Dynamic Retrospective Regression for Functional Data

Daniel Gervini

Department of Mathematical Sciences, University of Wisconsin–Milwaukee

3200 N Cramer St, Milwaukee, WI 53211

December 4, 2013

Abstract

Samples of curves, or functional data, usually present phase variability in addition to amplitude variability. Existing functional regression methods do not handle phase variability in an efficient way. In this paper we propose a functional regression method that incorporates phase synchronization as an intrinsic part of the model, and then attains better predictive power than ordinary linear regression in a simple and parsimonious way. The finite-sample properties of the estimators are studied by simulation. As an example of application, we analyze neuromotor data arising from a study of human lip movement. This paper has supplementary material online.

Key Words: Curve Registration; Functional Data Analysis; Hermite Spline; Spline Smoothing; Time Warping.

1 Introduction

Many statistical applications today involve modeling curves as functions of other curves. For example, the trajectories of CD4 cell counts over time in HIV patients can be modeled as functions of viral load trajectories (Liang et al. 2003, Wu and Liang 2004, Wu and Müller 2011); gene expression profiles of insects at the pupal stage can be modeled as functions of gene expression profiles at the embryonic stage (Müller et al. 2008); trajectories of systolic blood pressure over the years can be predicted to some extent from trajectories of body mass index (Yao et al. 2005); and daily electricity consumption trajectories can be predicted on the basis of previous days' consumption trajectories (Antoch et al. 2008). All of these examples fall into the relatively new area of functional regression, or regression methods for functional data.

Functional linear regression, in particular, is a more or less straightforward extension of multivariate linear regression to the functional-data framework (Ramsay and Silverman 2005, ch. 16). Recent developments in functional linear regression have focused on theoretical aspects such as rates of convergence (Cai and Hall 2006, Hall and Horowitz 2007, Crambes et al. 2009), sparse longitudinal data (Yao et al. 2005), and interpretability of the estimators (James et al. 2009). But a problem inherent to functional data that has received little attention in the regression context is the problem of phase variability.

As a motivating example, consider the data in Malfait and Ramsay (2003). The authors want to predict lip acceleration using electromyography (EMG) curves that measure neural activity in the primary muscle that depresses the lower lip, the depressor labii inferior. A person was asked to repeat the phrase "say Bob again" a few times, and the lip movement and associated EMG curve corresponding to the word "Bob" were recorded. Lip acceleration curves were obtained by differentiating the smoothed lip trajectories. The sample curves, time-standardized to 700 msec, are shown in Figure 1(a,b). Both samples follow regular patterns, but they show considerable variability in amplitude and timing of the main features. In fact, phase variability overwhelms amplitude variability in Figure

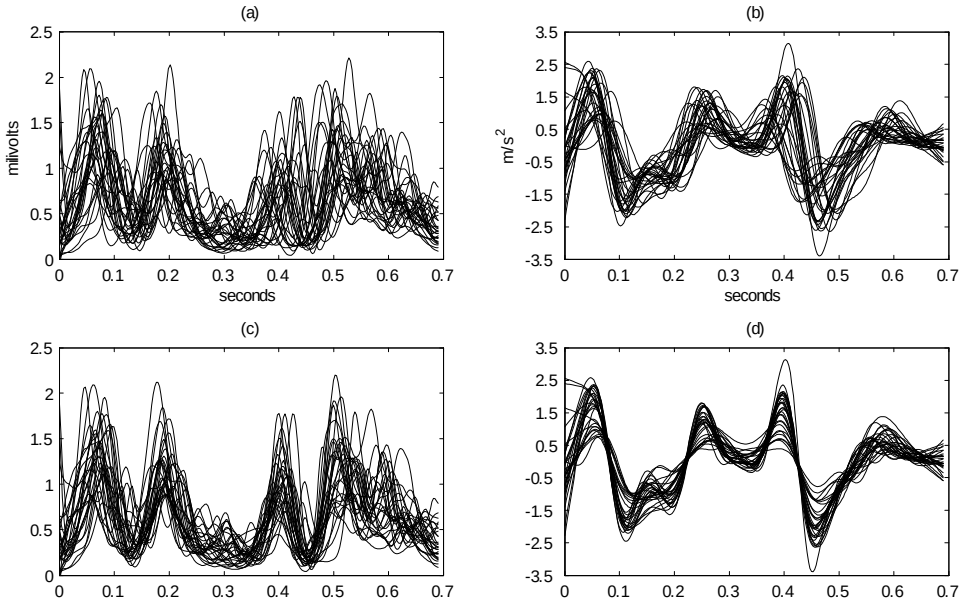

Figure 1: Lip Movement Example. (a) EMG curves; (b) lip acceleration curves; (c) synchronized EMG curves; (d) synchronized lip acceleration curves.

1(a), to the point that it is hard to tell how many systematic peaks a typical EMG curve has in the range .3–.7. A pair-by-pair analysis of the curves shows that the EMG spikes are aligned with certain features of the acceleration curves; therefore, the *timing* of the EMG spikes (not just their amplitude) is likely to provide valuable information for predicting lip acceleration.

Ordinary functional linear regression does not model phase variability explicitly. This creates some problems, because phase variability tends to spread the features of predictor and response curves over wide time ranges and as a result the regression function becomes very irregular and hard to interpret. On the other hand, if the curve features are synchronized, a simpler regression function will provide a good fit to the data.

Several methods of curve synchronization have been proposed over the years. We can mention Gervini and Gasser (2004, 2005), James (2007), Kneip et al. (2000), Kneip and Ramsay (2008), Liu and Müller (2004), Ramsay and Li (1998), Tang and Müller

(2008, 2009), and Wang and Gasser (1999), among others. But in a regression context, if covariate and response curves are synchronized independently it becomes impossible to predict new (un-warped) response curves from given (un-warped) covariate curves, since the associated warping functions cannot be predicted.

To address this problem, in this paper we propose a regression method that incorporates time warping as an intrinsic part of the model. Since we are going to apply this method to the lip movement data, we will focus on the retrospective regression model, or "historical" regression model as Malfait and Ramsay (2003) call it, where x and y are functions of time and for each t we model the response value $y(t)$ as a function of "past" covariate values $x(s)$ with $s \leq t$. In addition, we assume that the sample curves are smooth, as in Figure 1. Extending this model to sparse and irregular curves is something that will be addressed in other papers.

