Warped Functional Regression

Daniel Gervini
Department of Mathematical Sciences,
University of Wisconsin–Milwaukee,
PO Box 413, Milwaukee, Wisconsin 53201, USA
gervini@uwm.edu

May 22, 2014
Abstract

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.

Key Words: Functional Data Analysis; Random-Effect Models; 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 in statistics (Ramsay & Silverman, 2005). Many problems in this area involve modeling curves as functions of other curves. For example, Figure 1(a) shows daily trajectories 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 this type of 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 it 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 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 again is 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 two sources of variability, phase and amplitude, are confounded in the model.

The curves shown 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, and vice-versa. 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 questions. 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 \((x_1, y_1), \ldots, (x_n, y_n)\), where \(x_i(s)\) is the covariate and \(y_i(t)\) the response, with \(x_i : \mathcal{I} \to \mathbb{R}\) and \(y_i : \mathcal{T} \to \mathbb{R}\), and \(\mathcal{I}\) and \(\mathcal{T}\) 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

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

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

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 \( \omega_i(s) \) and the \( \zeta_i(t) \) are monotone increasing warping functions with \( \omega_i : \mathcal{S} \rightarrow \mathcal{S} \) and \( \zeta_i : \mathcal{T} \rightarrow \mathcal{T} \). 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 \( L^2(\mathcal{S}) \) and \( L^2(\mathcal{T}) \), 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 \( x_i(s) \) 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 $\xi_j$s are non-random basis functions, the $C_{ij}$s are random coefficients, and the $v_i$s are warping functions. Orthogonalizing the $\xi_j$s one obtains 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 $\mathcal{W}$ 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 are possible, such as integrated splines (Ramsay, 1988), monotone splines (Ramsay & Li, 1998) and constrained B-splines (Brumback & Lindstrom, 2004), monotone Hermite splines are better suited for the regression approach proposed here. Details about this family of warping functions are given in Appendix 5.1. We only mention here that, like other spline families, this is a finite-dimensional semiparametric family determined by a knot sequence chosen by the user. Thus, the family $\{\omega_i(s)\}$ will be determined by a knot sequence $\tau_{x0} = (\tau_{x01}, \ldots, \tau_{x0r_1})$ of strictly increasing points in $\mathcal{S}$, and each $\omega_i(s)$ will be determined by a corresponding sequence $\tau_{xi}$ of basis coefficients which satisfy $\omega_i(\tau_{x0j}) = \tau_{xij}$ for $j = 1, \ldots, r_1$. Similarly, the family $\{\zeta_i(t)\}$ will be determined by a knot sequence $\tau_{y0} = (\tau_{y01}, \ldots, \tau_{y0r_2})$ of strictly increasing points in $\mathcal{T}$ and each $\zeta_i(t)$ will be determined by basis coefficients $\tau_{yi}$ which satisfy $\zeta_i(\tau_{y0j}) = \tau_{yij}$ for $j = 1, \ldots, r_2$. The dual role of the $\tau_{xi}$s and the $\tau_{yi}$s as basis coefficients and as values of $\omega_i(s)$ and $\zeta_i(t)$ at the knots is what makes Hermite splines appealing. It is natural then to choose the knot sequences $\tau_{x0}$ and $\tau_{y0}$ to roughly correspond to the average location of the main features of the $x_i$s and the $y_i$s. Like $p_1$ and $p_2$ in (5) and (6), the dimensions $r_1$ and $r_2$ need not be large, since they will roughly correspond to the number of peaks and valleys of the $x_i$s and the $y_i$s, which will not be large for the type of applications we envision.

Unlike landmark registration, where the $\tau_{x_i}$s and the $\tau_{y_i}$s are individually estimated curve by curve, we will treat the $\tau_{x_i}$s and the $\tau_{y_i}$s as latent random effects, so they will
not be estimated directly. This is a big advantage in practice, since individual estimation
of the $\tau_{xi}$s and the $\tau_{yi}$s is difficult when the number of curves is large or when the curves
are sparsely sampled. A minor complication is that the $\tau_{xi}$s and the $\tau_{yi}$s are constrained
to be monotone increasing in $\mathcal{I}$ and $\mathcal{J}$, respectively, so for convenience we will work
with their Jupp transforms $\theta_{xi}$ and $\theta_{yi}$ instead, which are unconstrained vectors; the Jupp
transform is defined in Appendix 5.1.

Since the warping functions $\{\omega_{i}\}$ and $\{\zeta_{i}\}$ are determined by the random effects $\theta_{xi}$
and $\theta_{yi}$, 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 $y_{i}$s on the $x_{i}$s via the random
effects:

$$
\begin{bmatrix}
    v_{i} \\
    \theta_{yi}
\end{bmatrix} = 
\begin{bmatrix}
    0 \\
    \theta_{y0}
\end{bmatrix} + A \left( \begin{bmatrix}
    u_{i} \\
    \theta_{xi}
\end{bmatrix} - \begin{bmatrix}
    0 \\
    \theta_{x0}
\end{bmatrix} \right) + e_{i},
$$

(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 $\Sigma_{e}$ diagonal. For interpretability we split $A$ into four blocks

