The Annals of Statistics
1997, Vol. 25, No. 3, 1251–1276

ALIGNMENT OF CURVES BY DYNAMIC TIME WARPING[1]

BY KONGMING WANG and THEO GASSER

University of Zürich

When studying some process or development in different subjects or units—be it biological, chemical or physical—we usually see a typical pattern, common to all curves. Yet there is variation both in amplitude and dynamics between curves. Following some ideas of structural analysis introduced by Kneip and Gasser, we study a method—dynamic time warping with a proper cost function—for estimating the shift or warping function from one curve to another to align the two functions. For some models this method can identify the true shift functions if the data are noise free. Noisy data are smoothed by a nonparametric function estimate such as a kernel estimate. It is shown that the proposed estimator is asymptotically normal and converges to the true shift function as the sample size per subject goes to infinity. Some simulation results are presented to illustrate the performance of this method.

1. Introduction. When studying some process or development (e.g., a biological, chemical or physical process), we usually see a typical pattern which is common to different subjects or units, and yet there are variations in both amplitude and phase (or timing) between curves. One example is growth of humans or animals, where growth evolves at different intensities and at different paces in different individuals. Another example is speech signals, where the same words are spoken with varying loudness and varying speed. Classical statistical approaches such as repeated measure analysis of variance or principal component analysis deal exclusively with amplitude variation, and methods to deal with both are scarce. In Kneip and Gasser (1992), "structural analysis" was proposed to align or shift curves to a common average time scale before applying further statistics such as averaging curves. Estimating "structural average curves" proved to be successful to study human growth for variables which have been inaccessible because of their small size and relatively large residual variation [Gasser, Kneip, Binding, Prader and Molinari (1991) and Gasser, Kneip, Zieger, Molinari, Prader and Largo (1994)]. However, the step leading to individual shift functions for the alignment of curves is somewhat delicate and time-consuming.

In the engineering literature, a different approach, called dynamic time warping, was developed to align two signals with different dynamics [Parsons (1986), Rabiner and Schmidt (1980), Qi (1992)]. This method has been mainly

Received May 1995; revised August 1996.
[1]Research supported by Swiss NSF (21-36042.92).
AMS 1991 *subject classifications*. Primary 62G07; secondary 62H05.
Key words and phrases. Curves, shift functions, structural analysis, dynamic time warping, kernel estimation.

applied to speech analysis and speech recognition. We will give details in the next section and sketch here how time warping works. Suppose that two sequences $\{f(i),\ i = 1, \ldots, M\}$ and $\{g(j),\ j = 1, \ldots, N\}$ characterize two signals f and g, respectively. We want to find the best match between f and g by some alignment w, based on minimizing a cost function. The classical cost function is given by

$$\inf_{w} \sum_{(i,j)\in w} \left(f(i) - g(j)\right)^2.$$

Here $w = \{(i, j)\}$ is a warping path connecting $(1, 1)$ and (M, N) in a two-dimensional square lattice and satisfying monotonicity and connectedness. This means that both coordinates of the parametrized path $w = \{(i(k),\ j(k))$: $k = 1, \ldots, K;\ i(1) = j(1) = 1,\ i(K) = M,\ j(K) = N\}$ have to be nondecreasing, and that they can only increase by 0 or 1 when going from k to $k + 1$. Obviously this is a minimization problem which can be solved efficiently by using dynamic programming. What controls the warping result is the cost function. How a sample of functions can be aligned by dynamic time warping to some average time scale, however, needs some further thoughts.

We now describe briefly the methods proposed in the statistical literature to analyze samples of functions. Suppose the observed data y_{ij} fit the model

$$\begin{aligned} &y_{ij} = f_i(t_{ij}) + \varepsilon_{ij}, \\ &\quad j = 1, \ldots, n_i \text{ (timing for subject } i); \ i = 1, \ldots, m \text{ (subjects)}. \end{aligned} \tag{1}$$

Here $\{f_i\colon i = 1, \ldots, m\}$ are unknown smooth functions, $t_{ij} \in J \equiv [0, 1] \subset \mathbf{R}$ (J could be any closed interval), and $\{\varepsilon_{ij}\colon j = 1, \ldots, n_i;\ i = 1, \ldots, m\}$ are independent random variables with mean $E(\varepsilon_{ij}) = 0$ and variance $V(\varepsilon_{ij}) = \sigma_i^2 > 0$. Note that the $\{t_{ij}\}$ do not need to be equally spaced.