2 Dynamic retrospective regression

2.1 Model and estimation

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a sample of functions, where x_i is the covariate curve and y_i the response curve. We assume x_i and y_i are square-integrable functions on a common interval $[a, b]$. A linear predictor of $y_i(t)$ based on x_i has the form

$$L(t; x_i, \alpha, \beta) = \alpha(t) + \int_a^b \beta(s,t) x_i(s) ds, \tag{1}$$

where α is the intercept function and β the slope function. However, (1) employs the whole trajectory $\{x_i(s) : s \in [a, b]\}$ to predict $y_i(t)$, including "future" observations $x_i(s)$ with $s > t$. In many applications this is not reasonable. For example, for the lip movement data in Figure 1 it is clear that future neural activity cannot have an influence on past lip movement; therefore, prediction of $y_i(t)$ must be based only on the partially observed

curves $\{x_i(s) : s \in [a,t]\}$. Then instead of (1) we will use

$$RL(t; x_i, \alpha, \beta) = \alpha(t) + \int_a^t \beta(s,t) x_i(s) ds, \tag{2}$$

which can be seen as a particular case of (1) under the constraint $\beta(s,t) = 0$ for $s > t$. This model is called "historical linear model" by Malfait and Ramsay (2003), although we prefer the denomination "retrospective linear model".

As explained in the Introduction, ordinary functional linear regression works best for synchronized curves. Suppose, then, that for each pair (x_i, y_i) we have a warping function $w_i : [a,b] \to [a,b]$, that is, a strictly monotone increasing function that satisfies $w_i(a) = a$ and $w_i(b) = b$. Let $\tilde{x}_i = x_i \circ w_i$ and $\tilde{y}_i = y_i \circ w_i$ be the warped curves, where $(x_i \circ w_i)(t) = x_i(w_i(t))$; then we apply (2) to $(\tilde{x}_i, \tilde{y}_i)$ rather than (x_i, y_i), and define the *dynamic functional predictor* of $y_i(t)$ as $RL(w_i^{-1}(t); \tilde{x}_i, \alpha, \beta)$, obtaining

$$\hat{y}_i(t) = \alpha(w_i^{-1}(t)) + \int_a^{w_i^{-1}(t)} \beta(s, w_i^{-1}(t)) x_i(w_i(s)) ds. \tag{3}$$

Note that the same warping function w_i is used for x_i and y_i; this is reasonable for the type of applications we have in mind. Using a common warping function preserves the retrospective property of the model: the integral in (3) only involves values of $x_i(w_i(s))$ with $s \le w_i^{-1}(t)$, or equivalently, $x_i(s)$ with $s \le t$.

The estimators of α, β, and the w_is can be obtained by functional least squares, minimizing

$$\sum_{i=1}^{n} \| y_i \circ w_i - RL(\cdot; x_i \circ w_i, \alpha, \beta) \|^2 \tag{4}$$

with respect to α, β and the w_is, where $\|f\| = \{\int_a^b f^2(t)dt\}^{1/2}$ is the usual $L^2([a,b])$-norm. Note that for given β and w_is, the α that minimizes (4) is

$$\hat{\alpha}(t) = \overline{\tilde{y}}(t) - \int_a^t \beta(s,t) \overline{\tilde{x}}(s) ds, \tag{5}$$

so we can re-write (4) as

$$\sum_{i=1}^{n} \int_a^b \left[\tilde{y}_i(t) - \overline{\tilde{y}}(t) - \int_a^t \beta(s,t)\{\tilde{x}_i(s) - \overline{\tilde{x}}(s)\}ds \right]^2 dt, \tag{6}$$

eliminating the intercept α.

The estimation of β has to be done with care in order to avoid identifiability issues. To understand this problem, consider again the general linear predictor (1). For any function γ such that $\int_a^b \gamma(s,t)x_i(s)ds = 0$ for all t and all i, it is clear that $L(t; x_i, \alpha, \beta) = L(t; x_i, \alpha, \beta + \gamma)$ for all t and all i, so (1) cannot distinguish between β and $\beta + \gamma$. Since the space spanned by the x_is has dimension at most n, there is always going to be an infinite number of γs for which this occurs. The usual way to deal with this identifiability issue is to reduce the space of possible βs, so that the only γ that satisfies $\int_a^b \gamma(s,t)x_i(s)ds = 0$ for all t and all i in the reduced space is $\gamma \equiv 0$. An efficient way to do this is to use the tensor-product space of the principal components of the x_is and the y_is (e.g. as in Müller *et al.*, 2008), which is the functional equivalent of principal-component regression.

We briefly remind the reader what the functional principal components are. A continuous covariance function $\rho(s,t) = \mathrm{cov}\{x(s), x(t)\}$ admits the decomposition $\rho(s,t) = \sum_k \lambda_k \phi_k(s)\phi_k(t)$, where the ϕ_ks are orthogonal functions in $L^2([a,b])$ and the λ_ks are non-increasing positive scalars (the sequence may be finite or infinite, but in either case $\sum_k \lambda_k < \infty$). This is known as Mercer's Theorem (Gohberg *et al.*, 2003). The ϕ_ks are eigenfunctions of ρ, i.e. they satisfy $\int_a^b \rho(s,t)\phi_k(s)ds = \lambda_k \phi_k(t)$ for all t. The ϕ_ks are called the principal components of the x-space, since they are the functional equivalents of the multivariate principal components. In a similar way one obtains the principal components of the y-space, say $\{\psi_l\}$. To estimate the regression function β, one chooses the leading principal components of each space, say $\{\phi_1, \ldots, \phi_p\}$ and $\{\psi_1, \ldots, \psi_q\}$, and sets

$$\beta(s,t) = \sum_{k=1}^{p} \sum_{l=1}^{q} b_{kl} \phi_k(s)\psi_l(t). \tag{7}$$

The coefficients b_{kl} of β are estimated by least squares. This procedure can be adapted for the retrospective linear predictor (2) in a straightforward way, since the estimation of the principal components does not change.