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

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

$$y_{i}^{*}(t) - \mu_{y}(t) = \int \beta(s, t) \{x_{i}^{*}(s) - \mu_{x}(s)\} ds + \gamma_{1}(t)^{T}(\theta_{xi} - \theta_{x0}) + \delta_{i}(t), \quad (8)$$

and

$$\theta_{yi} - \theta_{y0} = \int \gamma_{2}(s) \{x_{i}^{*}(s) - \mu_{x}(s)\} ds + A_{22}(\theta_{xi} - \theta_{x0}) + e_{i2}, \quad (9)$$

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 $\{\psi_l(t)\}$. We will do that via B-splines. Let $b_x(s) = (b_{x1}(s), \ldots, b_{xq_1}(s))^T$ be a B-spline basis in $L^2(\mathcal{S})$ and $b_y(t) = (b_{y1}(t), \ldots, b_{yq_2}(t))^T$ a B-spline basis in $L^2(\mathcal{T})$. Let $\mu_x(s) = b^T_x(s)m_x$, $\mu_y(t) = b^T_y(t)m_y$, $\phi_k(s) = b^T_x(s)c_k$ and $\psi_l(t) = b^T_y(t)d_l$, for $m_x \in \mathbb{R}^{q_1}$, $m_y \in \mathbb{R}^{q_2}$, $c_k \in \mathbb{R}^{n_1}$ and $d_l \in \mathbb{R}^{q_2}$. The orthogonality restrictions on the $\phi_k$s and the $\psi_l$s can be expressed as $C^T J_x C = I_{p_1}$ and $D^T J_y D = I_{p_2}$, where $C = [c_1, \ldots, c_{p_1}] \in \mathbb{R}^{n_1 \times p_1}$, $D = [d_1, \ldots, d_{q_2}] \in \mathbb{R}^{q_2 \times p_2}$, $J_x = \int b_x(s)b_x^T(s)ds$ and $J_y = \int b_y(t)b_y^T(t)dt$.

If the curves $\{x_i\}$ and $\{y_i\}$ were observed on dense time grids and individual smoothing was 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. Then we will treat $u_i$, $v_i$, $\theta_{xi}$ and $\theta_{yi}$ as latent variables and estimate the model parameters by maximum likelihood. We assume $w_i = (u_{i1}^T, \theta_{x1}^T)^T$ is jointly multivariate Normal of dimension $d_1 = p_1 + r_1$, with mean and covariance given by

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

where $\theta_{x0}$ the Jupp transform of the knot vector $\tau_{x0}$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_{p_1})$. 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 and covariance given by

$$
\mu_z = \begin{bmatrix} 0 \\ \theta_{y0} \end{bmatrix}, \quad \Sigma_z = A \Sigma_w A^T + \Sigma_e,
$$

where $\theta_{y0}$ is the Jupp transform of the knot vector $\tau_{y0}$. Thus $v_i \sim N(0, \Gamma)$ with $\Gamma = A_1 \Sigma_w A_1^T + \Sigma_{e11}$, where $A_1 = [A_{11}, A_{12}]$ and $\Sigma_{e11}$ the $p_2 \times p_2$ upper-left diagonal block of $\Sigma_e$. Since $\Gamma$ has to be diagonal by model (6), and $\Sigma_e$ was assumed diagonal, it follows that $A_1 \Sigma_w A_1^T$ 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 $\Sigma_e$, the covariance matrix $\Sigma_w$ of the explanatory random effects $w_i$, the
spline coefficients $m_x$, $m_y$, $C$ and $D$ of the functional parameters, and the variances $\sigma_x^2$ and $\sigma_y^2$ of the random noise in (1) and (2). The derivation of the likelihood function and the EM algorithm to compute these estimators are discussed in Appendix 5.2 and 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 number 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 $x_i$s and the $y_i$s.

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 $\hat{E}(v_{n+1}|x_{n+1})$ and $\hat{E}(\theta_{y,n+1}|x_{n+1})$, which under model (7) come down to $\hat{v}_{n+1} = \hat{A}_{11}\hat{E}(u_{n+1}|x_{n+1}) + \hat{A}_{12}\{\hat{E}(\theta_{x,n+1}|x_{n+1}) - \theta_{x0}\}$ and $\hat{\theta}_{y,n+1} = \hat{A}_{21}\hat{E}(u_{n+1}|x_{n+1}) + \hat{A}_{22}\{\hat{E}(\theta_{x,n+1}|x_{n+1}) - \theta_{x0}\}$. 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 $\hat{y}_{n+1}(t) = \hat{y}_{n+1}^*\{\hat{\zeta}_{n+1}(t)\}$.

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. For this reason we will 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
\( \Sigma_e, \mathbf{m}_x, \mathbf{m}_y, C, D, \sigma^2_e \) and \( \sigma^2_\eta \) are fixed and known, because this assumption does not alter the asymptotic covariance matrix of \( \hat{A} \). However, in principle we cannot assume that \( \Sigma_w \) is fixed and known because \( \Sigma_w \) is part of the condition that \( A_1 \Sigma_w A_1^T \) 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,

\[
\zeta = \begin{bmatrix} \text{vec}(A^T) \\ v(\Sigma_w) \end{bmatrix},
\]

