Biometrika (2014), **xx**, x, *pp*. 1–26

Printed in Great Britain

Warped functional regression

BY DANIEL GERVINI

Department of Mathematical Sciences, University of Wisconsin–Milwaukee,

PO Box 413, Milwaukee, Wisconsin 53201, USA

gervini@uwm.edu

SUMMARY

A characteristic feature of functional data is the presence of phase variability in addition to amplitude variability. Existing functional regression methods do not handle time variability in an explicit and efficient way. In this paper we introduce a functional regression method that incorporates time warping as an intrinsic part of the model. The method achieves good predictive power in a parsimonious way and allows unified statistical inference about phase and amplitude components. The asymptotic distribution of the estimators is derived and the finite-sample properties are studied by simulation. An example of application involving ground-level ozone trajectories is presented.

Some key words: Functional Data Analysis; Random-Effects Model; Registration; Spline Smoothing; Time Warping.

1. INTRODUCTION

The analysis of data consisting of curves or other types of functions, rather than scalars or vectors, is increasingly common (Ramsay & Silverman, 2005). Many such problems involve modeling curves as functions of other curves. For example, Figure 1(a) shows daily trajectories

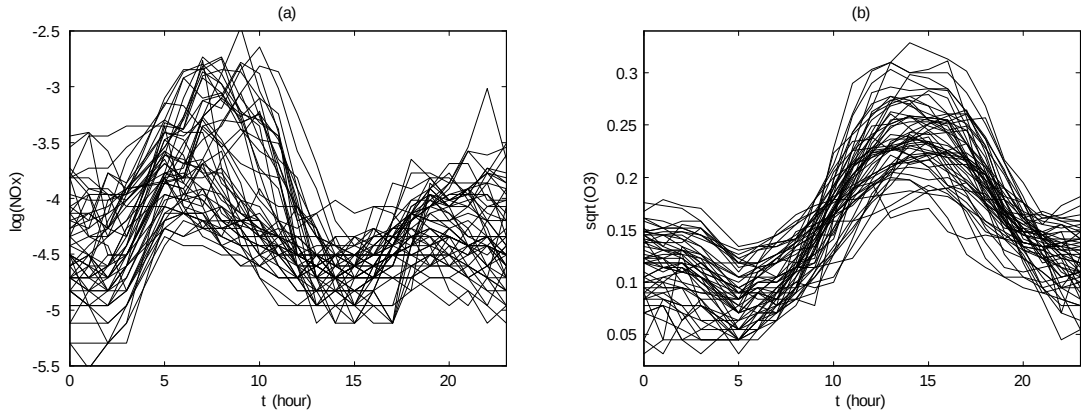

Fig. 1. Ozone Example. Daily trajectories of ground-level concentrations of (a) oxides of nitrogen and (b) ozone in the city of Sacramento in the Summer of 2005.

of oxides of nitrogen in the city of Sacramento, California, for 52 summer days in the year 2005, and Figure 1(b) shows the corresponding trajectories of ozone concentration. The goal is to predict ozone concentration from oxides of nitrogen.

Functional linear regression models are normally used for such problems (Ramsay & Silverman, 2005, ch. 16). Recent papers have studied different aspects of the functional linear regression model (Yao et al., 2005; Cai & Hall, 2006; Hall & Horowitz, 2007; Crambes et al., 2009; James et al., 2009). However, a characteristic feature of functional data that has not been widely investigated in a regression context is phase variability. Functional samples often present a few distinct features, such as peaks and valleys, which vary in amplitude and location from curve to curve, as is clear in Figure 1. Functional linear regression is usually based on functional principal components, which are well suited for fitting amplitude variability but not for location or phase variability. It may take an inordinate number of principal components to account even for very basic phase-variability processes (Ramsay & Silverman, 2005, ch. 7). A more efficient strategy is to model amplitude and phase variability separately: the former using traditional functional principal components and the latter using warping models. This approach is more efficient, because

the combined model often provides a better fit with fewer parameters than does the classical principal component decomposition. It is also more informative, because it provides direct information about the warping process, which classical principal components only do indirectly. Several warping methods have been proposed over the years (Gervini & Gasser, 2004, 2005; James, 2007; Kneip et al., 2000; Kneip & Ramsay, 2008; Liu & Müller, 2004; Ramsay & Li, 1998; Tang & Müller, 2008, 2009; Wang & Gasser, 1999).

Common functional linear regression models inherit the problems of functional principal components in presence of phase variability. Although a high-dimensional model based on a large number of principal components can provide a good fit to the data, the problem is again one of efficiency and interpretability, not just minimizing prediction error. It is usually hard to extract specific information about phase variability from a traditional functional regression model because the model confounds phase and amplitude variation.

The curves in Figure 1, for example, show peaks that vary not only in amplitude but also in location. It is reasonable to hypothesize that a large peak in oxides of nitrogen will be followed by a large peak in ozone concentration, and also that an early peak in oxides of nitrogen will be followed by an early peak in ozone level. Perhaps there may also be an interaction between timing and amplitude of the peaks. A common functional linear regression model of sufficiently high dimension will be able to fit these data well from the point of view of prediction error, but will not provide clear answers to these conjectures. A regression model that explicitly incorporates a warping component and does not confound the two sources of variability will be more useful for this, and that is what we propose in this paper.

2. THE WARPED FUNCTIONAL REGRESSION MODEL

2·1. *Model specification*

Consider a sample of functions $(x_1(s), y_1(t))$, ..., $(x_n(s), y_n(t))$, where $x_i(s)$ is the covariate and $y_i(t)$ the response, with $x_i : \mathscr{S} \to \mathbb{R}$ and $y_i : \mathscr{T} \to \mathbb{R}$, and $\mathscr{S}$ and $\mathscr{T}$ are closed intervals in $\mathbb{R}$. The functions $x_i(s)$ and $y_i(t)$ are usually not directly observable; instead we observe discretizations of them, with added random noise, at time grids $\{s_{ij} : j = 1, \ldots, \nu_{1i}\}$ and $\{t_{ij} : j = 1, \ldots, \nu_{2i}\}$. Thus the observed data consist of vectors $(x_1, y_1), \ldots, (x_n, y_n)$, with $x_i \in \mathbb{R}^{\nu_{1i}}$ and $y_i \in \mathbb{R}^{\nu_{2i}}$ with elements

$$x_{ij} = x_i(s_{ij}) + \varepsilon_{ij}, \;\; j = 1, \ldots, \nu_{1i}, \; i = 1, \ldots, n, \tag{1}$$

$$y_{ij} = y_i(t_{ij}) + \eta_{ij}, \;\; j = 1, \ldots, \nu_{2i}, \; i = 1, \ldots, n. \tag{2}$$

We will assume that the measurement errors $\{\varepsilon_{ij}\}$ and $\{\eta_{ij}\}$ are independent with $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ and $\eta_{ij} \sim N(0, \sigma_\eta^2)$.

The kind of curves we have in mind for our model will present a relatively small number of peaks and valleys that systematically appear in all curves but vary in amplitude and location. Then $\{x_i(s)\}$ and $\{y_i(t)\}$ can be thought of as compound processes

$$x_i(s) = x_i^*\{\omega_i^{-1}(s)\}, \tag{3}$$

$$y_i(t) = y_i^*\{\zeta_i^{-1}(t)\}, \tag{4}$$

where $\{x_i^*(s)\}$ and $\{y_i^*(t)\}$ account for amplitude variability and $\{\omega_i(s)\}$ and $\{\zeta_i(t)\}$ account for phase variability. The ω_is and the ζ_is are monotone increasing warping functions with $\omega_i : \mathscr{S} \to \mathscr{S}$ and $\zeta_i : \mathscr{T} \to \mathscr{T}$. The aligned processes $\{x_i^*(s)\}$ and $\{y_i^*(t)\}$ follow principal-

component decompositions

$$x_i^*(s) = \mu_x(s) + \sum_{k=1}^{p_1} u_{ik}\phi_k(s), \tag{5}$$

$$y_i^*(t) = \mu_y(t) + \sum_{l=1}^{p_2} v_{il}\psi_l(t), \tag{6}$$

with $\{\phi_k(s)\}$ and $\{\psi_l(t)\}$ orthonormal functions in $L^2(\mathscr{S})$ and $L^2(\mathscr{T})$, respectively, and $\{u_{ik}\}$ and $\{v_{il}\}$ uncorrelated zero-mean random variables.