Going back to the problem of minimizing (6), we proceed as follows. First note that the ϕ_ks and the ψ_ls are now the principal components of the warped functions $\{\tilde{x}_i\}$ and $\{\tilde{y}_i\}$, respectively, so we cannot estimate them separately in a preliminary step because they depend on the warping functions $\{w_i\}$, which are themselves estimated in the process. So we minimize (6) with respect to $\mathbf{B} = [b_{kl}]$ and the w_is, subject to the conditions (7) and

$$\int_a^b \rho_{\tilde{x}}(s,t)\phi_k(s)ds = \lambda_k\phi_k(t), \quad k = 1,\ldots,p, \tag{8}$$

$$\int_a^b \rho_{\tilde{y}}(s,t)\psi_k(s)ds = \xi_k\psi_k(t), \quad k = 1,\ldots,q, \tag{9}$$

where

$$\rho_{\tilde{x}}(s,t) = \frac{1}{n}\sum_{i=1}^{n}\{\tilde{x}_i(s) - \overline{\tilde{x}}(s)\}\{\tilde{x}_i(t) - \overline{\tilde{x}}(t)\},$$

$$\rho_{\tilde{y}}(s,t) = \frac{1}{n}\sum_{i=1}^{n}\{\tilde{y}_i(s) - \overline{\tilde{y}}(s)\}\{\tilde{y}_i(t) - \overline{\tilde{y}}(t)\}.$$

In addition, we assume that the ϕ_ks and the ψ_ls are orthonormal and that the sequences $\{\lambda_k\}$ and $\{\xi_k\}$ are positive and non-increasing. For identifiability of the warping functions, we also add the constraint $\bar{w}(t) \equiv t$.

It is convenient to model the functional parameters $\{\phi_k\}$, $\{\psi_l\}$ and $\{w_i\}$ using splines or similar basis functions, because this reduces the functional minimization problem to a more familiar multivariate minimization problem. Let $\boldsymbol{\gamma}(t) = (\gamma_1(t),\ldots,\gamma_\nu(t))$ be a spline basis (or some other system) in $L^2([a,b])$; then we assume $\phi_k(t) = \sum_{j=1}^{\nu} c_{kj}\gamma_j(t)$ and $\psi_l(t) = \sum_{j=1}^{\nu} d_{lj}\gamma_j(t)$ for coefficient vectors $\mathbf{c}_k$ and $\mathbf{d}_l$. The regression slope can then be expressed as

$$\beta(s,t) = \boldsymbol{\gamma}(s)^T\mathbf{C}^T\mathbf{B}\mathbf{D}\boldsymbol{\gamma}(t)I\{s \le t\}$$

and the functional constraints (8) and (9) turn into parametric constraints

$$\mathbf{\Omega}_{\tilde{x}}\mathbf{C} = \mathbf{J}_0\mathbf{C}\mathbf{\Lambda}, \tag{10}$$

$$\mathbf{\Omega}_{\tilde{y}}\mathbf{D} = \mathbf{J}_0\mathbf{D}\mathbf{\Xi}, \tag{11}$$

where $\mathbf{\Omega}_{\tilde{x}} = \iint \rho_{\tilde{x}}(s,t)\boldsymbol{\gamma}(s)\boldsymbol{\gamma}(t)^T\,ds\,dt$, $\mathbf{\Omega}_{\tilde{y}} = \iint \rho_{\tilde{y}}(s,t)\boldsymbol{\gamma}(s)\boldsymbol{\gamma}(t)^T\,ds\,dt$, $\mathbf{J}_0 = \int \boldsymbol{\gamma}(t)\boldsymbol{\gamma}(t)^T dt$, $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_p]$, $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_q]$, $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ and $\mathbf{\Xi} = \mathrm{diag}(\xi_1, \ldots, \xi_q)$. Note that $\mathbf{\Omega}_{\tilde{x}}$ and $\mathbf{\Omega}_{\tilde{y}}$ are functions of the w_is via $\rho_{\tilde{x}}$ and $\rho_{\tilde{y}}$, but we omit this in the notation for simplicity. In addition, we also have the orthogonality conditions $\mathbf{C}^T\mathbf{J}_0\mathbf{C} = \mathbf{I}_p$ and $\mathbf{D}^T\mathbf{J}_0\mathbf{D} = \mathbf{I}_q$.

Parameterizing the warping functions is more complicated due to their monotonicity. One possibility is to model the w_is as B-spline functions with monotone increasing coefficients, which guarantees that the w_is are monotone increasing (Brumback and Lindstrom, 2004); the boundary conditions $w_i(a) = a$ and $w_i(b) = b$ and the identifiability condition $\bar{w}(t) \equiv t$ can be expressed as linear constraints on the coefficients. Another possibility is to use the family of smooth monotone transformations (Ramsay and Li, 1998), where $\log\{w_i'(t)\}$ is modeled as an unconstrained B-spline function and $w_i(t)$ is computed by integration; if $\boldsymbol{\theta}_i \in \mathbb{R}^r$ are the spline coefficients of $\log\{w_i'(t)\}$, a convenient identifiability condition is the restriction $\bar{\boldsymbol{\theta}} = \mathbf{0}$, which approximately implies $\bar{w}(t) \equiv t$. A third possibility, which is the one we prefer in this paper, is to model the w_is as monotone interpolating cubic Hermite splines (Fritsch and Carlson, 1980). This family is specified by a vector of knots $\boldsymbol{\tau}_0 \in \mathbb{R}^r$ in (a, b) and each w_i is determined by a corresponding vector $\boldsymbol{\tau}_i$ such that $w_i(\boldsymbol{\tau}_0) = \boldsymbol{\tau}_i$. The $\boldsymbol{\tau}_i$s then become the parameters that determine w_i. The strategy, when using Hermite splines, is to place the knots $\boldsymbol{\tau}_0$ at locations of interest, such as the (approximate) average location of peaks and valleys. For example, for the lip movement data in Figure 1(a) a reasonable choice would be $\boldsymbol{\tau}_0 = (.1, .2, .4, .5)$, corresponding to the approximate average location of the peaks of the x_is (the peaks at $.4$ and $.5$ are hard to see

in Figure 1(a) but they are easy to spot when the curves are plotted individually; see also the aligned curves in Figure 1(c).) This way we obtain warping flexibility at the features of interest with a low-dimensional family of warping functions, since the $\boldsymbol{\tau}_i$s can take any value as long as $a < \tau_{i1} < \cdots < \tau_{ir} < b$. One technicality: due to this monotonicity restriction, it is computationally more convenient to use the Jupp transforms (Jupp, 1978) of the $\boldsymbol{\tau}_i$s,

$$\theta_{ij} = \log\{(\tau_{i,j+1} - \tau_{ij})/(\tau_{ij} - \tau_{i,j-1})\}, \;\; j = 1, \ldots, r,$$

as parameters, because the $\boldsymbol{\theta}_i$s are unconstrained vectors. The identifiability condition $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$ approximately implies that $\bar{\boldsymbol{\tau}} = \boldsymbol{\tau}_0$ and therefore $\bar{w}(t) \equiv t$. More details about monotone Hermite splines are given in the Technical Supplement.