(10)

where \( v(\Sigma_w) \) denotes the vec of the lower-triangular part of \( \Sigma_w \), including the diagonal. The dimension of \( \zeta \) is then \( d = d_1 d_2 + d_1 (d_1 + 1)/2 \). The restriction that \( A_1 \Sigma_w A_1^T \) be diagonal can be expressed as a system of \( m = (p_2 - 1)p_2/2 \) constraints of the form \( h_{ij}(\zeta) = 0 \), where \( h_{ij}(\zeta) = a_i^T \Sigma_w a_j \) and \( a_i^T \) is the \( i \)th 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 \( C = \{ \zeta \in \mathbb{R}^d : h(\zeta) = 0 \} \). The additional condition that \( \Sigma_w \) be positive definite does not alter the asymptotic distribution of the estimator because \( \Sigma_w \) lies in the interior of this space, not on the border. Let \( \zeta_0 \) be the true value of the parameter \( \zeta \). Since \( h(\zeta) \) is continuously differentiable, the tangent cone of \( C \) at \( \zeta_0 \) is \( \mathcal{T}_C(\zeta_0) = \{ \delta \in \mathbb{R}^d : D h(\zeta_0) \delta = 0 \} \), where \( D \) is the differential (Rockafellar & Wets, 1998, ch. 6.B). The asymptotic distribution of the constrained estimator \( \hat{\zeta}_n \) is simple in this case: it is just the usual asymptotic Normal distribution of an unconstrained maximum likelihood estimator, projected on \( \mathcal{T}_C(\zeta_0) \).

Specifically, let

\[
M(x, y) = E\{(w - \mu_w)(w - \mu_w)^T|(x, y)\},
\]

(11)

\[
N(x, y) = E\{(w - \mu_w)(z - \mu_z)^T|(x, y)\},
\]

(12)

and

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

(13)

where \( D_{d_1} \) is the duplication matrix that satisfies \( \text{vec}(\Sigma_w) = D_{d_1} v(\Sigma_w) \) (Magnus & Neudecker, 1999, ch. 3). It is shown in the Supplementary Material that \( U(x, y) \) is the
likelihood score function $\nabla_\zeta \log f(x, y; \zeta)$ at $\zeta = \zeta_0$. Let $B = Dh(\zeta_0)$, which is an $m \times d$ matrix of rank $m$ with rows

$$\nabla h_{ij}(\zeta)^T = [a_i^T \Sigma_w(e_j \otimes I_d) + a_j^T \Sigma_w(e_i \otimes I_d), 0_{rd_1}, (a_j^T \otimes a_i^T) D_{d_1}],$$

where $e_i$ is the $i$th canonical vector in $\mathbb{R}^p$. Let $\Xi$ 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 $I_d - B^T (B B^T)^{-1} B$; this matrix is not unique but Theorem 1 below is invariant under the choice of $\Xi$.

**Theorem 1**  
Under the above conditions, the asymptotic distribution of $\sqrt{n}(\hat{\zeta}_n - \zeta_0)$ 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

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

where the ‘hat’ in $U$ denotes that the true parameters in (13) are replaced by their estimators. The proof of Theorem 1 is given in the Appendix.

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 (13), identically distributed. In many applications, however, this will not be the case and the time grids will be unequal, giving $x_i \in \mathbb{R}^{\nu_1}$ and $y_i \in \mathbb{R}^{\nu_2}$ 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 (14) 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 exists, (11) and (12) and consequently (13) and (14) can still be computed with $(x_i, y_i)$s of unequal dimensions. The statement of Theorem 1 should then be re-expressed as

$$\sqrt{n}\{\Xi(\Xi^T \hat{V}_n \Xi)^{-1} \Xi^T\}^{-1/2}(\hat{\zeta}_n - \zeta_0) \longrightarrow N(0, I_d)$$

in distribution.
4 Simulations

4.1 Estimation accuracy

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

- Model 1: a one-dimensional amplitude and warping model, with $\mu_x(s) = .6\varphi(s, .3, .1) + .4\varphi(s, .6, .1)$, $\phi_1(s) = \varphi(s, .3, .1)/1.6796$, $\mu_y(t) = .6\varphi(t, .5, .1) + .4\varphi(t, .8, .1)$ and $\psi_1(t) = \varphi(t, .5, .1)/1.6796$, 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 $\tau_x = .3$ and $\tau_y = .5$. 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 $\Sigma_w = \text{diag}(.2^2, .1^2, .1^2, .1^2)$, $\Sigma_e = .07^2 I_2$, and $\sigma_\varepsilon = \sigma_\eta = .05$.

- Model 2: same as Model 1 but with a non-diagonal $A$; specifically, $a_{11} = a_{22} = 1$ and $a_{12} = a_{21} = .5$, so there was a relationship between covariate phase variability and response amplitude variability, and vice versa, in this model.

- Model 3: 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 $\varphi(s, .6, .1)$ orthogonalized with $\phi_1(s)$, and $\psi_1(t)$ the function $\varphi(t, .8, .1)$ orthogonalized with $\psi_1(t)$. The warping functions followed Hermite spline models with knots $\tau_{x0} = (.3, .6)$ and $\tau_{y0} = (.5, .8)$. 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 $\Sigma_w = \text{diag}(.2^2, .1^2, .1^2, .1^2)$, $\Sigma_e = .07^2 I_4$, and $\sigma_\varepsilon = \sigma_\eta = .05$.

- Model 4: same as Model 3 but with a non-diagonal regression matrix $A$, with blocks $A_{11} = A_{22} = I_2$ and $A_{12} = A_{21} = .5 I_2$.