When all subjects have the same number of measurements ($n = n_i$), the resulting multivariate data matrix can be analyzed by principal component analysis, as suggested by Rao (1958). In this way, m functions are reduced to a small number of "elementary" functions. This approach does not take into account the inherent smoothness of the functions f_i; and, more crucially, it does not account for "dynamic variability" but only for amplitude variability. The first drawback was eliminated in the proposal by Rice and Silverman (1991), incorporating a penalized smoothing approach into PCA [see also the related paper by Ramsay and Dalzell (1991)]. To account for the second drawback, Silverman (1995) also incorporated an individually constant shift parameter into PCA.

The cross-sectional average $\sum_{i=1}^m f_i/m$ is not a good estimate of the average dynamic and amplitude in general. Let us assume that $f_i(t) = s(u_i(t))$ in (1), where s is some shape function and $\{u_i\}$ is an i.i.d. random sample of shift functions satisfying $E(u_i(t)) = t$ and strict monotonicity. If s is nonlinear, then $E(f_1) \neq s$. By the law of large numbers, $m^{-1}\sum_{i=1}^m f_i \to E(f_1) \neq s$ in probability.

The structural analysis suggested by Kneip and Gasser (1992), leading in particular to a structural average curve, proceeds as follows.

1. Kernel estimates $\hat{f}_i$, $\hat{f}_i^{(1)}$, $\hat{f}_i^{(2)}$ of f_i, f_1', f_i'' are obtained.
2. Individual structural points are identified from $\hat{f}_i$, $\hat{f}_i^{(1)}$ and/or $\hat{f}_i^{(2)}$. Roughly speaking, structural points (called "landmarks" in shape analysis) are features that are common to all or most curves.
3. Shift functions $\hat{u}_i$ are constructed from the locations of the structural points such that individual structural points are shifted to the respective average location. In between, smooth monotonic interpolation is used.
4. A structural average $\hat{f}_0$ (in contrast to a cross-sectional average) is obtained by averaging aligned (smoothed) curves:

$$\hat{f}_0(\cdot) = \frac{1}{m} \sum_{i=1}^{m} \hat{f}_i(\hat{u}_i(\cdot)). \tag{2}$$

Figure 1 shows the alignment of the velocities of two growth curves of shoulder width (boys) by steps (2)–(3) and by dynamic time warping. The structural points used in steps (2)–(3) are simply extremes and inflection points. Dynamic time warping is introduced in the next section. Both methods produce rather good results in this example.

A delicate step is (2), and to a lesser extent (3); it is not easy to define features which are common to most curves and to determine them unequivocally from noisy data, such that they have an equivalent meaning. Dynamic time warping, which addresses a similar problem, does not need such prerequisites. The method is fully nonparametric. It is thus of interest to establish whether and when dynamic time warping leads to meaningful shift functions. This problem is dealt with here mainly in the context of aligning one function with respect to another. The possible extension to m functions is discussed but not treated exhaustively.

Section 2 is devoted to an introduction into dynamic time warping and to some improvements. First, a variational problem in continuous time is formulated in order to obtain smooth shift functions instead of a warping path. Secondly, new cost functions are introduced to offer an improvement. At the end of Section 2, we present approaches for aligning m curves to their common time scale so that a structural average curve can be computed. A theoretical analysis of this method is beyond the scope of this paper. In Section 3, various classes of models are introduced and it is shown that dynamic time warping identifies the correct shift functions in these classes. Asymptotic properties for the estimators of shift functions are derived in Section 4 (with proofs in the Appendix). Simulations with a small number of replications are presented in Section 5.

2. Dynamic time warping. Dynamic time warping has been designed for aligning one curve with respect to another, and it is well documented in the engineering literature. The article by Sakoe and Chiba (1978) is a good