The minimization of (6) has thus become a multivariate constrained minimization problem on the parameters $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\Xi}$, and $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$, which can be solved via standard optimization methods (see e.g. Nocedal and Wright, 2006, ch. 15). In our Matlab programs we use the interior-point algorithm as implemented in Matlab's function "fmincon". This type of algorithms converge only to a local minimum, so it is important to select a good starting point to increase the chances of actually finding the global minimum, or at least a "good" local solution. One approach we have found successful is to do a quick (separate) synchronization of the x_is and the y_is and use the resulting principal components as initial estimators of the ϕ_ks and the ψ_ls, and the warping parameters of either sample as initial $\boldsymbol{\theta}_i$s. Another alternative is to try several random starting points, but this is much more time consuming.

Once the estimators $\hat{\alpha}$, $\hat{\beta}$ and $\hat{w}_1, \ldots, \hat{w}_n$ have been obtained, it is possible to predict a response function y_{n+1} for a given covariate function x_{n+1}. The natural predictor of y_{n+1} given x_{n+1} is $\hat{y}_{n+1}(t) = RL(\hat{w}_{n+1}^{-1}(t); \tilde{x}_{n+1}, \hat{\alpha}, \hat{\beta})$, but $\hat{w}_{n+1}$ cannot be obtained by minimizing the integrated squared residual because that involves the unobserved response y_{n+1}. Instead, $\hat{w}_{n+1}$ can be obtained from x_{n+1} alone by synchronizing x_{n+1} to the mean

of the warped x_is:

$$\hat{w}_{n+1} = \arg\min_{w} \left\| x_{n+1} \circ w - \overline{\tilde{x}} \right\|^2 . \tag{12}$$

Note that $\overline{\tilde{x}}$ here is fixed, so the "pinching" or "overwarping" problem associated with least-squares registration [discussed in Ramsay and Silverman (2005, ch. 7.6) and Kneip and Ramsay (2008)] will not be a serious issue.

2.2 Selection of meta-parameters

In addition to the parameters estimated by least squares, there are some meta-parameters that also need to be specified. For example, the number and placement of knots of the spline bases used for the ϕ_ks, the ψ_ls and the w_is, and most importantly p and q, the number of principal components to be included in (7). The simplest approach would be to minimize computationally simple criteria such as the "generalized cross validation" criterion

$$\mathrm{GCV}(p, q) = \mathrm{MSE}(p, q)/(1 - pq/n)^2$$

(Wahba, 1990) or the "corrected Akaike criterion"

$$\mathrm{AICC}(p, q) = \mathrm{MSE}(p, q) \exp\{1 + 2(pq + 1)/(n - pq - 2)\}$$

(Hurvitz *et al.*, 1998), where $\mathrm{MSE} = n^{-1} \sum_{i=1}^{n} \|\tilde{y}_i^* - \widehat{\tilde{y}}_i^*\|^2$ and $\tilde{y}_i^*(t) = \tilde{y}_i(t) - \overline{\tilde{y}}(t)$. Unfortunately these criteria did not perform well in our simulations, so a more computationally complex approach such as k-fold cross-validation (Hastie *et al.*, 2009) needs to be explored.

Regarding the spline bases for the ϕ_ks and the ψ_ls, we note that since the method "borrows strength" across curves to estimate the coefficients $\{\mathbf{c}_k\}$ and $\{\mathbf{d}_k\}$, the spline dimension ν can be relatively large and the resulting ϕ_ks and the ψ_ls will still be reasonably regular. Thus we simply take a fairly large number of equally spaced points in (a, b) as

knots. If necessary, roughness penalty terms can be added to control the regularity of the ϕ_ks and the ψ_ls. In our implementation, we do that by adding the penalty to the constraints (as in Silverman, 1996), substituting $\mathbf{J}_0$ in (10), (11) and in the orthogonality constraints by $\mathbf{J}_0 + \eta\mathbf{J}_2$, where $\mathbf{J}_2 = \int \boldsymbol{\gamma}''(t)\boldsymbol{\gamma}''(t)^T dt$ and η is a roughness-penalty parameter, that in practice is chosen subjectively.

Regarding the warping functions, the approach to follow will depend on the warping family. If interpolating Hermite splines are used, a small number of knots r placed nearby the salient landmarks usually provide ample warping flexibility, and r can be chosen by cross-validation along with p and q within a small range of triplets (p, q, r). If smooth monotone transformations are used, for which spline knots are not identified with meaningful landmarks, a better approach is to use several equally-spaced knots and add the roughness-penalty term $\eta \sum_{i=1}^{n} \int \{(\log w_i')'\}^2$ to the objective function, as in Ramsay and Li (1998); in that case the smoothing parameter η must be chosen with care, because it determines the effective dimension of the warping space and therefore there is going to be an interplay between p, q and η.

3 Simulations

We ran two sets of simulations to assess the performance of the proposed method. The first set was designed to compare the dynamic regression estimator with the ordinary functional least squares estimator, to determine to what extent the estimators are able to reconstruct the true regression function β. The data was generated as follows. We generated $\tilde{x}_i$s following the shape-invariant model $\tilde{x}_i(s) = z_i e^{-30(s-.4)^2}$ with z_i i.i.d. $N(1, .2^2)$, which is a one-component model with $\mu_{\tilde{x}}(s) = e^{-30(s-.4)^2}$ and $\phi_1 = \mu_{\tilde{x}}/\|\mu_{\tilde{x}}\|$. The $\tilde{y}_i$s were generated as

$$\tilde{y}_i(t) = \int_0^t \beta(s,t)\tilde{x}_i(s)ds + \varepsilon_i(t) \tag{13}$$

with $\beta(s,t) = 5e^{-50\{(s-.4)^2+(t-.6)^2\}}$, and $\varepsilon_i(t) = u_i \sin(6\pi t)$ with u_i i.i.d. $N(0,\sigma^2)$. We considered two possibilities: a model without random error, where $\sigma = 0$, and a model with $\sigma = .10$. The effect of the regression function β is, basically, to shift the peak from .4 to .6. Note that (13) induces a one-component model for the $\tilde{y}_i$s in the $\sigma = 0$ case, with $\mu_{\tilde{y}}(t) = \int_0^t \beta(s,t)\mu_{\tilde{x}}(s)ds$ and $\psi_1 = \mu_{\tilde{y}}/\|\mu_{\tilde{y}}\|$; whereas it induces a two-component model in the $\sigma = .10$ case.