A few comments about (3)–(6) are in order, because models (3) and (4) may seem unidentifiable and models (5) and (6) may seem too restrictive for finite p_1 and p_2. These issues are extensively discussed in Kneip & Ramsay (2008, sec. 2.3) and in the Supplementary Material. Proposition 1 in Kneip & Ramsay (2008) shows that if the x_is have at most K peaks and valleys and their derivatives $x_i'(t)$ have at most K zeros, then $x_i(t)$ admits the decomposition $x_i(t) = \sum_{j=1}^{p} C_{ij}\xi_j\{v_i(t)\}$ for some $p \leq K + 2$, where the ξ_js are non-random basis functions, the C_{ij}s are random coefficients, and the v_is are warping functions. Orthogonalizing the ξ_js leads to model (5). Then p_1 in (5) and p_2 in (6) need not be large if the number of features to be aligned is small. The identifiability of (3) and (4) given amplitude models (5) and (6) and given certain conditions on the warping family $\mathscr{W}$ is shown in the Supplementary Material. If the summations in (5) and (6) were allowed to be infinite, then (3) and (4) would be unidentifiable. The practical effect of large p_1 and p_2 in (5) and (6) is that the sample curves tend to present a large and unequal number of features, and then it does not make sense to try to align them; in such cases amplitude and phase variability essentially become indistinguishable. Samples like that do occur in practice, but the methods we propose in this paper are not intended for those situations.

The warping functions $\{\omega_i(s)\}$ and $\{\zeta_i(t)\}$ will be modelled as monotone Hermite splines (Fritsch & Carlson, 1980). Although other families, such as integrated splines (Ramsay, 1988), monotone splines (Ramsay & Li, 1998) and constrained B-splines (Brumback & Lindstrom,

2004) could be used, monotone Hermite splines are better suited for the regression approach proposed here. Details about this family of warping functions are given in the Appendix. We only mention here that, like other spline families, this one is a finite-dimensional semiparametric family determined by a knot sequence chosen by the user. Thus, the family $\{\omega_i(s)\}$ is determined by a knot sequence $\tau_{x0} = (\tau_{x01}, \ldots, \tau_{x0r_1})$ of strictly increasing points in $\mathscr{S}$, and each $\omega_i(s)$ is determined by a corresponding sequence τ_{xi} of basis coefficients that satisfy $\omega_i(\tau_{x0j}) = \tau_{xij}$ for $j = 1, \ldots, r_1$. Similarly, the family $\{\zeta_i(t)\}$ is determined by a knot sequence $\tau_{y0} = (\tau_{y01}, \ldots, \tau_{y0r_2})$ of strictly increasing points in $\mathscr{T}$ and each $\zeta_i(t)$ is determined by basis coefficients τ_{yi} that satisfy $\zeta_i(\tau_{y0j}) = \tau_{yij}$ for $j = 1, \ldots, r_2$. The dual role of the τ_{xi}s and the τ_{yi}s as basis coefficients and as values of $\omega_i(s)$ and $\zeta_i(t)$ at the knots is what makes Hermite splines appealing. We choose the knot sequences τ_{x0} and τ_{y0} as the approximate average location of the main features of the x_is and the y_is, such as peaks and valleys, so r_1 and r_2 will not be large for the type of applications we envision.

Unlike landmark registration, where the τ_{xi}s and the τ_{yi}s are individually estimated curve by curve, we will treat the τ_{xi}s and the τ_{yi}s as latent random effects, so they will not be estimated directly. This is a big advantage in practice, since individual estimation of the τ_{xi}s and the τ_{yi}s is difficult when the number of curves is large or when the curves are sparsely sampled. A minor complication is that the τ_{xi}s and the τ_{yi}s are constrained to be monotone increasing in $\mathscr{S}$ and $\mathscr{T}$, respectively, so for convenience we will work with their Jupp transforms θ_{xi} and θ_{yi} instead, which are unconstrained vectors; the Jupp transform is defined in the Appendix.

Since the warping functions $\{\omega_i\}$ and $\{\zeta_i\}$ are determined by the random effects θ_{xi} and θ_{yi}, and the amplitude functions $\{x_i^*\}$ and $\{y_i^*\}$ are determined by the random effects u_i and v_i, we

can specify an indirect regression model of the y_is on the x_is via the random effects:

$$\begin{pmatrix} v_i \\ \\ \theta_{yi} \end{pmatrix} = \begin{pmatrix} 0 \\ \\ \theta_{y0} \end{pmatrix} + A\left\{ \begin{pmatrix} u_i \\ \\ \theta_{xi} \end{pmatrix} - \begin{pmatrix} 0 \\ \\ \theta_{x0} \end{pmatrix} \right\} + e_i, \tag{7}$$

where A is the $(p_2 + r_2) \times (p_1 + r_1)$ regression matrix and e_i is an error term, which we assume $N(0, \Sigma_e)$ with Σ_e diagonal. For interpretability we split A into blocks corresponding to u_i, θ_{xi}, v_i and θ_{yi}:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ \\ A_{21} & A_{22} \end{pmatrix},$$

with $A_{11} \in \mathbb{R}^{p_2 \times p_1}$, $A_{12} \in \mathbb{R}^{p_2 \times r_1}$, $A_{21} \in \mathbb{R}^{r_2 \times p_1}$ and $A_{22} \in \mathbb{R}^{r_2 \times r_1}$. Then (5), (6) and (7) imply that

$$y_i^*(t) - \mu_y(t) = \int \beta(s,t)\{x_i^*(s) - \mu_x(s)\}\, \mathrm{d}s + \gamma_1(t)^T(\theta_{xi} - \theta_{x0}) + \delta_i(t), \tag{8}$$

$$\theta_{yi} - \theta_{y0} = \int \gamma_2(s)\{x_i^*(s) - \mu_x(s)\}\, \mathrm{d}s + A_{22}(\theta_{xi} - \theta_{x0}) + e_{i2}, \tag{9}$$

where $\beta(s,t) = \psi(t)^T A_{11} \phi(s)$, $\gamma_1(t)^T = \psi(t)^T A_{12}$, $\gamma_2(s) = A_{21}\phi(s)$ and $\delta_i(t) = \psi(t)^T e_{i1}$. Thus, for example, $A_{12} = 0$ implies that $\gamma_1(t) = 0$ and then the amplitude variability of the responses is unrelated to the time variability of the covariates; similarly, $A_{21} = 0$ implies that $\gamma_2(s) = 0$ and then the time variability of the responses is unrelated to the amplitude variability of the covariates.