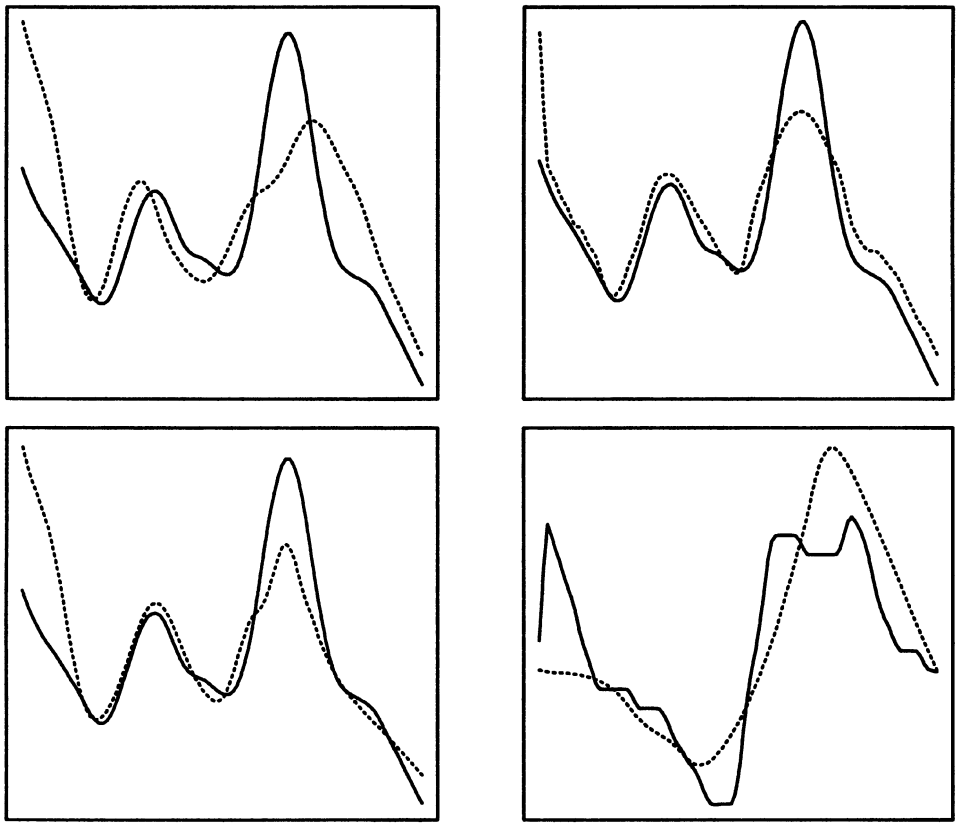

FIG. 1. *Example of aligning two curves*: *top left gives two smoothed velocity curves*; *top right is the alignment by dynamic time warping*; *bottom left is the alignment by steps* (2)–(3); *bottom right gives differences of shift functions and the identity function. The shift functions are produced by dynamic time warping* (*solid line*) *and by steps* (2)–(3) (*dotted*).

reference for basic ideas, and we present this approach in Section 2.1. A new cost function is introduced there. In Section 2.2, we deal with the problem of aligning m regression functions.

2.1. *Aligning two regression functions.* In speech recognition, a word or a sentence can be expressed as a sequence of features by feature extraction methods. This feature vector, rather than the original audio signal, is then further analyzed. The recognition process consists of comparing the recorded word with words in a template set. If its feature sequence matches closely to the feature sequence of a word in the template set, then the word (or speech) is recognized. For this comparison, the time-axis fluctuations between the given word and a template have to be eliminated. The template for a word might be what we call a structural average here and it is obtained by averaging samples spoken by many people.

We describe briefly the version of dynamic time warping given in Sakoe and Chiba (1978). Let $\mathscr{F} = (f(1), \ldots, f(M))$ and $\mathscr{G} = (g(1), \ldots, g(N))$ be two feature vectors. Whether the two speech patterns are of the same category is similar to asking whether there is a mapping w of the form

$$w = ((i(1), j(1)), \ldots, (i(K), j(K))) \tag{3}$$

such that the discrepancy

$$C_0(\mathscr{F}, \mathscr{G}, w) \equiv \sum_{k=1}^{K} d(f(i(k)), g(j(k)))r(k) \tag{4}$$

is small enough. The warping path w has to satisfy several side conditions.

CONDITIONS. (i) Monotoncity: $i(k) \leq i(k+1)$ and $j(k) \leq j(k+1)$.
(ii) Continuity: $i(k+1) - i(k) \leq 1$ and $j(k+1) - j(k) \leq 1$.
(iii) Boundary: $i(1) = j(1) = 1$, $i(K) = M$ and $j(K) = N$.
(iv) Window: $|i(k) - j(k)| \leq$ a given positive integer.
(v) Slope contraint: neither too steep nor too gentle a gradient should be allowed.

In (4), the function $d(\cdot, \cdot)$ is a distance measure and r is a nonnegative weighting function [usually one takes $r(k) \equiv 1$]. The length K of the warping path w is determined by the warping process. Note that this cost function is symmetric in $\mathscr{F}$ and $\mathscr{G}$.