Regarding the warping functions, we also considered two situations: data without warping, where $(x_i, y_i) = (\tilde{x}_i, \tilde{y}_i)$, and warped data $(x_i, y_i) = (\tilde{x}_i \circ w_i^{-1}, \tilde{y}_i \circ w_i^{-1})$, with warping functions $w_i(t) = (e^{a_i t} - 1)/(e^{a_i} - 1)$ where the a_is are uniformly distributed in $[-1, 1]$. Ten random pairs (x_i, y_i) of the latter case are shown in Figure 2(a,b) for illustration. Two sample sizes were considered: $n = 50$ and $n = 100$. We will refer to this model as "Model 1".

For this set of simulations we implemented the dynamic functional regression estimator with Hermite splines, using a single knot at $\tau_0 = .5$. The ϕ_ks and ψ_ls were modeled as cubic B-splines with equally spaced knots; two cases were considered: four and nine knots, giving $\nu = 8$ and $\nu = 13$ respectively. The same spline bases were used for the ϕ_ks and ψ_ls of the ordinary least squares estimator. Regarding the choice of dimensions (p, q), we considered four combinations: $(1,1)$, $(1,2)$, $(2,1)$, and $(2,2)$. Given that the true $\tilde{x}_i$s are one-dimensional and the true $\tilde{y}_i$s are either one-dimensional (when $\sigma = 0$) or two-dimensional (when $\sigma = .10$), we expect the optimal estimators to correspond to models $(1,1)$ and $(1,2)$, respectively.

We would also expect the GCV or the AICC criteria to choose these models as optimal, if they were useful for model selection. Tables 1 and 2 report mean integrated absolute errors, $\mathrm{MIAE}(\hat{\beta}) = \mathrm{E}\{\iint |\hat{\beta}(s,t) - \beta(s,t)|\, ds\, dt\}$, based on 300 Monte Carlo replications, for $\sigma = 0$ and $\sigma = .10$ respectively. The MIAEs of the models selected by AICC and GCV were very similar, so we only report the results for AICC. We see that in the absence of warping the dynamic regression estimator is comparable to ordinary least squares, so

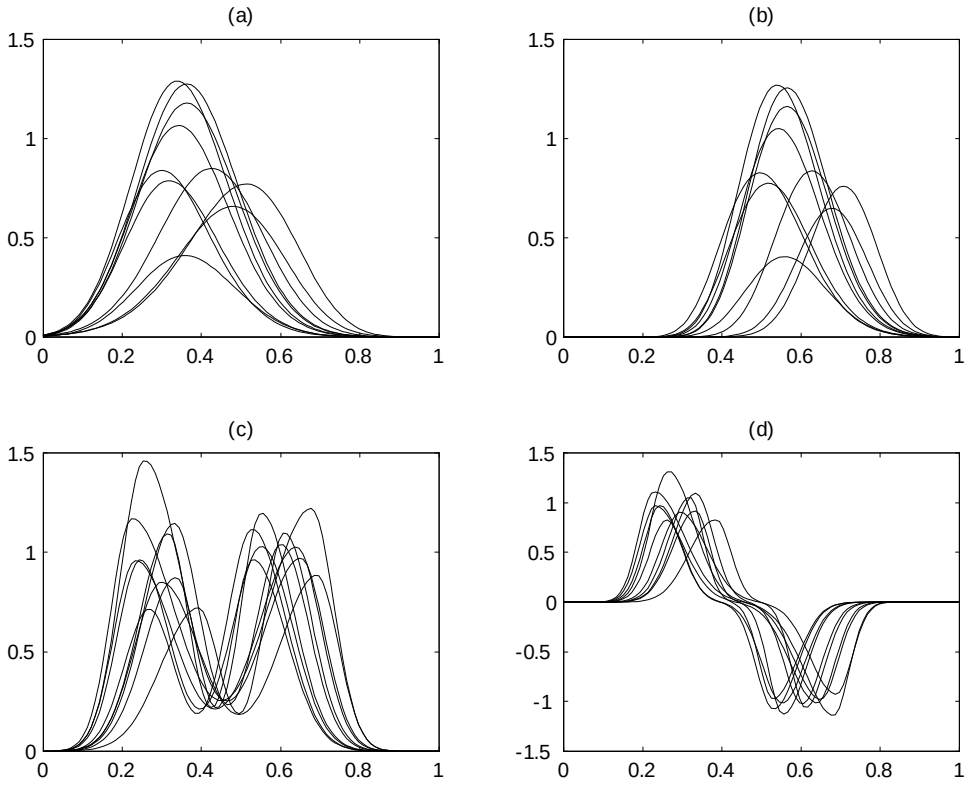

Figure 2: Simulated data. Ten illustrative sample curves (x_i, y_i) for Model 1 [(a) covariates, (b) responses] and Model 2 [(c) covariates, (d) responses].

Model without warping

	4 knots				9 knots			
	$n = 50$		$n = 100$		$n = 50$		$n = 100$	
(p, q)	D	L	D	L	D	L	D	L
$(1, 1)$	.074	.066	.072	.066	.142	.062	.115	.062
$(1, 2)$	.078	.066	.079	.066	.129	.063	.120	.063
$(2, 1)$	.070	.103	.070	.101	.130	.219	.149	.222
$(2, 2)$	.069	.138	.092	.131	.123	.240	.128	.231
AICC	.071	.131	.069	.129	.121	.238	.103	.236

Model with warping

	4 knots				9 knots			
	$n = 50$		$n = 100$		$n = 50$		$n = 100$	
(p, q)	D	L	D	L	D	L	D	L
$(1, 1)$	.085	.650	.082	.656	.134	.593	.132	.598
$(1, 2)$	.082	.647	.080	.653	.144	.595	.143	.600
$(2, 1)$	.148	.669	.175	.672	.136	.612	.177	.614
$(2, 2)$	.158	.522	.317	.522	.209	.506	.287	.505
AICC	.203	.522	.349	.522	.207	.506	.295	.505

Table 1: Simulation results for Model 1, case $\sigma = 0$ (no error term). Mean integrated absolute errors of the slope estimators $\hat{\beta}$ are given, for dynamic regression (D) and ordinary linear regression (L).