2·2. *Estimation and prediction*

Models (5) and (6) depend on functional parameters that need to be estimated: the mean functions $\mu_x(s)$ and $\mu_y(t)$ and the principal components $\{\phi_k(s)\}$ and $\{\psi_l(t)\}$. Let $b_x(s) = (b_{x1}(s), \ldots, b_{xq_1}(s))^T$ be a B-spline basis in $L^2(\mathscr{S})$ and $b_y(t) = (b_{y1}(t), \ldots, b_{yq_2}(t))^T$ a B-spline basis in $L^2(\mathscr{T})$. Let $\mu_x(s) = b_x^T(s) m_x$, $\mu_y(t) = b_y^T(t) m_y$, $\phi_k(s) = b_x^T(s) c_k$ and $\psi_l(t) = b_y^T(t) d_l$, for $m_x \in \mathbb{R}^{q_1}$, $m_y \in \mathbb{R}^{q_2}$, $c_k \in \mathbb{R}^{q_1}$ and $d_l \in \mathbb{R}^{q_2}$. The orthogonality re-

strictions on the ϕ_ks and the ψ_ls can be expressed as $C^T J_x C = I_{p_1}$ and $D^T J_y D = I_{p_2}$, where $C = (c_1, \ldots, c_{p_1}) \in \mathbb{R}^{q_1 \times p_1}$, $D = (d_1, \ldots, d_{p_2}) \in \mathbb{R}^{q_2 \times p_2}$, $J_x = \int b_x(s) b_x^T(s) \mathrm{d}s$ and $J_y = \int b_y(t) b_y^T(t) \mathrm{d}t$.

If the curves $\{x_i\}$ and $\{y_i\}$ were observed on dense time grids and individual smoothing were possible, the spline coefficients and the rest of the model parameters could be estimated by least squares. However, we are more interested in applications where the trajectories are not densely sampled, so we will treat u_i, v_i, θ_{xi} and θ_{yi} as latent variables and estimate the model parameters by maximum likelihood. We assume $w_i = (u_i^T, \theta_{xi}^T)^T$ is jointly multivariate normal of dimension $d_1 = p_1 + r_1$, with mean and covariance

$$\mu_w = \begin{pmatrix} 0 \\ \\ \theta_{x0} \end{pmatrix}, \quad \Sigma_w = \begin{pmatrix} \Lambda & \Sigma_{u\theta_x} \\ \\ \Sigma_{u\theta_x}^T & \Sigma_{\theta_x} \end{pmatrix},$$

where θ_{x0} the Jupp transform of the knot vector τ_{x0} and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_{p_1})$. This and model (7) imply that $z_i = (v_i^T, \theta_{yi}^T)^T$ is multivariate normal of dimension $d_2 = p_2 + r_2$ with mean $\mu_z = (0^T, \theta_{y0}^T)^T$ and covariance $\Sigma_z = A \Sigma_w A^T + \Sigma_e$, where θ_{y0} is the Jupp transform of the knot vector τ_{y0}. Thus $v_i \sim N(0, \Gamma)$ with $\Gamma = A_{1\cdot} \Sigma_w A_{1\cdot}^T + \Sigma_{e,11}$, where $A_{1\cdot} = [A_{11}, A_{12}]$ and $\Sigma_{e,11}$ the $p_2 \times p_2$ upper-left diagonal block of Σ_e. Since Γ has to be diagonal by model (6), and Σ_e was assumed diagonal, it follows that $A_{1\cdot} \Sigma_w A_{1\cdot}^T$ must be diagonal, which imposes an additional restriction on the parameters.

To summarize, the parameters of this model are the regression matrix A, the residual covariance matrix Σ_e, the covariance matrix Σ_w of the explanatory random effects w_i, the spline coefficients m_x, m_y, C and D of the functional parameters, and the variances σ_ε^2 and σ_η^2 of the random noise in (1) and (2). The derivations of the likelihood function and the EM algorithm to compute these estimators are discussed in the Supplementary Material.

In addition to the model parameters there are meta-parameters that need to be chosen by the user, such as the dimension and knot placement of the B-spline bases for the functional parameters. This can be done either subjectively or by cross-validation. Since the method borrows strength across curves, it is possible to use a larger number of knots than would be practical for single-curve smoothing. The other meta-parameters that need to be specified are the numbers of components in models (5) and (6), p_1 and p_2, and the warping dimensions r_1 and r_2. As already discussed, these quantities should roughly correspond to the number of salient features of the x_is and the y_is, which gives a range of reasonable values for p_1, p_2, r_1 and r_2; within this range, the final choice may be done objectively by cross-validation, as we do in § 5.

In addition to parameter estimation, it is usually of interest to predict a response curve for a given covariate curve. This can be done in a straightforward way. Given a covariate data vector x_{n+1}, obtained by discretizing a covariate curve $x_{n+1}(s)$ on some time grid, the predictors $\hat{v}_{n+1}$ and $\hat{\theta}_{y,n+1}$ of the response random effects are given by $\hat{E}(v_{n+1} \mid x_{n+1})$ and $\hat{E}(\theta_{y,n+1} \mid x_{n+1})$, which under model (7) come down to $\hat{v}_{n+1} = \hat{A}_{11}\hat{E}(u_{n+1} \mid x_{n+1}) + \hat{A}_{12}\{\hat{E}(\theta_{x,n+1} \mid x_{n+1}) - \theta_{x0}\}$ and $\hat{\theta}_{y,n+1} = \hat{A}_{21}\hat{E}(u_{n+1} \mid x_{n+1}) + \hat{A}_{22}\{\hat{E}(\theta_{x,n+1} \mid x_{n+1}) - \theta_{x0}\}$. With $\hat{v}_{n+1}$ and $\hat{\theta}_{y,n+1}$ we compute $\hat{y}^*_{n+1}(t)$ and $\hat{\zeta}_{n+1}(t)$ respectively, and then $\hat{y}_{n+1}(t) = \hat{y}^*_{n+1}\{\hat{\zeta}^{-1}_{n+1}(t)\}$.

3. INFERENCE

Consider now the asymptotic distribution of $\hat{A}$ when the number of curves n goes to infinity. For simplicity, we will assume that the time grids are equal for all individuals, which makes the raw data vectors $(x_1, y_1), \ldots, (x_n, y_n)$ independent and identically distributed. We will also assume that the functional parameters belong to the spline space used for estimation, whose dimension is held fixed.

The asymptotic analysis is not entirely straightforward due to the parameter constraints, so we use the results of Geyer (1994). Since we are only interested in the marginal asymptotic distribution of $\hat{A}$ and not in the asymptotic covariance between $\hat{A}$ and the rest of the parameters, we can assume without loss of generality that Σ_e, m_x, m_y, C, D, σ_ε^2 and σ_η^2 are fixed and known, because this assumption does not alter the asymptotic covariance matrix of $\hat{A}$. However, in principle we cannot assume that Σ_w is fixed and known because Σ_w is part of the condition that $A_{1\cdot}\Sigma_w A_{1\cdot}^T$ be diagonal. So we will derive the joint asymptotic distribution of $\hat{A}$ and $\hat{\Sigma}_w$, even though we are only interested in the marginal distribution of $\hat{A}$.

The parameter of interest is then, in vector form, $\xi = (\mathrm{vec}(A^T)^T, \mathrm{v}(\Sigma_w)^T)^T$, where $\mathrm{v}(\Sigma_w)$ denotes the vec of the lower-triangular part of Σ_w, including the diagonal. The dimension of ξ is $d = d_1 d_2 + d_1(d_1+1)/2$. The restriction that $A_{1\cdot}\Sigma_w A_{1\cdot}^T$ be diagonal can be expressed as a system of $m = (p_2 - 1)p_2/2$ constraints of the form $h_{ij}(\xi) = 0$, where $h_{ij}(\xi) = a_i^T \Sigma_w a_j$ and a_i^T is the ith row of A. The functions h_{ij} can be stacked together into a single vector-valued function $h : \mathbb{R}^d \to \mathbb{R}^m$, and the constrained parameter space can be expressed as $C = \{\xi \in \mathbb{R}^d : h(\xi) = 0\}$. The additional condition that Σ_w be positive definite does not alter the asymptotic distribution of the estimator because Σ_w lies in the interior of this space, not on the border. Let ξ_0 be the true value of the parameter ξ. Since $h(\xi)$ is continuously differentiable, the tangent cone of C at ξ_0 is $T_C(\xi_0) = \{\delta \in \mathbb{R}^d : \mathrm{D}h(\xi_0)\delta = 0\}$, where D is the differential (Rockafellar & Wets, 1998, ch. 6.B). The asymptotic distribution of the constrained estimator $\hat{\xi}_n$ is simple in this case: it is just the usual asymptotic normal distribution of an unconstrained maximum likelihood estimator, projected on $T_C(\xi_0)$.