The time-normalized distance between the speech patterns $\mathscr{F}$ and G is defined as the solution of the following minimization problem:

$$D_0(\mathscr{F}, \mathscr{G}) = \inf_w C_0(\mathscr{F}, \mathscr{G}, w). \tag{5}$$

Conditions (i) and (ii) are natural. Condition (iii) is posted since the start and end of words are detected before time normalization. Condition (iv) is posted according to the concept that time-axis fluctuations should not lead to too excessive differences in timing. Finally (v) is used to prevent unrealistic warping if $|M - N|$ is relatively large. Figure 2 shows what time warping does.

In principle a cost function could be any functional of the two input sequences and a warping path, depending on the purpose of the application. The cost functions used in speech recognition and pattern classification are varieties of the classical quadratic cost function (see Table 1 for some examples).

Some explanations might be helpful. In the second line of the table, $P(w)$ is a penalty function which penalizes jumps and flat spots on the warping path w, and β is a coefficient which controls how severe the penalty is [$\beta = 0.0075$ in Roberts, Lawrence, Eisen and Hoirch (1987), chosen by trial

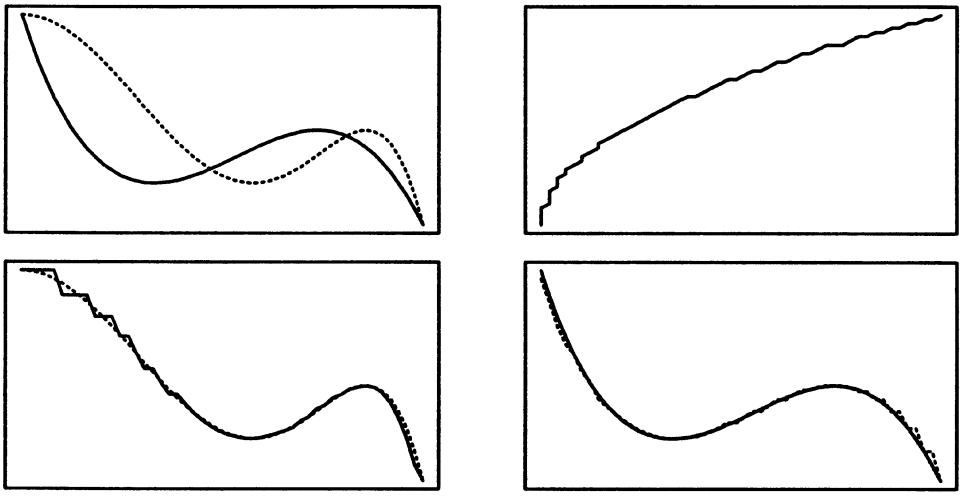

FIG. 2. *Tutorial illustration of dynamic time warping. Top left: two given curves; top right: warping function; bottom figures: the warping of one curve to the other.*

and error]. From the equality $ab = -(a-b)^2/2 + (a^2+b^2)/2$, we see that

$$-\sum_{(i,j)\in\omega} f(i)g(j) = \tfrac{1}{2}\sum_{(i,j)\in w}\big(f(i)-g(j)\big)^2 - \tfrac{1}{2}\sum_{(i,j)\in w}\big(f(i)^2+g(j)^2\big).$$

The first term is just the classical cost function. Since the second term tends to make the warping path longer, a penalty is needed. A version of $P(w)$ given in Roberts, Lawrence, Eisen and Hoirch (1987) is the following. Write $w = \{(i(k), j(k)): k = 1, \ldots, K\}$ where K is the length of w. Then

$$P(w) = \sum_{k=2}^{K} p(w,k),$$

TABLE 1
Cost functions

$\sum_{(i,j)\in w}(f(i)-g(j))^2$	Classical
$-\left[\sum_{(i,j)\in w} f(i)g(j) - \beta P(w)\right]$	Roberts, Lawrence, Eisen and Hoirch
$\int_0^1\left[\alpha^2\left(\frac{f(t)}{\|f\|}-\frac{g(u(t))}{\|g\|}\right)^2 + (1-\alpha)^2\left(\frac{f'(t)}{\|f'\|}-\frac{g'(u(t))}{\|g'\|}\right)^2 + \phi(u'(t))\right]dt$	Proposed here

where

$$p(w,k)=\begin{cases}(k-r_k-1)^2,\\ \qquad \text{if } i(k)=i(k-1),\ r_k=\min\{r\le k-1\colon i(r)=i(k)\},\\ (k-r_k-1)^2,\\ \qquad \text{if } j(k)=j(k-1),\ r_k=\min\{r\le k-1\colon j(r)=j(k)\},\\ 0,\quad \text{otherwise.}\end{cases}$$

With the penalty term added, the cost function is no longer a convex function.