- Model 5: 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 belonged 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 \( b(s)^T c_0 = s \), the identity, then we generated \( c_i \sim N(c_0, .05^2 I_9) \) 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: 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 \( \{s_{i1}, \ldots, s_{i\nu_1}\} \) and \( \{t_{i1}, \ldots, t_{i\nu_2}\} \) were random and irregular, with \( \nu_1 \) and \( \nu_2 \) 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 10 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 \( \tau_{x0} = .45 \) and \( \tau_{y0} = .65 \), 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 \( \tau_{x0} = (.3, .6) \) and \( \tau_{y0} = (.5, .8) \), and for Model 6 we used Hermite splines with knots at \( \tau_{x0} = .45 \) and \( \tau_{y0} = .65 \); 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 \( (\int [E\{\hat{\mu}_x(s)\} - \mu_x(s)]^2 ds)^{1/2} \) and as ‘root mean squared error’ the quantity \( (\int E[\{\hat{\mu}_x(s) - \mu_x(s)\}]^2 ds)^{1/2} \). 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 Tables 1 and 2; for \( \hat{\mu}_x \) and \( \hat{\mu}_y \) the quantities have been multiplied by 10 to eliminate leading zeros.

We see in Tables 1 and 2 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 are going to have higher variances. The question 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 exception is Model 6, where covariates and responses are warped independently and the warped regression estimator cannot fully show its advantages. However, even in this unfavorable case the root mean squared error of the warped regression estimator of \( \beta \) is not much higher than that of the ordinary least squares estimator, and for the other functional parameters it is actually smaller. 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 Section 4.1; for Models 5 and 6 prediction did not make much 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 \( (\sum_{i=1}^{n^*} \|y_i - \hat{y}_i\|^2 / \nu n^*)^{1/2} \). For each model we computed the same estimators as in Section 4.1, and in addition ordinary linear regression estimators with more principal
| Param. | Model 1 |     |     | Model 2 |     |     |
|--------|---------|-----|-----|---------|-----|-----|
|        | bias    | rmse| bias| rmse    | bias| rmse|
| β      | 0.12    | 0.19| 0.21| 0.30    | 0.11| 0.69|
| μₓ     | 0.10    | 0.19| 0.34| 0.37    | 0.12| 0.19|
| μᵧ     | 0.13    | 0.32| 0.42| 0.51    | 0.16| 0.59|
| φ₁     | 0.05    | 0.06| 0.15| 0.18    | 0.08| 0.05|
| ψ₁     | 0.15    | 0.21| 0.22| 0.34    | 0.09| 0.83|

| Model 3 | Model 4 |     |     |         |     |     |
|---------|---------|-----|-----|---------|-----|-----|
| β      | 0.37    | 1.00| 1.15| 1.14    | 0.47| 1.23|
| μₓ     | 0.14    | 0.27| 0.46| 0.47    | 0.13| 0.26|
| μᵧ     | 0.16    | 0.38| 0.56| 0.58    | 0.19| 0.65|
| φ₁     | 0.92    | 0.99| 1.23| 1.40    | 0.96| 0.99|
| φ₂     | 0.25    | 0.93| 0.59| 1.06    | 0.22| 0.96|
| ψ₁     | 0.99    | 0.99| 1.40| 1.40    | 0.99| 0.99|
| ψ₂     | 0.17    | 0.87| 0.47| 1.21    | 0.20| 0.62|

| Model 5 | Model 6 |     |     |         |     |     |
|---------|---------|-----|-----|---------|-----|-----|
| β      | 0.18    | 0.73| 0.73| 0.78    | 0.80| 1.05|
| μₓ     | 0.44    | 0.94| 0.84| 1.10    | 0.55| 0.94|
| μᵧ     | 0.49    | 0.86| 0.88| 1.03    | 0.52| 0.87|
| φ₁     | 0.18    | 0.68| 0.50| 0.86    | 0.98| 0.99|
| φ₂     | —       | —   | —   | —       | —   | —   |
| ψ₁     | 0.17    | 0.62| 0.47| 0.75    | 0.99| 0.99|
| ψ₂     | —       | —   | —   | —       | —   | —   |