Model without warping

	4 knots				9 knots			
	$n = 50$		$n = 100$		$n = 50$		$n = 100$	
(p, q)	D	L	D	L	D	L	D	L
$(1, 1)$	.112	.124	.110	.115	.171	.140	.165	.127
$(1, 2)$	.094	.076	.090	.072	.145	.074	.152	.070
$(2, 1)$	.105	.271	.132	.236	.180	.513	.198	.561
$(2, 2)$	.373	.193	.342	.172	.296	.256	.310	.245
AICC	.333	.083	.337	.072	.276	.076	.317	.069

Model with warping

	4 knots				9 knots			
	$n = 50$		$n = 100$		$n = 50$		$n = 100$	
(p, q)	D	L	D	L	D	L	D	L
$(1, 1)$	.118	.647	.114	.650	.161	.589	.156	.596
$(1, 2)$	.100	.643	.097	.646	.143	.591	.174	.597
$(2, 1)$	.172	.668	.228	.671	.165	.608	.205	.610
$(2, 2)$	.480	.529	.516	.526	.399	.518	.378	.514
AICC	.343	.529	.531	.526	.325	.518	.375	.514

Table 2: Simulation results for Model 1, case $\sigma = .10$. Mean integrated absolute errors of the slope estimators $\hat{\beta}$ are given, for dynamic regression (D) and ordinary linear regression (L).

nothing is lost by using a more complex estimator. But in presence of warping the dynamic estimator is clearly better. We know that warping distorts the principal component estimators and, as a consequence, the ordinary least squares estimator cannot produce a good estimator of β unless too many components are used, and in that case overfitting is a problem. On the other hand, the dynamic estimator successfully recovers the principal components ϕ_k and ψ_l of the $\tilde{x}_i$s and $\tilde{y}_i$s, and therefore it provides an accurate estimator of β, especially for the optimal models $(p, q) = (1, 1)$ and $(p, q) = (1, 2)$. Unfortunately GCV and AICC do not provide very useful guidance for model selection, since the MIAEs of estimators chosen by AICC are often close to or even larger than the MIAEs of estimators corresponding to fixed but overparameterized models (for example, in Table 1, for 9 knots and $n = 100$, the MIAE is $.295$ if the model is chosen by AICC but $.287$ if the model $(2, 2)$ is always used). Then alternative procedures like k-fold cross-validation should be explored. We did not study the performance of k-fold cross-validation by simulation but it did prove useful for the Lip Movement data analysis in Section 4.

We also ran a second set of simulations where we varied the dimension of the warping spaces used for estimation. The data was generated from a more complex model that we will call "Model 2". The warped covariates $\{\tilde{x}_i\}$ followed a two-component model $\tilde{x}_i(s) = \mu_{\tilde{x}}(s) + \sum_{k=1}^{2} z_{ik}\phi_k(s)$ with $\mu_{\tilde{x}}(s) = e^{-100(s-.3)^2} + e^{-100(s-.6)^2}$ and each ϕ_k proportional to a peak (more specifically, if $g_1(s) = e^{-100(s-.3)^2}$ and $g_2(s) = e^{-100(s-.6)^2}$, we took $\phi_1 = g_1/\|g_1\|$ and $\phi_2 = c(g_2 - \langle\phi_1, g_2\rangle\phi_1)$ with c a normalizing constant). The z_{ik}s were i.i.d. $N(0, .07^2)$ and $N(0, .05^2)$, respectively. As regression slope we took $\beta(s, t) = \{\phi_1(s)\psi_1(t) + \phi_2(s)\psi_2(t)\}\mathbb{I}\{s \leq t\}$, with $\psi_1 = \phi_1$ and $\psi_2 = -\phi_2$; the mean of the warped responses $\{\tilde{y}_i\}$ was set as $\mu_{\tilde{y}}(t) = e^{-100(t-.3)^2} - e^{-100(t-.6)^2}$. So the $\tilde{y}_i$s have a peak and a valley; the height of the peak is proportional to the height of the first peak of $\tilde{x}_i$, and the depth of the valley is proportional to the height of the second peak of $\tilde{x}_i$. No random error $\varepsilon_i(t)$ was used for Model 2, since the results for Model 1 were similar for models with or without random error. The pair $(\tilde{x}_i, \tilde{y}_i)$ was then warped with a $w_i(t)$

	$\mathrm{MIAE}(\hat{\beta})$		$\mathrm{MIAE}(\hat{w}) \times 10$	
(p,q,r)	H	MS	H	MS
$(1,1,1)$	.637	.226	.171	.108
$(1,1,2)$	.220	.215	.068	.081
$(2,2,2)$	.150	.138	.066	.079
$(2,2,3)$	.215	.171	.081	.084
$(3,3,3)$	.231	.184	.081	.084

Table 3: Simulation results for Model 2. Mean integrated absolute errors of the slope estimators $\hat{\beta}$ and the warping functions are given, for dynamic regression estimators using Hermite splines (H) or monotone smooth transformations (MS) as warping functions.

that had two independent warping knots, one at each peak. Specifically, we generated $\tau_{i1} \sim U(.2,.4)$ and $\tau_{i2} \sim U(.5,.7)$ independently and constructed a piecewise linear $w_i(t)$ such that $w_i(0) = 0$, $w_i(.3) = \tau_{i1}$, $w_i(.6) = \tau_{i2}$ and $w_i(1) = 1$. A sample of ten pairs (x_i, y_i) is shown in Figure 2(c,d) for illustration. We generated samples of size $n = 50$ and 300 replications were run.

We compared the performance of dynamic functional regression estimators with two different families of warping functions: Hermite splines and smooth monotone functions (Ramsay and Li, 1998). We considered three knot sequences $\boldsymbol{\tau}_0$ of increasing dimensions: $.50$, $(.33,.66)$, and $(.25,.50,.75)$. For Hermite splines we did not penalized the roughness of the w_is, but for smooth monotone functions we did, since the algorithm tended to produce degenerate warping functions otherwise (the smoothing parameter was chosen subjectively and the same value was used in all cases). The ϕ_ks and ψ_ls were modeled as cubic B-splines with nine equally spaced knots, as before. Overall, we considered five combinations (p,q,r): $(1,1,1)$, $(1,1,2)$, $(2,2,2)$, $(2,2,3)$, and $(3,3,3)$; the model closest to the truth is $(2,2,2)$.