Specifically, let

$$M(x,y) = E\{(w-\mu_w)(w-\mu_w)^T|(x,y)\}, \tag{10}$$

$$N(x,y) = E\{(w-\mu_w)(z-\mu_z)^T|(x,y)\}, \tag{11}$$

and

$$U(x,y) = \begin{bmatrix} \mathrm{vec}\{N(x,y)\Sigma_{e,0}^{-1}\} - \mathrm{vec}\{M(x,y)A_0^T\Sigma_{e,0}^{-1}\} \\ (-1/2)D_{d_1}^T \mathrm{vec}\{\Sigma_{w,0}^{-1} - \Sigma_{w,0}^{-1}M(x,y)\Sigma_{w,0}^{-1}\} \end{bmatrix}, \tag{12}$$

where D_{d_1} is the duplication matrix that satisfies $\mathrm{vec}(\Sigma_w) = D_{d_1}\mathrm{v}(\Sigma_w)$ (Magnus & Neudecker, 1999, ch. 3). It is shown in the Supplementary Material that $U(x,y)$ is the likelihood score function $\nabla_\xi \log f(x,y;\xi)$ at $\xi = \xi_0$. Let $B = \mathrm{D}h(\xi_0)$, which is an $m \times d$ matrix of rank m with rows $\nabla h_{ij}(\xi)^T = (a_i^T\Sigma_w(e_j \otimes I_{d_1}) + a_j^T\Sigma_w(e_i \otimes I_{d_1}), 0_{r_2 d_1}^T, (a_j^T \otimes a_i^T)D_{d_1})$, where e_i is the ith canonical vector in $\mathbb{R}^{p_2}$. Let Ξ be an orthogonal $d \times (d-m)$ matrix of rank $d-m$ such that $B\Xi = 0$, which can be computed for instance via the singular value decomposition of the orthogonal projector $I_d - B^T(BB^T)^{-1}B$; this matrix is not unique but Theorem 1 below is invariant under the choice of Ξ.

THEOREM 1. *Under all of the above conditions, the asymptotic distribution of* $n^{1/2}(\hat{\xi}_n - \xi_0)$ *is* $N\{0, \Xi(\Xi^T V \Xi)^{-1}\Xi^T\}$, *where* $V = E\{U(x,y)U(x,y)^T\}$.

Matrix V in Theorem 1 is Fisher's information matrix for this model and can be estimated by

$$\hat{V}_n = \frac{1}{n}\sum_{i=1}^{n} \hat{U}(x_i,y_i)\hat{U}(x_i,y_i)^T, \tag{13}$$

where the hat in U denotes that the true parameters in (12) are replaced by their estimators. The proof of Theorem 1 is given in the Supplementary Material. Note that a parametric rate of convergence is obtained because the functional parameters are assumed to belong to spline spaces of dimensions that do not increase with n.

The assumption that the time grids were equal for all individuals was a simplification to make the data vectors (x_i, y_i), and consequently the likelihood scores (12), identically distributed. In many applications this will not be the case, however, and the time grids will be unequal, giving $x_i \in \mathbb{R}^{\nu_{1i}}$ and $y_i \in \mathbb{R}^{\nu_{2i}}$ which are still independent but not identically distributed due

to the different dimensions. Usually this does not affect the final asymptotic result as long as (13) does not become degenerate, as shown for instance by Pollard (1990, Ch. 11) in the context of regression with non-random covariates. Although the Fisher information matrix V as such does not exist, (10) and (11) and consequently (12) and (13) can still be computed with (x_i, y_i)s of unequal dimensions. The statement of Theorem 1 should then be re-expressed as $n^{1/2}\{\Xi(\Xi^T\hat{V}_n\Xi)^{-1}\Xi^T\}^{-1/2}(\hat{\xi}_n - \xi_0) \longrightarrow N(0, I_d)$ in distribution as $n \to \infty$. The finite-sample adequacy of the asymptotic results of this section, particularly for hypothesis testing, were studied by simulations which are reported in the Supplementary Material.

4. SIMULATIONS

4·1. *Estimation accuracy*

To study the finite-sample accuracy of the proposed estimators we simulated data from the following six models.

Model 1 is a one-dimensional amplitude and warping model, with $\mu_x(s) = 0.6\varphi(s, 0.3, 0.1) + 0.4\varphi(s, 0.6, 0.1)$, $\phi_1(s) = \varphi(s, 0.3, 0.1)/1.6796$, $\mu_y(t) = 0.6\varphi(t, 0.5, 0.1) + 0.4\varphi(t, 0.8, 0.1)$ and $\psi_1(t) = \varphi(t, 0.5, 0.1)/1.6796$, for s and t in $[0, 1]$, where $\varphi(s, \mu, \sigma)$ denotes the $N(\mu, \sigma^2)$ density function. The warping functions followed Hermite spline models with knots $\tau_{x0} = 0.3$ and $\tau_{y0} = 0.5$. Thus, although $\mu_x(s)$ and $\mu_y(t)$ have two peaks, phase and amplitude variability are concentrated on the main peak. The regression matrix A was the identity matrix, so there was no relationship between covariate phase variability and response amplitude variability, or vice versa, in this model. The other parameters were $\Sigma_w = \mathrm{diag}(0.2^2, 0.1^2)$, $\Sigma_e = 0.07^2 I_2$, and $\sigma_\varepsilon = \sigma_\eta = 0.05$. Model 2 is the same as Model 1 but with a non-diagonal A; specifically, $a_{11} = a_{22} = 1$ and $a_{12} = a_{21} = 0.5$, so there was a relationship between covariate phase variability and response amplitude variability, and vice versa, in this model.

Model 3 is a two-dimensional amplitude and warping model, with $\mu_x(s)$, $\mu_y(t)$, $\phi_1(s)$ and $\psi_1(t)$ as in Model 1, $\phi_2(s)$ the function $\varphi(s, 0.6, 0.1)$ orthogonalized with $\phi_1(s)$, and $\psi_1(t)$ the function $\varphi(t, 0.8, 0.1)$ orthogonalized with $\psi_1(t)$. The warping functions followed Hermite spline models with knots $\tau_{x0} = (0.3, 0.6)$ and $\tau_{y0} = (0.5, 0.8)$. This model, then, has amplitude and phase variability at both peaks of $\mu_x(s)$ and $\mu_y(t)$. The regression matrix A was the identity, and the other parameters were $\Sigma_w = \mathrm{diag}(0.2^2, 0.1^2, 0.1^2, 0.1^2)$, $\Sigma_e = 0.07^2 I_4$, and $\sigma_\varepsilon = \sigma_\eta = 0.05$. Model 4 is the same as Model 3 but with a non-diagonal regression matrix A, with blocks $A_{11} = A_{22} = I_2$ and $A_{12} = A_{21} = 0.5 I_2$.

