + V_{ij})\phi(t)$ and the classical ANOVA F-test can be applied to $\{\hat{U}_i + \hat{V}_{ij}\}$. This gives a very significant F-value 11.03 with p-value 0.00, confirming that the father effect is very strong on the amplitude variability of the growth curves.

As indicated at the end of Section 3, a more general model with correlations between amplitude and warping components can be set up. We fitted a one-component model with correlations for these data and obtained estimators $\hat{\phi}_1$ and $\hat{\psi}_1$ very similar to the ones obtained above, and correlations $\widehat{\mathrm{corr}}(U_i, \boldsymbol{\eta}_i) = .18$ and $\widehat{\mathrm{corr}}(V_{ij}, \boldsymbol{\xi}_{ij}) = .17$, which do not seem very significant. The statistical significance of these correlations could be studied, for instance, by bootstrap confidence intervals, but for brevity's sake we will not do it here.

Finally, it is important to note that for the unaligned raw data, variation in the length of the larval period and the onset of the wandering phase resulted in crossing of family curves late in the larval period (Irwin and Carter, 2013). After application of the warping method,

the warped curves are aligned by peak body mass at the onset of the wandering phase, resulting in family curves late in the larval period that maintain relative positions similar to early in the larval period. This realignment undoubtedly will facilitate estimation of genetic components of variance, a proposition that we can test in the future.

7 Supplementary materials

Web Appendices referenced in Sections 3–6 and Matlab programs implementing the proposed estimators are available with this paper at the Biometrics website on Wiley Online Library.

Acknowledgements

This research was supported by National Science Foundation grants DMS-1006281 to Daniel Gervini and EF-0328594 to Patrick A. Carter, and by a grant from the National Institute for Mathematical and Biological Synthesis to Patrick A. Carter.

References

Ash, R.B., and Gardner, M.F. (1975). *Topics in Stochastic Processes*. Academic Press.

Bigot, J., and Gadat, S. (2010). A deconvolution approach to estimation of a common shape in a shifted curves model. *The Annals of Statistics* **38** 2422–2464.

Bookstein, F. L. (1997). *Morphometric tools for landmark data: geometry and biology*. Cambridge University Press.

Brumback, L.C., and Lindstrom, M. (2004). Self modeling with flexible, random time transformations. *Biometrics* **60** 461–470.

Chen, H., and Wang, Y. (2011). A penalized spline approach to functional mixed effects model analysis. *Biometrics* **67** 861–870.

Claeskens, G., Silverman, B. W., and Slaets, L. (2010). A multiresolution approach to time warping achieved by a Bayesian prior–posterior transfer fitting strategy. *Journal of the Royal Statistical Society Series B* **72** 673–694.

Di, C. Z., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2009). Multilevel functional principal component analysis. *The Annals of Applied Statistics* **3** 458–488.

Fritsch, F. N., and Carlson, R. E. (1980). Monotone piecewise cubic interpolation. *SIAM Journal of Numerical Analysis* **17** 238–246.

Gervini, D. (2009). Detecting and handling outlying trajectories in irregularly sampled functional datasets. *The Annals of Applied Statistics* **3** 1758–1775.

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

Gervini, D., and Gasser, T. (2005). Nonparametric maximum likelihood estimation of the structural mean of a sample of curves. *Biometrika* **92** 801–820.

Gohberg, I., Goldberg, S., and Kaashoek, M. A. (2003). *Basic Classes of Linear Operators*. Basel: Birkhäuser Verlag.

Gomulkiewicz, R., and Beder, J. H. (1996). The selection gradient of an infinite-dimensional trait. *SIAM Journal of Applied Mathematics* **56** 509–523.

Guo, W. (2002). Functional mixed effects models. *Biometrics* **58** 121–128.

Heckman, N. E. (2003). Functional data analysis in evolutionary biology. In *Recent Advances and Trends in Nonparametric Statistics.* Elsevier.