Table 1: Simulation Results. Bias and root mean squared errors of warped functional regression (W) and ordinary functional regression (O) for sample size $n = 50$. 
| Param. | Model 1 |  |  |  |  | Model 2 |  |  |  |  |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
|        | bias   | rmse   | bias   | rmse   | bias   | rmse   | bias   | rmse   | bias   | rmse   |
| $\beta$ | 0.12   | 0.18   | 0.18   | 0.24   | 0.12   | 0.70   | 0.29   | 0.72   |        |        |
| $\mu_x$ | 0.10   | 0.19   | 0.27   | 0.31   | 0.13   | 0.19   | 0.29   | 0.30   |        |        |
| $\mu_y$ | 0.14   | 0.33   | 0.32   | 0.43   | 0.19   | 0.60   | 0.40   | 0.68   |        |        |
| $\phi_1$ | 0.05   | 0.05   | 0.11   | 0.13   | 0.07   | 0.05   | 0.19   | 0.12   |        |        |
| $\psi_1$ | 0.16   | 0.20   | 0.19   | 0.28   | 0.10   | 0.84   | 0.18   | 0.85   |        |        |
| $\beta$ | 0.38   | 1.06   | 0.83   | 1.13   | 0.41   | 1.26   | 0.88   | 1.31   |        |        |
| $\mu_x$ | 0.13   | 0.27   | 0.34   | 0.38   | 0.11   | 0.27   | 0.34   | 0.38   |        |        |
| $\mu_y$ | 0.16   | 0.38   | 0.40   | 0.49   | 0.18   | 0.66   | 0.45   | 0.75   |        |        |
| $\phi_1$ | 0.55   | 0.99   | 0.79   | 1.40   | 0.48   | 0.99   | 0.70   | 1.40   |        |        |
| $\phi_2$ | 0.22   | 1.04   | 0.46   | 1.09   | 0.15   | 1.04   | 0.40   | 1.09   |        |        |
| $\psi_1$ | 0.84   | 0.98   | 1.19   | 1.39   | 0.81   | 0.99   | 1.15   | 1.40   |        |        |
| $\psi_2$ | 0.12   | 0.92   | 0.33   | 1.13   | 0.16   | 0.63   | 0.34   | 1.00   |        |        |
| $\beta$ | 0.17   | 0.74   | 0.60   | 0.77   | 0.85   | 1.05   | 1.25   | 1.08   |        |        |
| $\mu_x$ | 0.43   | 0.95   | 0.69   | 1.04   | 0.53   | 0.95   | 0.74   | 1.03   |        |        |
| $\mu_y$ | 0.48   | 0.88   | 0.73   | 0.97   | 0.50   | 0.88   | 0.74   | 0.97   |        |        |
| $\phi_1$ | 0.15   | 0.76   | 0.42   | 0.87   | 0.99   | 0.99   | 1.40   | 1.40   |        |        |
| $\phi_2$ | —      | —      | —      | —      | 0.92   | 1.18   | 1.13   | 1.27   |        |        |
| $\psi_1$ | 0.16   | 0.66   | 0.40   | 0.72   | 0.97   | 0.99   | 1.38   | 1.40   |        |        |
| $\psi_2$ | —      | —      | —      | —      | 0.47   | 1.14   | 0.70   | 1.23   |        |        |

Table 2: Simulation Results. Bias and root mean squared errors of warped functional regression (W) and ordinary functional regression (O) for sample size $n = 100$. 
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 3 shows the results. The table indicates the overall dimension of the estimators: for example, O-9 is the ordinary regression estimator based on 3 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.

We see that, generally speaking, the ordinary linear regression estimator needs an additional principal component to attain a comparable or smaller prediction error 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 3 or other simulation summaries.
Random grids

|       | $n = 50$ |       | $n = 200$ |
|-------|----------|-------|-----------|
| $Q$   | 0.09     | $Q$   | 0.10      |
| $Z_{11}$ | 0.08    | $Z_{11}$ | 0.10  |
| $Z_{12}$ | 0.08    | $Z_{12}$ | 0.09  |
| True variance | 0.09 | Asymptotic | 0.10 |
| Asymptotic     | 0.34    | Asymptotic | 0.33  |
| Bootstrap       | 0.25    | Bootstrap | 0.25  |
|                | 0.16    |                | 0.14  |
|                | 0.13    |                | 0.11  |

Equally spaced grids

|       |       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|-------|
| $Q$   | 0.11  | $Q$   | 0.10  |
| $Z_{11}$ | 0.08    | $Z_{11}$ | 0.10  |
| $Z_{12}$ | 0.08    | $Z_{12}$ | 0.07  |
| True variance | 0.10 | Asymptotic | 0.27 |
| Asymptotic     | 0.36    | Asymptotic | 0.27  |
| Bootstrap       | 0.33    | Bootstrap | 0.29  |
|                | 0.18    |                | 0.11  |
|                | 0.21    |                | 0.23  |

Table 4: Simulation Results. Tail probabilities of test statistics, true value is 0.10.

because they are graphical in nature, so we are going to study them by example in §5.

4.3 Asymptotic accuracy

We also studied by simulation the finite-sample adequacy of the asymptotic results of §3, particularly for hypothesis testing. We simulated data from Model 1 with $A = 0$, and also from a similar model that uses equally-spaced time grids of size 15 instead of the random time grids of Model 1. Two sample sizes were considered in each case, $n = 50$ and $n = 200$. Each scenario was replicated 500 times.

The warped regression estimator was computed using the same specifications as above. The covariance matrix of $\text{vec}((\hat{A})^T)$ was estimated by the asymptotic formulas of §3 and by bootstrap, using 50 bootstrap samples. The ‘true’ covariance matrix of $\text{vec}((\hat{A})^T)$ was computed as the sample covariance of the 500 replicated estimators. Since we are interested in testing, we computed tail probabilities of $Q = \text{vec}((\hat{A})^T)^T\hat{\Sigma}^{-1}\text{vec}((\hat{A})^T)$, where $\hat{\Sigma}$ is the respective covariance estimator of $\text{vec}((\hat{A})^T)$, and of $Z_{1j} = \hat{a}_{1j}/\hat{sd}(\hat{a}_{1j})$ for $j = 1, 2$. Specifically, we report $P(Q \geq 7.78$) and $P(|Z_{1j}| \geq 1.645$) for $j = 1, 2$, which should be close to 0.10.