Model 5 is a one-dimensional amplitude model like Model 1 but with warping functions that do not follow a regression model and do not belong to the Hermite-spline family; they belong to a generic B-spline family with monotone increasing coefficients, which produces monotone increasing functions (Brumback & Lindstrom, 2004). Specifically, if $b(s)$ are cubic B-splines with 7 equally-spaced knots in $(0, 1)$ and c_0 is such that $b(s)^T c_0 \equiv s$, the identity, then we generated $c_i \sim N(c_0, 0.05^2 I_9)$ and took $\omega_i^{-1}(s) = \{g_i(s) - g_i(0)\}/\{g_i(1) - g_i(0)\}$, with $g_i(s) = b(s)^T c_{(i)}$ and $c_{(i)}$ the coefficients of c_i sorted in increasing order. The inverse warping functions of the responses, the $\zeta_i^{-1}(t)$s, were generated in an analogous way and were independent of the $\omega_i^{-1}(s)$s. Model 6 is a two-dimensional amplitude model like Model 3 with a non-Hermite warping model like Model 5.

Two sample sizes, $n = 50$ and $n = 100$, were considered for each model. Each scenario was replicated 500 times. In all cases the time grids $\{s_{i1}, \ldots, s_{i\nu_{1i}}\}$ and $\{t_{i1}, \ldots, t_{i\nu_{2i}}\}$ were random and irregular, with ν_{1i} and ν_{2i} uniformly distributed between 10 and 20, and independent of one another, and s_{ij} and t_{ij} uniformly distributed on $[0, 1]$.

For each sample we computed the proposed warped functional regression estimator using cubic B-splines with ten equally spaced knots for the functional parameters, with the number of

principal components p_1 and p_2 equal to the true model quantities, that is, $p_1 = p_2 = 1$ for Models 1, 2 and 5, and $p_1 = p_2 = 2$ for Models 3, 4 and 6. The specification of the warping functions, although always in a Hermite-spline family, varied from model to model. For Models 1 and 2 we used the same family used for estimation. For Models 3 and 4, however, we used Hermite-spline families with single knots at $\tau_{x0} = 0.45$ and $\tau_{y0} = 0.65$, so as to study the behavior of the estimator when the number of warping knots is underspecified. For Model 5 we used Hermite splines with knots at $\tau_{x0} = (0.3, 0.6)$ and $\tau_{y0} = (0.5, 0.8)$, and for Model 6 we used Hermite splines with knots at $\tau_{x0} = 0.45$ and $\tau_{y0} = 0.65$; this allows us to study the advantages of doing some kind of warping as opposed to not doing any warping at all, since the true warping processes of Models 5 and 6 do not follow a regression model and do not belong to the Hermite spline family.

For comparison we also computed ordinary functional regression estimators based on principal components, as in, e.g., Müller et al. (2008), with the difference that the principal components were computed by maximum likelihood via B-spline models, as in James et al. (2000), rather than by kernel smoothing.

As measures of performance we computed bias and root mean squared errors of $\hat{\beta}(s,t)$, $\hat{\mu}_x(s)$, $\hat{\mu}_y(t)$, $\{\hat{\phi}_j(s)\}$ and $\{\hat{\psi}_j(t)\}$. We defined as bias of $\hat{\mu}_x$ the quantity $(\int [E\{\hat{\mu}_x(s)\} - \mu_x(s)]^2 \mathrm{ds})^{1/2}$ and as root mean squared error the quantity $(\int E[\{\hat{\mu}_x(s) - \mu_x(s)\}^2] \mathrm{ds})^{1/2}$. For $\hat{\mu}_y(t)$ and $\hat{\beta}(s,t)$ the definitions were analogous, with double integrals for the latter. For the principal component estimators, which have undefined signs, we actually computed the bias and root mean squared errors of the bivariate functions $\hat{\phi}_j(s)\hat{\phi}_j(s')$ and $\hat{\psi}_j(t)\hat{\psi}_j(t')$, which are sign-invariant. These are reported in Table 1; for $\hat{\mu}_x$ and $\hat{\mu}_y$ the quantities have been multiplied by ten to eliminate leading zeros. For reasons of space we only report here the results for Models 3, 4 and 6; the full report is given in the Supplementary Material.

Table 1. *Simulation Results. Biases and root mean squared errors*

	$n = 50$											
	Model 3				Model 4				Model 6			
	bias		rmse		bias		rmse		bias		rmse	
Parameter	W	O	W	O	W	O	W	O	W	O	W	O
β	0.37	1.00	1.15	1.14	0.47	1.23	1.39	1.32	0.80	1.05	1.56	1.11
μ_x	0.14	0.27	0.46	0.47	0.13	0.26	0.47	0.46	0.55	0.94	0.93	1.11
μ_y	0.16	0.38	0.56	0.58	0.19	0.65	0.61	0.81	0.52	0.87	0.92	1.05
ϕ_1	0.92	0.99	1.23	1.40	0.96	0.99	1.36	1.40	0.98	0.99	1.39	1.40
ϕ_2	0.25	0.93	0.59	1.06	0.22	0.96	0.58	1.07	0.86	1.08	1.18	1.25
ψ_1	0.99	0.99	1.40	1.40	0.99	0.99	1.40	1.39	0.99	0.99	1.40	1.40
ψ_2	0.17	0.87	0.47	1.21	0.20	0.62	0.48	1.03	0.53	1.01	0.87	1.20
	$n = 100$											
β	0.38	1.06	0.83	1.13	0.41	1.26	0.88	1.31	0.85	1.05	1.25	1.08
μ_x	0.13	0.27	0.34	0.38	0.11	0.27	0.34	0.38	0.53	0.95	0.74	1.03
μ_y	0.16	0.38	0.40	0.49	0.18	0.66	0.45	0.75	0.50	0.88	0.74	0.97
ϕ_1	0.55	0.99	0.79	1.40	0.48	0.99	0.70	1.40	0.99	0.99	1.40	1.40
ϕ_2	0.22	1.04	0.46	1.09	0.15	1.04	0.40	1.09	0.92	1.18	1.13	1.27
ψ_1	0.84	0.98	1.19	1.39	0.81	0.99	1.15	1.40	0.97	0.99	1.38	1.40
ψ_2	0.12	0.92	0.33	1.13	0.16	0.63	0.34	1.00	0.47	1.14	0.70	1.23

W, warped functional regression; O, ordinary functional regression.

Table 1 shows that warped functional regression estimators have smaller biases than ordinary functional regression estimators in practically all cases, which is not surprising since the model has more parameters; for the same reason they have higher variances. The questions is whether the smaller bias outweighs the higher variance. Root mean squared errors show that this is indeed the case: warped regression estimators beat ordinary least squares estimators in practically all cases. The one exception is Model 6, where the higher variability of $\hat{\beta}$ outweighs its lower

bias, at least for $n \leq 100$, but even in this case, the root mean squared errors of the other functional parameters is still smaller for the warped estimator. Therefore, from the point of view of estimation accuracy, the warped functional regression estimator is advantageous in presence of phase variability.

4·2. *Prediction accuracy*

Another aspect of the regression problem is prediction, or the estimation of a response function $y(t)$ for a new covariate curve $x(s)$. We compared prediction accuracy of warped and ordinary regression estimators by simulating data from Models 1–4 of § 4·1; for Models 5 and 6 prediction did not make sense because covariate and response warping functions were independent. In addition to training samples of sizes $n = 50$ and $n = 100$, we generated prediction samples of size $n^* = 100$ on equally-spaced time grids of size $\nu = 20$ and measured the prediction accuracy by the root mean squared error $\{E(\sum_{i=1}^{n^*} \|y_i - \hat{y}_i\|^2 / \nu n^*)\}^{1/2}$. For each model we computed the same estimators as in § 4·1 and in addition ordinary linear regression estimators with more principal components. Specifically, for the one-dimensional Models 1 and 2 we considered ordinary least squares estimators with 1, 2 and 3 components, and for the two-dimensional Models 3 and 4 we considered estimators with 2, 3 and 4 components.