Huey, R. B., and Kingsolver, J. G. (1989). Evolution of thermal sensitivity of ectotherm performance. *Trends in Ecology and Evolution* **4** 131–135.

Irwin, K.K., and Carter, P.A. (2013). Constraints on the evolution of function-valued traits: a study of growth in *Tribolium casteneum*. *Journal of Evolutionary Biology* (in press).

Izem, R., and Kingsolver, J. G. (2005). Variation in continuous reaction norms: quantifying directions of biological interest. *The American Naturalist*, **166**, 277–289.

Jupp, D. L. B. (1978) Approximation to data by splines with free knots. *SIAM Journal of Numerical Analysis* **15** 328–343.

Kingsolver, J. G., Gomulkiewicz, R., and Carter, P. A. (2002). Variation, selection and evolution of function-valued traits. In *Microevolution Rate, Pattern, Process*, pp. 87–104. Springer.

Kirkpatrick, M., and Heckman, N. (1989). A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. *Journal of Mathematical Biology* **27** 429–450.

Kneip, A., and Engel, J. (1995). Model estimation in nonlinear regression under shape invariance. *The Annals of Statistics* **23** 551–570.

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

Kneip, A., and Ramsay, J. O. (2008). Combining registration and fitting for functional models. *Journal of the American Statistical Association* **103** 1155–1165.

Manikkam, M., Guerrero-Bosagna, C., Tracey, R., Haque, M., and Skinner, M. (2012). Transgenerational actions of environmental compounds on reproductive disease and identification of epigenetic biomarkers of ancestral exposures. *PLoS One 7*.

Meyer, K., and Kirkpatrick, M. (2005). Up hill, down dale: quantitative genetics of curvaceous traits. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**, 1443–1455.

Morris, J. S., and Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society Series B* **68** 179–199.

Müller, H. G. (2008). Functional modeling of longitudinal data. In *Longitudinal data analysis. Handbooks of modern statistical methods*. Chapman & Hall/CRC, New York, pp. 223–252.

Ragland, G.J., and Carter, P.A. (2004). Genetic constraints on the evolution of growth and life history traits in the salamander Ambystoma macrodactylum. *Heredity* **92** 569–578.

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

Rice, J. A. (2004). Functional and longitudinal data analysis: Perspectives on smoothing. *Statistica Sinica* **14** 631–648.

Sangalli, L.M., Secchi, P., Vantini, S., and Vitelli, V. (2010). k-mean alignment for curve clustering. *Computational Statistics and Data Analysis* **54** 1219–1233.

Skinner, M. K., Manikkam, M., and Guerrero-Bosagna, C. (2010). Epigenetic transgenerational actions of environmental factors in disease etiology. *Trends in Endocrinology and Metabolism* **21** 214–222.

Tang, R., and Müller, H. G. (2008). Pairwise curve synchronization for functional data. *Biometrika* **95** 875–889.

Telesca, D., and Inoue, L. Y. (2008). Bayesian hierarchical curve registration. *Journal of the American Statistical Association* **103** 328–339.