In addition to the mean integrated absolute errors of the $\hat{\beta}$s we wanted to assess the warping quality, so we also computed $\mathrm{MIAE}(\hat{w}) = \mathrm{E}\{n^{-1}\sum_{i=1}^{n}\int|\hat{w}_i(t) - w_i(t)|dt\}$. They are shown in Table 3. We see that the optimal model is $(2,2,2)$ as expected, and that monotone smooth transformations generally produce smaller estimation errors for β than Hermite splines, although the latter produce smaller warping errors, probably because the

(p, q)	D	L
$(1, 1)$	.244	.293
$(2, 2)$	.223	.275
$(3, 3)$	.222	.257
$(4, 4)$	.236	.238
$(5, 5)$	.220	.231
$(6, 6)$	—	.232
$(7, 7)$	—	.232

Table 4: Lip Movement Example. Cross-validated mean prediction errors for several models of dynamic functional regression (D) and ordinary linear regression (L).

true warping functions were also splines. Monotone smooth transformations seem to be more robust to misspecification of the warping knots, although this comes at the price of having to select a smoothing parameter.

4 Application: Lip Movement Data

In this section we apply the new estimation method to the data of Malfait and Ramsay (2003). As explained in the Introduction, the goal is to predict lip acceleration (Figure 1(b)) using lip neural activity (Figure 1(a)). This data is hard to analyze for a number of reasons: the curves have sharp peaks and valleys, the first EMG spike occurs very close to the origin, there is substantial phase variability, and there are only 29 sample curves left after removing 3 obvious outliers. We computed dynamic and ordinary retrospective regression estimators with different numbers of components (p, q) and chose the best model by five-fold cross-validation (see Table 4). The principal components were modeled as cubic B-splines with knots at $\{.05, .10, \ldots, .65\}$. As warping functions we used Hermite splines with knots $\boldsymbol{\tau}_0 = (.08, .2, .4, .5)$, which approximately correspond to the average location of the EMG peaks.

According to Table 4, the dynamic regression estimator with smallest cross-validated error corresponds to a five-component model, but we choose the three-component model since it attains a comparable error with fewer components; while the best ordinary lin-

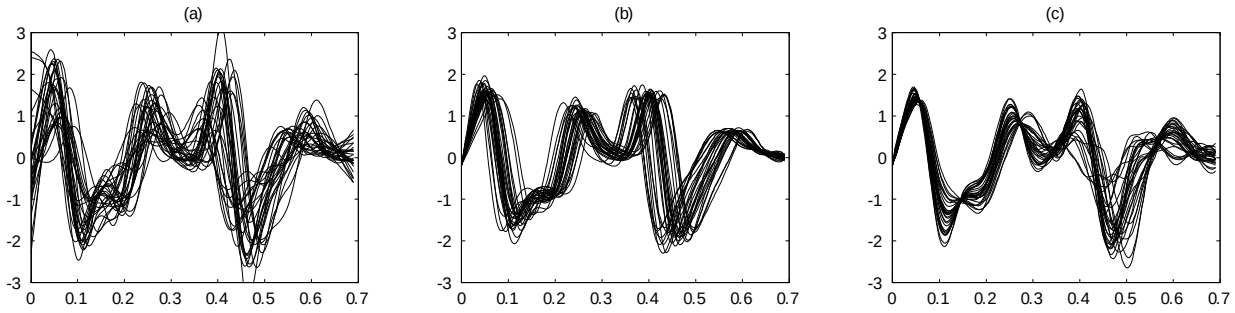

Figure 3: Lip Movement Example. (a) Response curves; (b) fitted curves obtained by dynamic regression; (c) fitted curves obtained by ordinary linear regression.

ear regression estimator is unequivocally given by a five-component model. Figure 1(c,d) shows the warped sample curves. We see that dynamic regression does a good job at synchronizing the curves. The features of both explanatory and response curves emerge very clearly. In particular, the peaks of the EMG curves around $t = .4$ and $t = .5$, which were barely discernible in Figure 1(a), are plain to see in Figure 1(c). These peaks correspond to the agonistic and antagonistic actions of the lower-lip muscle at the beginning and the end of the second 'b' in 'Bob'.

Figure 3 shows the lip acceleration curves $\{y_i\}$ together with the fitted curves $\{\hat{y}_i\}$. We see that ordinary least squares produces a substantially worse fit; the mean prediction error of dynamic regression is .0898 while the mean prediction error of ordinary regression is .1648, almost twice as large. Even though ordinary regression uses two more principal components than dynamic regression to estimate β, it is clear that these extra components cannot make up for the lack of a time-warping mechanism, and adding more components actually makes prediction worse, as Table 4 shows. So this is a situation where the data clearly calls for a model that includes a time-warping mechanism, and dynamic regression then represents a substantial improvement over ordinary linear regression.

Interpreting $\hat{\beta}(s,t)$ is harder but also interesting. To determine which features of the $\hat{\beta}$s are actually statistically significant, we estimated the variance of $\hat{\beta}(s,t)$, $\hat{v}(s,t)$, by bootstrap (using residual resampling). Contour plots of the filtered estimators $\hat{\beta}(s,t)\mathbb{I}\{|\hat{\beta}(s,t)| \geq$

$2\sqrt{\hat{v}(s,t)}\}$ are shown in Figure 4. The dynamic regression estimator shows significant features outside the diagonal, implying that lip acceleration can be predicted not only by neural activity immediately preceding the event, but also by neural activity further in the past. For example, consider predicting the sharp deceleration of the $\tilde{y}_i$s at $t = .45$, which is given by $\mu_{\tilde{y}}(.45) + \int_0^{.45} \beta(s, .45)\{\tilde{x}_i(s) - \mu_{\tilde{x}}(s)\}ds$. In Figure 4(a) we see that $\beta(s, .45)$ not only has a peak near the diagonal, which is unsurprising because it corresponds to the immediately preceding neural activity, but also at $s = .1$ (where the valley between the first two peaks of the $\tilde{x}_i$s occur), and troughs before and after that peak (where the first two peaks of the $\tilde{x}_i$s occur). This implies that if the first two spikes of $\tilde{x}_i$ (related to the *first* 'b') are sharper than the mean, the integral $\int_0^{.45} \beta(s, .45)\{\tilde{x}_i(s) - \mu_{\tilde{x}}(s)\}ds$ will tend to be negative and then $\tilde{y}_i(.45)$, the deceleration of the lips at the *second* 'b', will tend to be stronger than the mean. Off-diagonal features of the ordinary least squares estimator can also be seen in Figure 4(b), and were also observed by Malfait and Ramsay (2003) using a different approach to ordinary least squares (based on a triangular-basis expansion for β rather than on a tensor-product principal-component expansion), but they are harder to interpret because they are applied to non-synchronized curves.