Table 2 shows the results. The table indicates the overall dimension of the estimators: for example, O-9 is the ordinary regression estimator based on three principal components for covariates and responses, which has overall dimension 9. Prediction errors of ordinary linear regression estimators will decrease as the number of principal components increases, and eventually they will be smaller than prediction errors of warped regression estimators of fixed dimension. The point is that given comparable prediction errors, a low-dimensional warped regression model that neatly separates the two sources of variability will be preferable to a higher-dimensional ordinary linear model that confounds them.

Table 2. *Simulation Results. Prediction errors for new responses*

	Model 1		Model 2	
Estimator	$n = 50$	$n = 100$	$n = 50$	$n = 100$
W-1	0.14	0.13	0.15	0.14
O-1	0.19	0.19	0.20	0.20
O-4	0.14	0.13	0.15	0.15
O-9	0.14	0.13	0.15	0.15
	Model 3		Model 4	
W-4	0.20	0.19	0.21	0.20
O-4	0.21	0.20	0.23	0.23
O-9	0.17	0.17	0.20	0.19
O-16	0.17	0.16	0.19	0.18

W, warped functional regression; O, ordinary functional regression.

Generally speaking, the ordinary linear regression estimator needs an additional principal component to attain a prediction error comparable to or smaller than the warped regression estimator, although sometimes a strictly smaller prediction error is not attained, as in Models 1 and 2. For Models 3 and 4 the ordinary least squares estimator does attain smaller prediction errors, but in order to attain an error that is only 10% smaller it needs to use four times as many parameters as the warped regression model, which makes it extremely impractical from the point of view of interpretability. Interpretability issues cannot be directly gleaned from Table 2 or other simulation summaries because they are graphical in nature, so we study them by example in § 5.

5. APPLICATION: MODELING GROUND-LEVEL OZONE CONCENTRATION

Ground-level ozone causes serious health problems. Unlike other pollutants, ozone is not emitted directly into the air but is a result of complex chemical reactions in the atmosphere that include, among other factors, volatile organic compounds and oxides of nitrogen. Oxides of nitrogen are emitted by combustion engines, power plants and other industrial sources. The modeling of ground-level ozone formation has been an active topic of air-quality studies for many years.

In this article we will use data from the California Environmental Protection Agency online database. Hourly concentration of pollutants at many locations in California are available for the years 1980–2009. We will analyze trajectories of oxides of nitrogen, NOx, and ozone, O3, in the city of Sacramento in the Summer of 2005. We omit weekends and holidays because NOx and O3 levels are substantially lower and follow different patterns. We also removed some outlying trajectories, so the final sample consisted of 52 days between June 6 and August 26, shown in Figure 1.

Both NOx and O3 trajectories follow simple regular patterns. NOx curves tend to peak around 7am, and O3 curves around 2pm. Therefore we fitted warped regression models with single warping knots, trying several values of τ_{x0} and τ_{y0} around 7am and 2pm respectively. The results were similar in all cases; the estimators reported here correspond to $\tau_{x0} = 7$ and $\tau_{y0} = 14$. As basis functions we used cubic B-splines with 7 equally spaced knots, one knot every 3 hours; we also tried 10 knots but the results were not substantially different. Three warped regression models were fitted: (i) a model with one principal component for x and one for y, (ii) a model with two principal components for x and one for y, and (iii) a model with one principal component for x and two for y. The log-likelihood values were 44.44, 45.21 and 52.04, respectively; and the cross-validation prediction errors were 0.139, 0.136 and 0.134, respectively. Model (ii) has the same overall dimension as model (iii) but is worse both from the point of view of the

likelihood and of the prediction error, so we discarded it. For model (i) the estimated regression coefficients and the bootstrap standard deviations, based on 200 resamples, were

$$\hat{A} = \begin{pmatrix} 0.73 & 0.09 \\ 0.19 & 0.44 \end{pmatrix}, \quad \mathrm{std}(\hat{A}) = \begin{pmatrix} 0.07 & 0.02 \\ 0.08 & 0.06 \end{pmatrix},$$

and for model (iii) were

$$\hat{A} = \begin{pmatrix} 0.36 & 0.12 \\ 0.01 & 0.02 \\ 0.18 & 0.54 \end{pmatrix}, \quad \mathrm{std}(\hat{A}) = \begin{pmatrix} 0.08 & 0.06 \\ 0.04 & 0.10 \\ 0.06 & 0.11 \end{pmatrix}.$$

For model (iii) the coefficients of the second principal component of the response, $\hat{a}_{21}$ and $\hat{a}_{22}$, are not significant, while for model (i) all coefficients are significant even allowing for underestimation of the standard deviations, with the possible exception of $\hat{a}_{21}$ which is a borderline case. For this reason we prefer (i) as our final model.

To interpret the principal components, Figure 2(a) shows $\hat{\mu}_x$ and $\hat{\mu}_x \pm c_1\hat{\phi}_1$ for some constant c_1, and Figure 2(b) shows $\hat{\mu}_y$ and $\hat{\mu}_y \pm c_2\hat{\psi}_1$ for another constant c_2. Both principal components are shape components: curves with positive scores tend to have sharper features than the mean while curves with negative scores tend to have flatter features than the mean. The fact that the diagonal coefficients of $\hat{A}$ are positive indicates that the component scores $\hat{u}_i$ and $\hat{v}_i$ are positively correlated, as Figure 2(c) shows, and the warping landmarks $\hat{\tau}_{xi}$ and $\hat{\tau}_{yi}$, which can roughly be interpreted as peak locations, are also positively correlated, as Figure 2(f) shows. Amplitude and warping factors are also positively cross-correlated, since the off-diagonal elements of $\hat{A}$ are also positive. In particular $\hat{a}_{12}$ is highly significant, so late NOx peaks tend to be associated with high peaks of O3 and vice-versa, as Figure 2(d) shows.

An ordinary functional regression fit is shown in Figure 3; the plot shows $\hat{\mu}_x$, $\hat{\mu}_y$, $\hat{\mu}_x \pm c_1\hat{\phi}_j$ and $\hat{\mu}_y \pm c_2\hat{\psi}_j$ for a three-component model, or overall dimension nine, which has cross-