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

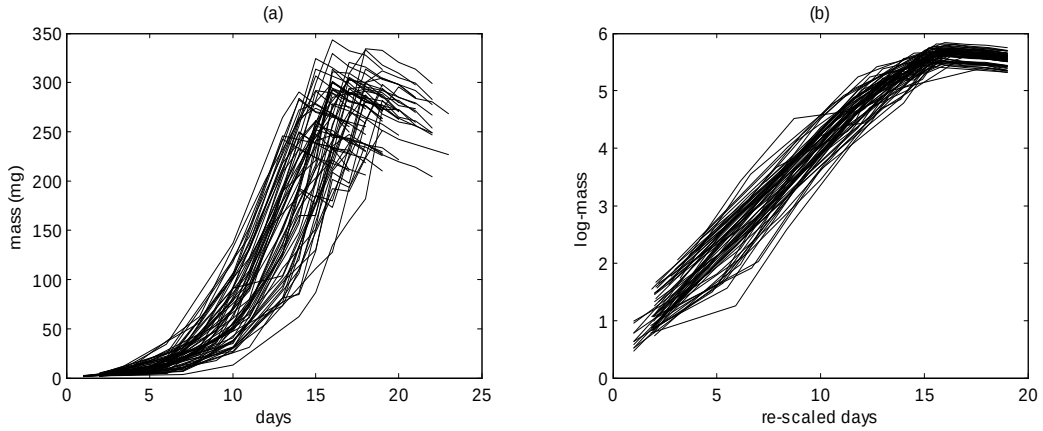

Figure 1: Flour Beetle Growth Example. (a) Raw mass trajectories; (b) log-trajectories re-scaled to common endpoint.

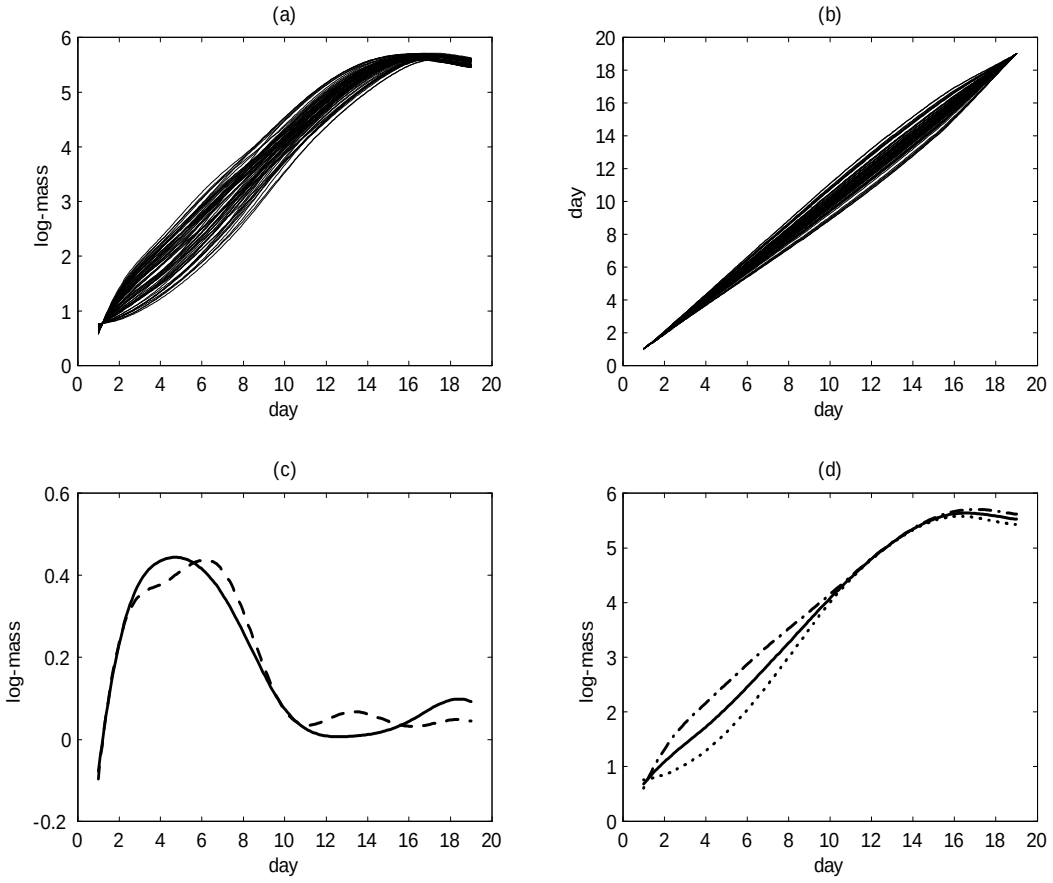

Figure 2: Flour Beetle Growth Example. (a) Fitted trajectories using warped ANOVA model; (b) warping functions; (c) principal component of the main factor, $\hat{\phi}(t)$ (solid line), and of the residual term, $\hat{\psi}(t)$ (dashed line); (d) estimated mean $\hat{\mu}(t)$ (solid line), $\hat{\mu}(t) + \hat{\phi}(t)$ (dash-dot line), and $\hat{\mu}(t) - \hat{\phi}(t)$ (dotted line).