More information about the dynamics of the process can be extracted from the warping functions themselves. The use of interpolating Hermite splines facilitates this, because the estimated parameters $\hat{\boldsymbol{\tau}}_i$ roughly correspond to the locations of the landmarks $\boldsymbol{\tau}_0$ on the respective sample curve. For our choice of $\boldsymbol{\tau}_0$, the $\hat{\boldsymbol{\tau}}_i$s will roughly correspond to the location of the four characteristic peaks of the EMG curves. Thus $d_{i1} = \hat{\tau}_{i2} - \hat{\tau}_{i1}$ indicates the duration of the first 'b', $d_{i2} = \hat{\tau}_{i3} - \hat{\tau}_{i2}$ the duration of the 'o', and $d_{i3} = \hat{\tau}_{i4} - \hat{\tau}_{i3}$ the duration of the second 'b'. The pairwise correlations of the ds are $\rho_{12} = -.46$, $\rho_{13} = .66$ and $\rho_{23} = -.24$, indicating that there is a significant negative correlation between the duration of the first 'b' and the 'o', and a significant positive correlation between the durations of the two 'b's. More accurate information about the phonemes' duration could be obtained by estimating the exact peak locations curve by curve, but that would be

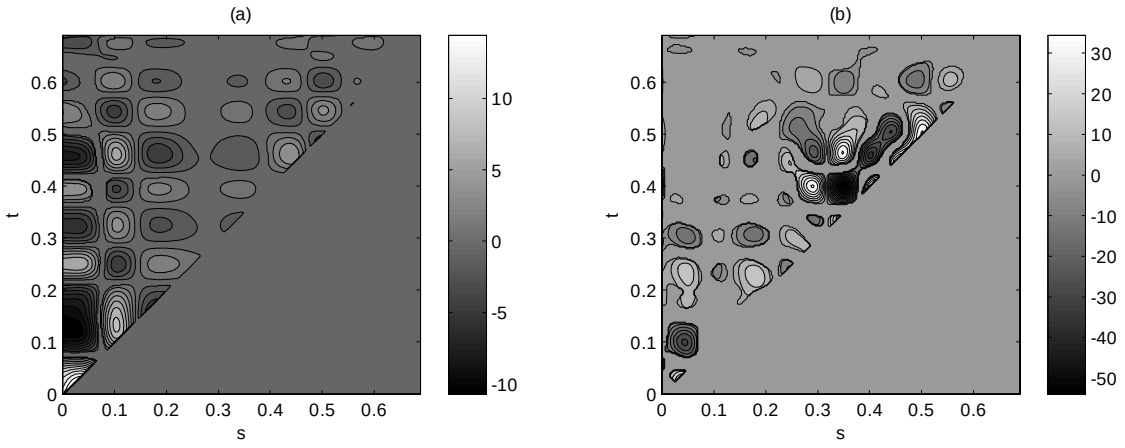

Figure 4: Lip Movement Example. Contour plots of estimated slope functions $\hat{\beta}(s,t)$ [(a) dynamic regression estimator, (b) ordinary least squares estimator].

unfeasible for larger datasets. An advantage of Hermite-spline warping is that the $\boldsymbol{\tau}_i$s are estimated automatically as a by-product of the procedure.

Supplementary material

Technical supplement: Contains technical details of implementation, such as derivatives of objective functions and constraints. (PDF file)

Matlab package DRFR: Matlab package "DRFR" containing code to compute the estimators introduced in this article and the Lip Movement data. (ZIP file)

Acknowledgements

This research was partially supported by NSF grant DMS 10-06281.

References

Antoch, J., Prchal, L., De Rosa, M.R. and Sarda, P. (2008) Functional linear regression with functional response: application to prediction of electricity consumption. In *Functional and Operational Statistics*, pp. 23–29, eds. S. Dabo-Niang and F. Ferraty. Heidelberg: Physica-Verlag.

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

Cai, T. and Hall, P. (2006). Prediction in functional linear regression. *The Annals of Statistics* **34** 2159–2179.

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

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

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

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

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

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

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Second Edition.* Springer, New York.

Hurvich, C.M., Simonoff, J.S., and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society (Series B)* **60** 271–293.

James, G., Wang, J. and Zhu, J. (2009). Functional linear regression that's interpretable. *The Annals of Statistics* **37** 2083–2108.

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

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. and Ramsay, J.O. (2000). Curve registration by local regression. *Canadian Journal of Statistics* **28** 19–30.

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

Liang, H., Wu, H., and Carroll, R. J. (2003). The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effects varying-coefficient models with measurement error. *Biostatistics* **4** 297–312.

Liu, X. and Müller, H.-G. (2004). Functional convex averaging and synchronization for time-warped random curves. *Journal of the American Statistical Association* **99** 687–699.

Malfait, N. and Ramsay, J.O. (2003). The historical functional linear model. *Canadian Journal of Statistics* **31** 115–128.

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

Nocedal, J. and Wright, S.J. (2006). *Numerical Optimization. Second Edition*. Springer, New York.

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

Ramsay, J.O. and Silverman, B. (2005). *Functional Data Analysis. Second Edition.* Springer, New York.

Silverman, B. (1996). Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics* **24** 1–24.

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

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

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

Wahba, G. (1990). *Spline Models for Observational Data.* CBMS-NSF regional conference series in applied mathematics, SIAM, Philadelphia.

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

Wu, H. and Liang, H. (2004). Backfitting random varying-coefficient models with time-dependent smoothing covariates. *Scandinavian Journal of Statistics* **31** 3–19.

Wu, S. and Müller, H.-G. (2011). Response-additive regression for longitudinal data. *Biometrics* **67** 852–860.