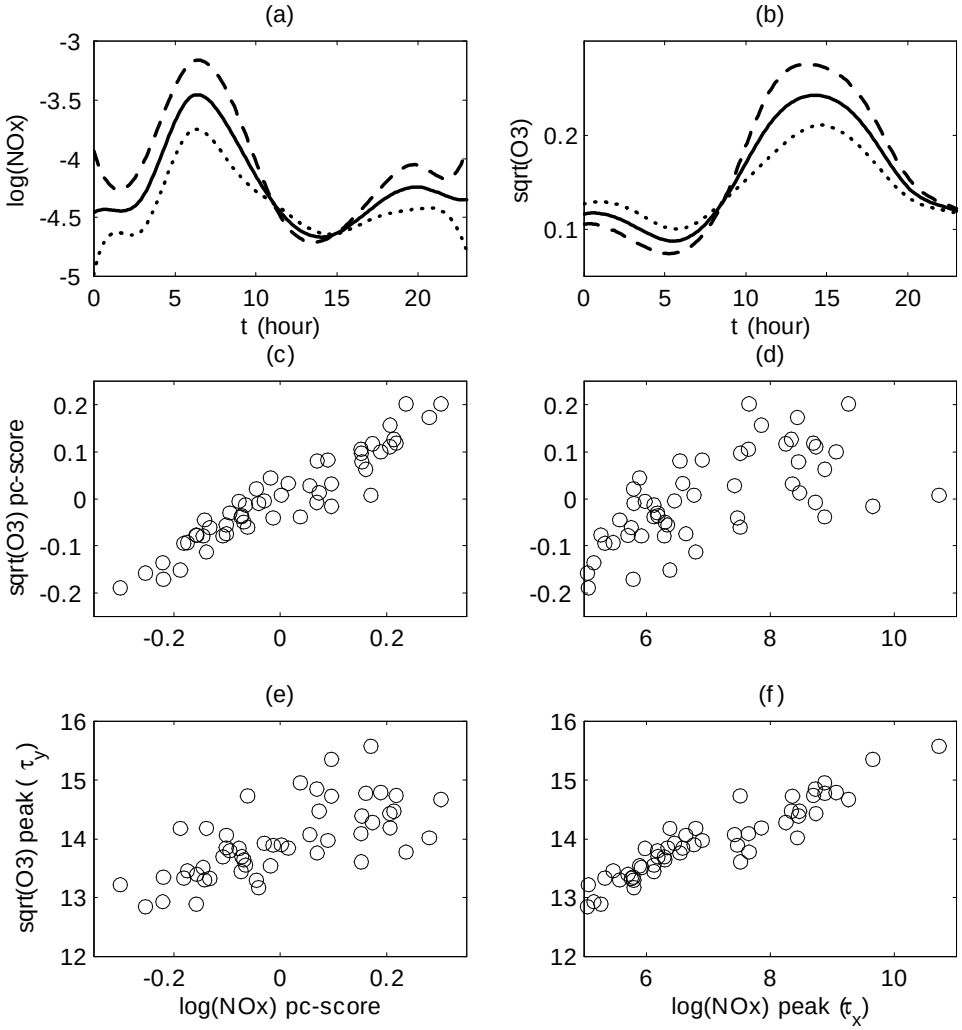

Fig. 2. Ozone Example. Warped Functional Regression fit. (a) Log-NOx mean (solid line), and mean plus (dashed line) and minus (dotted line) the principal component; (b) same as (a) for the square root of O3; (c) covariate versus response pc-scores; (d) covariate peak versus response pc-score; (e) covariate pc-score versus response peak; (f) covariate versus response peaks.

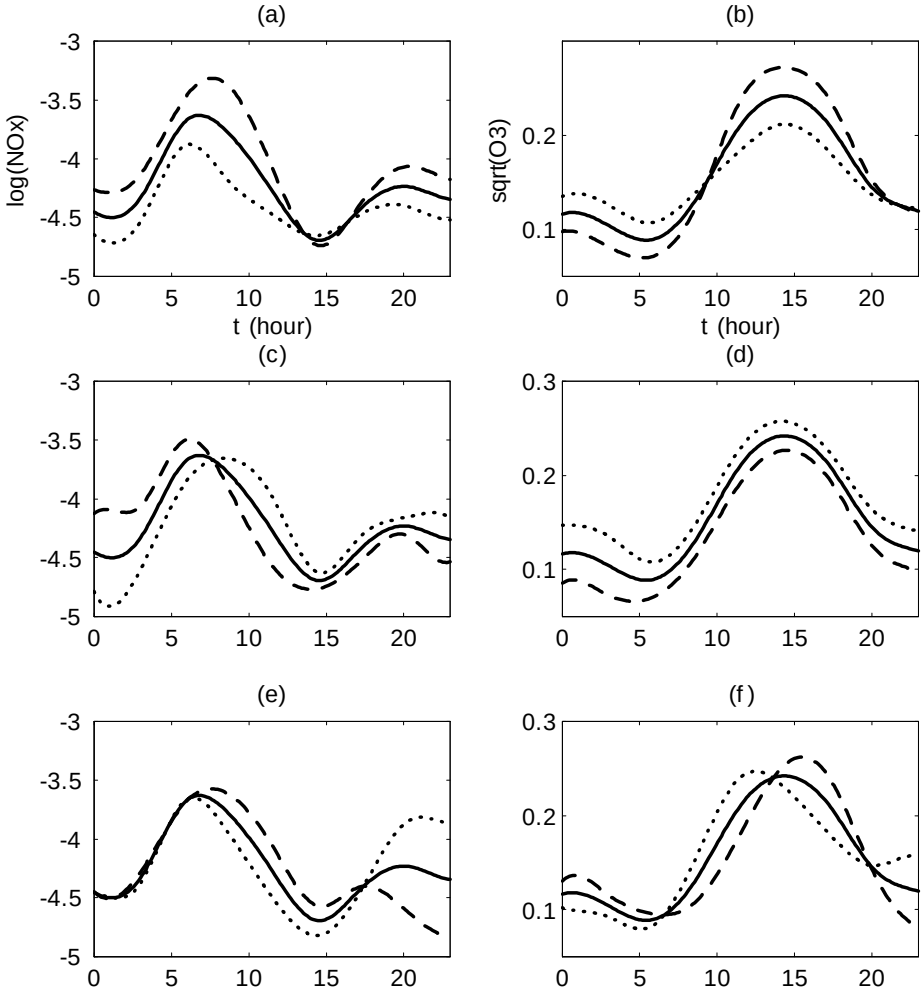

Fig. 3. Ozone Example. Ordinary Functional Regression fit. (a,c,e) Mean (solid line), and mean plus (dashed line) and minus (dotted line) the first [(a)], second [(c)] and third [(e)] principal components of explanatory curves; (b,d,f) same as (a,c,e), respectively, for response curves.

validation prediction error 0.126. A two-component model, of overall dimension four and thus comparable to the warped regression model, would correspond to the upper four panels of Figure 3, and has a cross-validation prediction error 0.132. Time variability in the explanatory

curves is explained by the second x-component (Figure 3(c)), but phase variability in the response curves is not accounted for until the third component (Figure 3(f)), so it really takes a nine-dimensional ordinary regression model to explain the phase-variability features that a four-dimensional warped model would explain. From a quantitative point of view the nine-dimensional ordinary regression model offers a lower cross-validation prediction error, 0.126, than the four-dimensional warped model, 0.139; but this modest decrease of 10% in the prediction error comes at the price of a higher model dimension and a less clear-cut interpretation of the components. We see in Figure 3(c,f) that the predominantly time-related principal components are also associated with some kinds of amplitude variability, and we see in Figure 3(a) that the predominantly amplitude-related first x-component is somewhat influenced by time variability. The question 'Is the timing of the NOx peak a strong predictor of the timing of the O3 peak?' cannot be answered as easily with ordinary functional regression as with warped regression. This blurring of the principal components is avoided by warped functional regression, which neatly separates the sources of variability and offers a more easily interpretable model. And in cases where phase variability is even more pronounced, warped regression does attain a lower cross-validation error than ordinary functional regression, as shown in the additional example in the Supplementary Material.

ACKNOWLEDGEMENT

This research was partially supported by a grant from the National Science Foundation.

SUPPLEMENTARY MATERIAL

Supplementary material available online includes further discussion of model identifiability, the derivation of the EM algorithm for estimation, detailed derivation of formulae involved in the

asymptotic distribution of the estimator, the proof of Theorem 1, additional simulation results, an additional data analysis, and a detailed treatment of monotone Hermite splines.