	Model 1									Model 2								
	bias			sd			rmse			bias			sd			rmse		
	C	2s	ML	C	2s	ML	C	2s	ML	C	2s	ML	C	2s	ML	C	2s	ML
$\hat{\mu}$	.008	.016	.008	.055	.055	.055	.055	.057	.056	.008	.017	.008	.051	.052	.057	.051	.054	.058
$\hat{\phi}_1$	.007	.127	.048	.064	.067	.069	.064	.143	.084	.172	.219	.130	.078	.077	.101	.189	.233	.165
$\hat{\psi}_1$	.014	.127	.032	.120	.121	.116	.121	.175	.121	.013	.184	.050	.112	.111	.136	.113	.215	.145
	Model 3									Model 4								
$\hat{\mu}$	.073	.033	.013	.081	.083	.088	.109	.089	.089	.071	.032	.014	.085	.085	.093	.111	.091	.094
$\hat{\phi}_1$	.175	.129	.051	.530	.109	.145	.559	.168	.154	.109	.213	.087	.422	.108	.173	.436	.239	.194
$\hat{\psi}_1$	.327	.126	.027	.736	.145	.167	.806	.192	.169	1.095	.192	.057	.293	.135	.226	1.134	.235	.233
	Model 5									Model 6								
$\hat{\mu}$	.106	.041	.024	.090	.094	.119	.139	.103	.122	.104	.048	.023	.088	.094	.113	.136	.106	.116
$\hat{\phi}_1$	.226	.233	.114	.603	.135	.302	.644	.269	.323	.179	.282	.047	.521	.130	.245	.551	.310	.250
$\hat{\psi}_1$	.468	.230	.082	.838	.163	.263	.960	.282	.275	1.013	.380	.154	.329	.281	.362	1.065	.473	.393
	Model 7									Model 8								
$\hat{\mu}$	.070	.042	.020	.093	.099	.129	.116	.108	.130	.072	.036	.016	.083	.086	.100	.110	.093	.101
$\hat{\phi}_1$	.082	.244	.080	.322	.122	.195	.332	.273	.211	.132	.242	.096	.385	.111	.182	.407	.266	.206
$\hat{\psi}_1$	.874	.205	.043	.389	.154	.221	.957	.257	.225	1.001	.199	.069	.354	.126	.205	1.061	.235	.217
	Model 9									Model 10								
$\hat{\mu}$	.072	.035	.019	.086	.084	.105	.112	.091	.106	.106	.059	.042	.097	.100	.136	.143	.116	.143
$\hat{\phi}_1$	.242	.206	.172	.633	.337	.559	.678	.395	.585	.346	.262	.236	.663	.353	.637	.748	.439	.679
$\hat{\phi}_2$	.357	.602	.418	.571	.903	.801	.673	1.085	.904	.503	.708	.469	.742	.825	.845	.896	1.087	.967
$\hat{\psi}_1$	.383	.258	.139	.761	.209	.427	.852	.332	.449	.521	.312	.156	.835	.232	.510	.984	.389	.533
$\hat{\psi}_2$	.839	.321	.387	.888	.284	.771	1.222	.429	.863	.780	.552	.448	.910	.471	.826	1.211	.726	.940

Table 1: Simulation Results. Bias, standard deviation and root mean squared error for common (C), two-step (2s) and maximum likelihood (ML) ANOVA estimators.