Table 4 shows the results. There are two aspects of the asymptotics that we are trying to assess: the adequacy of the normal approximation and the adequacy of the variance estimators. The first aspect can be best assessed using the true variance in the test statistics, so the variance estimation error is not a confounding factor. In this regard we see in Table 4...
that the asymptotic approximation is good even for \( n = 50 \), both for the global \( Q \)-test and for the marginal \( Z \)-tests. In the more realistic cases where the variance is estimated, we see that bootstrap variance estimators generally work better than the asymptotic-variance formula; although both underestimate the true variances, bootstrap tends to underestimate them less, especially for random time grids.

5 Application: Modeling Ground-Level Ozone Concentration

Ground-level ozone is an air pollutant known to cause 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 (site 3011 in the database) 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 \( \tau_{x0} \) and \( \tau_{y0} \) 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. The second model did not seem to represent much of an improvement over the first one, so we discarded it. For models (i) and (iii) the estimated regression coefficients and the bootstrap standard deviations, based on 200 resamples, were

\[
\hat{A} = \begin{bmatrix} 0.73 & 0.09 \\ 0.19 & 0.44 \end{bmatrix}, \quad \text{std}(\hat{A}) = \begin{bmatrix} 0.07 & 0.02 \\ 0.08 & 0.06 \end{bmatrix},
\]

\[
\hat{A} = \begin{bmatrix} 0.36 & 0.12 \\ 0.01 & 0.02 \\ 0.18 & 0.54 \end{bmatrix}, \quad \text{std}(\hat{A}) = \begin{bmatrix} 0.08 & 0.06 \\ 0.04 & 0.10 \\ 0.06 & 0.11 \end{bmatrix}.
\]

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 9. A two-component model, of overall dimension 4 and thus comparable to the warped regression model, would correspond to the upper four panels of Figure 3. 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)).
Figure 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.
Figure 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.
so it really takes a 9-dimensional ordinary regression model to explain the phase-variability features that a 4-dimensional warped model would explain. And the predominantly time-related principal components, Figure 3(c,f), are also associated with some kinds of amplitude variability. Likewise, principal components that are predominantly amplitude-related, like the first \( x \)-component, Figure 3(a), are somewhat influenced by time variability. This blurring of the components is avoided by warped functional regression, which neatly separates the sources of variability and offers not only a more easily interpretable model but also a lower-dimensional one.

**Acknowledgement**

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

**Supplementary material**

Supplementary material available online includes a more thorough discussion of model identifiability, the derivation of the EM algorithm for estimation, detailed derivation of formulae involved in the asymptotic distribution of the estimator, 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 \( \mathcal{S} = [a, b] \) and \( a < \tau_{01} < \cdots < \tau_{0r} < b \) be a sequence of \( r \) knots in \( \mathcal{S} \). Define the basis functions \( \{\alpha_j(s; \tau_0)\} \) and \( \{\beta_j(s; \tau_0)\} \) as follows: let \( h_{00}(s) = (1 + 2s)(1 - s)^2 \) and \( h_{10}(s) = s(1 - s)^2 \); then

\[
\alpha_0(s; \tau_0) = \begin{cases} 
0 & \text{if } s < a \text{ or } s > \tau_{01} \\
\frac{s-a}{\tau_{01}-a} & \text{if } a \leq s \leq \tau_{01},
\end{cases}
\]
\[ \alpha_j(s; \tau_0) = \begin{cases} 0 & \text{if } s < \tau_{0,j-1} \text{ or } s > \tau_{0,j+1} \\ h_0 \left( \frac{\tau_{0,j-s}}{\tau_{0,j-1} - \tau_{0,j-1}} \right) & \text{if } \tau_{0,j-1} \leq s \leq \tau_{0,j} \\ h_0 \left( \frac{s-\tau_{0,j}}{\tau_{0,j+1} - \tau_{0,j}} \right) & \text{if } \tau_{0,j} \leq s \leq \tau_{0,j+1} \end{cases} \]

for \( j = 1, \ldots, r \),

\[ \alpha_{r+1}(s; \tau_0) = \begin{cases} 0 & \text{if } s < \tau_{0,r} \text{ or } s > b \\ h_0 \left( \frac{b-s}{b-\tau_{0,r}} \right) & \text{if } \tau_{0,r} \leq s \leq b, \end{cases} \]

\[ \beta_0(s; \tau_0) = \begin{cases} 0 & \text{if } s < a \text{ or } s > \tau_{0,1} \\ (\tau_{0,1} - a)h_{10} \left( \frac{s-a}{\tau_{0,1} - a} \right) & \text{if } a \leq s \leq \tau_{0,1}, \end{cases} \]