APPENDIX

5·1. *Monotone Hermite splines*

In this section we explain how the warping functions $\omega_i(s)$ are constructed; the $\zeta_i(t)$s are constructed in a similar way. Let $\mathscr{S} = [a, b]$ and $a < \tau_{01} < \cdots < \tau_{0r} < b$ be a sequence of r knots in $\mathscr{S}$. For $h_{00}(s) = (1 + 2s)(1 - s)^2$, define the basis functions $\{\alpha_j(s; \tau_0)\}$ as

$$\alpha_0(s;\tau_0) = \begin{cases} 0, & s < a \text{ or } s > \tau_{01} \\ h_{00}\left(\frac{s-a}{\tau_{01}-a}\right), & a \le s \le \tau_{01}, \end{cases}$$

$$\alpha_j(s;\tau_0) = \begin{cases} 0, & s < \tau_{0,j-1} \text{ or } s > \tau_{0,j+1} \\ h_{00}\left(\frac{\tau_{0j}-s}{\tau_{0j}-\tau_{0,j-1}}\right), & \tau_{0,j-1} \le s \le \tau_{0j} \\ h_{00}\left(\frac{s-\tau_{0j}}{\tau_{0,j+1}-\tau_{0j}}\right), & \tau_{0j} \le s \le \tau_{0,j+1} \end{cases}$$

for $j = 1, \ldots, r$, and

$$\alpha_{r+1}(s;\tau_0) = \begin{cases} 0, & s < \tau_{0r} \text{ or } s > b \\ h_{00}\left(\frac{b-s}{b-\tau_{0r}}\right), & \tau_{0r} \le s \le b. \end{cases}$$

For $h_{10}(s) = s(1 - s)^2$, define the basis functions $\{\beta_j(s; \tau_0)\}$ as

$$\beta_0(s;\tau_0) = \begin{cases} 0, & s < a \text{ or } s > \tau_{01} \\ (\tau_{01}-a)h_{10}\left(\frac{s-a}{\tau_{01}-a}\right), & a \le s \le \tau_{01}, \end{cases}$$

$$\beta_j(s;\tau_0) = \begin{cases} 0, & s < \tau_{0,j-1} \text{ or } s > \tau_{0,j+1} \\ -(\tau_{0j}-\tau_{0,j-1})h_{10}\left(\frac{\tau_{0j}-s}{\tau_{0j}-\tau_{0,j-1}}\right), & \tau_{0,j-1} \le s \le \tau_{0,j} \\ (\tau_{0,j+1}-\tau_{0,j})h_{10}\left(\frac{s-\tau_{0,j}}{\tau_{0,j+1}-\tau_{0,j}}\right), & \tau_{0,j} \le s \le \tau_{0,j+1} \end{cases}$$

for $j = 1, \ldots, r$, and

$$\beta_{r+1}(s;\tau_0) = \begin{cases} 0, & s < \tau_{0r} \text{ or } s > b \\ -(b-\tau_{0r}) h_{10}\left(\frac{b-s}{b-\tau_{0r}}\right), & \tau_{0r} \le s \le b. \end{cases}$$

The function $\omega_i(s) = \sum_{j=0}^{r+1} \tau_{ij}\alpha_j(s;\tau_0) + \sum_{j=0}^{r+1} d_{ij}\beta_j(s;\tau_0)$, where $\tau_{i0} = a$ and $\tau_{i,r+1} = b$, is a differentiable piecewise-cubic function that satisfies $\omega_i(\tau_{0j}) = \tau_{ij}$ and $\omega_i'(\tau_{0j}) = d_{ij}$ for $j = 1, \ldots, r$. Thus the τ_{ij}s play the dual role of basis coefficients and values of $\omega_i(s)$ at the knots. For $\omega_i(s)$ to be strictly monotone increasing the d_{ij}s must satisfy certain necessary and sufficient conditions given in Fritsch & Carlson (1980). For situations like ours where no particular values of the d_{ij}s are specified, Fritsch & Carlson provide a simple algorithm to compute, from given τ_{ij}s, values of the d_{ij}s that satisfy the monotonicity constraints. This algorithm is given in the Supplementary Material. Since the algorithm is deterministic, the d_{ij}s are functions of the τ_{ij}s and then $\omega_i(s)$ is entirely parameterized by $\tau_i = (\tau_{i1}, \ldots, \tau_{ir})$, thus forming an r-dimensional space.

The Jupp transform (Jupp, 1978) is defined as $\theta_{ij} = \log\{(\tau_{i,j+1} - \tau_{ij})/(\tau_{ij} - \tau_{i,j-1})\}$ for $j = 1, \ldots, r$, with inverse given by $\tau_{ij} = a + (b-a)s_{ij}/(1+s_{ir})$ for $j = 1, \ldots, r$, where $s_{ij} = \sum_{k=1}^{j} \exp(\theta_{i1} + \cdots + \theta_{ik})$. For any r-dimensional unconstrained vector θ the inverse Jupp transform yields a vector τ of strictly increasing knots in (a, b). In particular, for $\theta = 0$ the corresponding τ is a sequence of r equally spaced knots in (a, b).

REFERENCES

ASH, R. B. & GARDNER, M. F. (1975). *Topics in Stochastic Processes*. New York: Academic Press.

BRUMBACK, L. C. & LINDSTROM, M. J. (2004). Self modeling with flexible, random time transformations. *Biometrics* **60** 461–470.

CAI, T. & HALL, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34** 2159–2179.

CRAMBES, C., KNEIP, A., & SARDA, P. (2009). Smoothing splines estimators for functional linear regression. *Ann. Statist.* **37** 35–72.

FRITSCH, F. N. & CARLSON, R. E. (1980). Monotone piecewise cubic interpolation. *SIAM J. Numer. Anal.* **17** 238–246.

GERVINI, D. & GASSER, T. (2004). Self-modeling warping functions. *J. R. Statist. Soc. B* **66** 959–971.

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

GEYER, C. J. (1994). On the asymptotics of constrained M-estimators. *Ann. Statist.* **22** 1993–2010.

HALL, P. & HOROWITZ, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35** 70–91.

JAMES, G. M. (2007). Curve alignment by moments. *Ann. Appl. Statist.* **1** 480–501.

JAMES, G. M., HASTIE, T. J. & SUGAR, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **87** 587–602.

JAMES, G., WANG, J. & ZHU, J. (2009). Functional linear regression that's interpretable. *Ann. Statist.* **37** 2083–2108.

JUPP, D. L. B. (1978). Approximation to data by splines with free knots. *SIAM J. Numer. Anal.* **15** 328–343.

KNEIP, A., LI, X., MACGIBBON, B. & RAMSAY, J. O. (2000). Curve registration by local regression. *Canadian J. Statist.* **28** 19–30.

KNEIP, A. & RAMSAY, J. O. (2008). Combining registration and fitting for functional models. *J. Amer. Statist. Assoc.* **103** 1155–1165.

LIU, X. & MULLER, H.-G. (2004). Functional convex averaging and synchronization for time-warped random curves. *J. Amer. Statist. Assoc.* **99** 687–699.

MAGNUS, J. R. & NEUDECKER, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics (Second Edition)*. New York: Wiley.

MULLER, H.-G., CHIOU, J.-M., & LENG, X. (2008). Inferring gene expression dynamics via functional regression analysis. *BMC Bioinformatics* **9** 60.

POLLARD, D. (1990). *Empirical Processes: Theory and Applications*. Hayward, California: Institute of Mathematical Statistics.

RAMSAY, J. O. (1988). Monotone regression splines in action (with discussion). *Statist. Science* **3** 425–461.

RAMSAY, J. O. & LI, X. (1998). Curve registration. *J. R. Statist. Soc. B* **60** 351–363.

RAMSAY, J. O. & SILVERMAN, B. (2005). *Functional Data Analysis (Second Edition).* Springer, New York.

ROCKAFELLAR, R. & WETS, R. (1998). *Variational Analysis*. New York: Springer.

TANG, R. & MULLER, H.-G. (2008). Pairwise curve synchronization for functional data. *Biometrika* **95** 875–889.

TANG, R. & MULLER, H.-G. (2009). Time-synchronized clustering of gene expression trajectories. *Biostatistics* **10** 32–45.

YAO, F., MULLER, H.-G. & WANG, J.-L. (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33** 2873–2903.

WANG, K. & GASSER, T. (1999). Synchronizing sample curves nonparametrically. *Ann. Statist.* **27** 439–460.