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

for \( j = 1, \ldots, r \), and

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

The function

\[ \omega_i(s) = \sum_{j=0}^{r+1} \tau_{ij} \alpha_j(s; \tau_0) + \sum_{j=0}^{r+1} d_{ij} \beta_j(s; \tau_0), \tag{16} \]

where \( \tau_{i0} = a \) and \( \tau_{i,r+1} = b \), is a differentiable piecewise-cubic function that satisfies \( \omega_i(\tau_{0,j}) = \tau_{ij} \) and \( \omega_i'(\tau_{0,j}) = d_{ij} \) for \( j = 1, \ldots, r \). Thus the \( \tau_{ij} \)s play the dual role of basis coefficients and values of \( \omega_i(s) \) at the knots. For (16) 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 \( \tau_{ij} \)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 \( \tau_{ij} \)s and then (16) 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

$$
\theta_{ij} = \log \left( \frac{\tau_{ij+1} - \tau_{ij}}{\tau_{ij} - \tau_{ij-1}} \right), \quad j = 1, \ldots, r,
$$

with inverse given by

$$
\tau_{ij} = a + (b - a) \cdot \frac{\sum_{k=1}^{j} \exp(\theta_{i1} + \cdots + \theta_{ik})}{1 + \sum_{k=1}^{r} \exp(\theta_{i1} + \cdots + \theta_{ik})}, \quad j = 1, \ldots, r.
$$

Note that for any \( r \)-dimensional unconstrained vector \( \theta \) the inverse Jupp transform yields a vector \( \tau \) of strictly increasing knots in \((a, b)\). In particular, for \( \theta = 0 \) the corresponding \( \tau \) is a sequence of \( r \) equally spaced knots in \((a, b)\).

### 5.2 Likelihood function

Under the distributional assumptions in Section 2.2, the likelihood function is derived as follows. The joint density function of the data vectors \((x_i, y_i)\) and the latent random effects \((w_i, z_i)\) can be factorized as

$$
f(x_i, y_i, w_i, z_i) = f(x_i, y_i | w_i, z_i) f(z_i | w_i) f(w_i)
$$

since \( y_i \) depends on \( w_i \) only through \( z_i \), according to (7). Clearly \( w_i \sim N(\mu_w, \Sigma_w) \) and \( z_i | w_i \sim N(\mu + A(w_i - \mu_w), \Sigma_e) \). The conditional distributions \( x_i | w_i \) and \( y_i | z_i \) are derived as follows. Given \( w_i = (u_i^T, \theta_{x1}^T) \) and \( z_i = (v_i^T, \theta_{y1}^T) \), the values of \( \theta_{x1} \) and \( \theta_{y1} \) determine the warping functions \( \omega_i(s) \) and \( \zeta_i(t) \) and consequently two warped time grids \( s_{ij} = \omega_i^{-1}(s_{ij}), \quad j = 1, \ldots, \nu_1 \), and \( t_{ij} = \zeta_i^{-1}(t_{ij}), \quad j = 1, \ldots, \nu_2 \). Let \( B_{x1}^* \in \mathbb{R}^{\nu_1 \times q_1} \) and \( B_{y1}^* \in \mathbb{R}^{\nu_2 \times q_2} \) be the B-spline bases evaluated at the warped time grids, that is \( B_{x1,i,jk} = b_{xk}(s_{ij}) \) and \( B_{y1,i,jk} = b_{yk}(t_{ij}) \). Then, in view of model specifications (11)-(16) we have \( x_i | w_i \sim N(B_{x1}^* m_x + B_{x1}^* C_u, \sigma_e^2 I_{\nu_1}) \) and \( y_i | z_i \sim N(B_{y1}^* m_y + B_{y1}^* D v_i, \sigma_\eta^2 I_{\nu_2}) \). The maximum likelihood estimators maximize

$$
\ell(A, \Sigma_e, \Sigma_w, m_x, m_y, C, D, \sigma_e^2, \sigma_\eta^2) = \sum_{i=1}^{n} \log \int \int f(x_i, y_i, w, z) \, dw \, dz \quad (17)
$$

23
but the integrals in (17) do not have closed forms so we use the EM algorithm to find
the optimum, treating the random effects \((w_i, z_i)\) as missing data. Most of the updating
equations of the EM algorithm are easy to derive but the restrictions on the parameters \(C, \ D, \text{ and } A\) pose some difficulties. This is discussed in detail in the Supplementary Material.

**Proof of Theorem 1**

This proof is a direct application of Theorem 4.4 of Geyer (1994); note that Theorem
5.2 of Geyer (1994), which pertains to consistent local minimizers instead of global min-
imizers, can also be applied because our \(T_{C}(\zeta_0)\) satisfies the stronger condition of be-
ing Clarke-regular (Rockafellar & Wets, 1998, ch. 6.B). Following Geyer’s notation, let
\(F(\zeta) = E\{-\log f(x, y; \zeta)\}\) and \(F_n(\zeta) = -(1/n) \sum_{i=1}^{n} \log f(x_i, y_i; \zeta)\). Then \(\hat{\zeta}_n = \arg\min_{\zeta \in \mathcal{C}} F_n(\zeta)\) and \(\zeta_0 = \arg\min_{\zeta \in \mathcal{C}} F(\zeta)\). Assumption A of Geyer (1994) is that
\[
F(\zeta) = F(\zeta_0) + \frac{1}{2}(\zeta - \zeta_0)^T V(\zeta - \zeta_0) + o(\|\zeta - \zeta_0\|),
\]
with \(V = \nabla^2 F(\zeta_0)\) positive definite. This is satisfied in our case because \(\nabla F(\zeta_0) = -E\{\nabla \log f(x, y; \zeta_0)\} = 0\) and \(\nabla^2 F(\zeta_0) = E\{U(x, y)U(x, y)^T\}\). To see that the latter is positive definite, note that for \(\zeta\) as in (10) we have
\[
U(x, y)^T \zeta = \text{tr}\{\Sigma_{w,0}^{-1}N(x, y)^T A^T\} - \text{tr}\{\Sigma_{e,0}^{-1}A_0 M(x, y)A^T\}
- \frac{1}{2}\text{tr}\{\Sigma_{e,0}^{-1} \Sigma_{w,0} - \Sigma_{w,0}^{-1}M(x, y)\Sigma_{w,0}^{-1} \Sigma_{w,0}\}
= E\{(w - \mu_w)^T A^T \Sigma_{e,0}^{-1} e(x, y)\}
- \frac{1}{2}\text{tr}\{\Sigma_{w,0}^{-1} \Sigma_{w,0}\} + \frac{1}{2}E\{(w - \mu_w)^T \Sigma_{w,0}^{-1} \Sigma_{w,0}^{-1} \Sigma_{w,0} (w - \mu_w)\|(x, y)\},
\]
where \(e = z - \mu_z + A_0 (w - \mu_w)\), then \(\zeta^TV\zeta = E[\{U(x, y)^T \zeta\}^2] \geq 0\) and it is equal
to zero only if \(U(x, y)^T \zeta = 0\) with probability one, which can only happen if \(\zeta = 0\).

Assumption B of Geyer, in our case, is that
\[
-\log f(x, y; \zeta) = -\log f(x, y; \zeta_0) + (\zeta - \zeta_0)^T D(x, y) + ||\zeta - \zeta_0|| r(x, y, \zeta)
\]
for some \(D(x, y)\) such that the remainder \(r(x, y, \zeta)\) is stochastically equicontinuous. This
is satisfied by $D(x, y) = -\nabla \log f(x, y; \zeta_0)$; the fact that $r(x, y, \zeta)$ is stochastically equicontinuous follows from Pollard (1984, pp. 150–152). Clearly $D(x, y)$ satisfies a Central Limit Theorem with asymptotic covariance matrix $A$ that in this case is equal to $V$, so Assumption C of Geyer is also satisfied. Then Theorem 4.4 of Geyer can be applied. It states that the asymptotic distribution of $\sqrt{n}(\hat{\zeta}_n - \zeta_0)$ is the same as the distribution of $\hat{\delta}(Z)$, the minimizer of

$$q_Z(\delta) = \delta^T Z + \frac{1}{2} \delta^T V \delta$$

over $\delta \in T_C(\zeta_0)$, where $Z \sim N(0, A)$.

In our case $\hat{\delta}(Z)$ can be obtained in closed form, due to the simplicity of $T_C(\zeta_0)$. Concretely, $T_C(\zeta_0)$ is the space of $\delta$s such that $B\delta = 0$. Let $\Omega = [\Xi^*, \Xi]$ be a $d \times d$ orthogonal matrix whose first $m$ columns $\Xi^*$ span the space generated by the rows of $B$ and whose last $d - m$ columns $\Xi$ are orthogonal to the rows of $B$. Then $\delta \in T_C(\zeta_0)$ if and only if $\delta = \Omega\beta$ with $\beta_1 = \cdots = \beta_m = 0$; that is, $\delta = \Xi\beta_2$ with $\beta_2$ the subvector containing the last $d - m$ coordinates of $\beta$. Then for $\delta \in T_C(\zeta_0)$ we can write

$$q_Z(\delta) = \beta_2^T \Omega^T \Xi + \frac{1}{2} \beta_2^T \Omega^T V \Omega \beta_2$$

$$= \beta_2^T \Xi^T Z + \frac{1}{2} \beta_2^T \Xi^T V \Xi \beta_2,$$

which is clearly minimized by $\hat{\beta}_2 = (\Xi^T V \Xi)^{-1} \Xi^T Z$. Therefore $\hat{\delta}(Z) = \Xi(\Xi^T V \Xi)^{-1} \Xi^T Z$, and since $A = V$, the result of the theorem follows.

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. *The Annals of Statistics* 34 2159–2179.
Crambes, C., Kneip, A., & Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *The Annals of Statistics* **37** 35–72.

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

Gervini, D. & Gasser, T. (2004). Self-modeling warping functions. *Journal of the Royal Statistical Society (Series 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. *The Annals of Statistics* **22** 1993–2010.

Hall, P. & Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics* **35** 70–91.

James, G.M. (2007). Curve alignment by moments. *The Annals of Applied Statistics* **1** 480–501.

James, G., Hastie, T. G. & 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. *The Annals of Statistics* **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 Journal of Statistics* **28** 19–30.

Kneip, A. & Ramsay, J.O. (2008). Combining registration and fitting for functional models. *Journal of the American Statistical Association* **103** 1155–1165.
Liu, X. & Müller, H.-G. (2004). Functional convex averaging and synchronization for time-warped random curves. *Journal of the American Statistical Association* 99 687–699.

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

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

Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer, New York.

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). *Statistical Science* 3 425–461.

Ramsay, J.O. & Li, X. (1998). Curve registration. *Journal of the Royal Statistical Society (Series 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. & Müller, H.-G. (2008). Pairwise curve synchronization for functional data. *Biometrika* 95 875–889.

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

Yao, F., Müller, H.-G. & Wang, J.-L. (2005). Functional linear regression analysis for longitudinal data. *The Annals of Statistics* 33 2873–2903.

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