The third line of Table 1 is the cost function proposed in this paper. Details are given below. Our cost function is inspired by Sobolev norms and by the least squares principle. The normalization with the sup-norm $\|\cdot\|$ in (6) is intended to reduce the differences in amplitudes of the curves when estimating the shift functions. This should prevent us from explaining amplitude variability between curves in terms of dynamic variability. There are other cost functions used in speech analysis. For example, a cost function defined in terms of the linear predictive coding features sets of signals is used in Höhne, Coker, Levinson and Rabiner (1983).

For our purpose, where aligning maxima, minima, and inflection points of curves is important, we incorporate derivatives of functions into the cost function. Apart from heuristics, theoretical and simulation analysis lends support to this idea. Incorporating higher order derivatives (the second derivative in particular) is possible in principle, but problems of estimating higher order derivatives from noisy data might arise.

Now, we give details of the new cost function. Define a functional F of the functions f, f', g, g', u and a real variable α by

$$\begin{aligned}F(f,f',g,g',u,\alpha)(t)\equiv{}&\alpha^2\left(\frac{f(t)}{\|f\|}-\frac{g(u(t))}{\|g\|}\right)^2\\ &+(1-\alpha)^2\left(\frac{f'(t)}{\|f'\|}-\frac{g'(u(t))}{\|g'\|}\right)^2.\end{aligned}\tag{6}$$

Here f' is the derivative of f and $\|f'\|$ is the supremum norm of f'. Then the cost function is defined by

$$C(f,f',g,g',u,\alpha)\equiv\int_0^1\left[F(f,f',g,g',u,\alpha)(t)+\phi(u'(t))\right]dt.\tag{7}$$

The function ϕ serves as a penalty function which plays a role similar to the side conditions (i)–(v). It is specified as follows. Let $M>\delta>0$ be constants and define ϕ to be a convex function satisfying the following conditions: $\phi(x)=0$ for $x\in[\delta+r,M-r]$ with a small positive number r; $\phi(\delta^+)=\phi(M^-)=\infty$; $\phi(x)=\infty$ for $x\in(\delta,M)^c$; and $\phi\in C^4(\delta,M)$. An example of

such a ϕ is given by

$$\phi(t) = c\Big[(\delta + r - t)^5 I_{(\delta,\,\delta+r]}(t)/(t-\delta) + (t - M + r)^5 I_{[M-r,\,M)}(t)/(M-t)\Big], \qquad t \in (\delta, M)$$

for a constant c. Here $I_A(t)$ is the indicator function of a set A.

The best warping or shift function between two functions f and g is given by the solution of the following variational problem

$$\inf\{C(f, f', g, g', u, \alpha) : u \in C^1,\ \alpha \in \mathbf{R}\}. \tag{8}$$

The cost function (7) is motivated not only by the least squares principle. For two important classes of models, true shift functions can be recovered by using the cost function; compare Section 3. Simulations in Section 5 also show its usefulness.

It is easy to verify via Euler's equation that if the optimal solution u of (8) satisfies $\delta + r < u'(t) < M - r$, $t \in (0, 1)$, then any extrema of g is aligned to a stationary point of f where $f' = 0$.

Note that we do not require $u(0) = 0$. This flexibility allows us to study certain models in more detail (Section 3). In applications one usually has $u(0) = 0$. This is the case in speech recognition, where the start and end points of words or sentences are detected before time warping. It is also the case when growth curves are analyzed. The case $u(0) \neq 0$ arises, for example, when one signal has not been observed from the beginning.

The warping path in a discrete setting is in parametrized curve form $\{(i(k), j(k)) : k = 1, \ldots, K\}$, while the warping function u in our continuous setting is not. The reason is that the warping path in a discrete setting is not a one-to-one mapping. Note that the parametric form makes the variational problem symmetric in the two functions being matched. Since we require that the shift function is in C^1 and that the optimal shift function is strictly increasing, no parametrization in curve form is needed.

The variational problem (8) can be solved as follows. For a given α, use dynamic programming to find an optimal u corresponding to this α. Then the minimization in α can be restricted to [0, 1] (see the proof of Lemma 4.1) and can be done, say, by grid search.

2.2. *Aligning m regression functions.* We now go back to model (1) of m regression functions and present a method for aligning all curves to their average time scale based on aligning one function to another. Once this is done, further statistical analysis such as structural averaging is then straightforward.

Dynamic time warping produces a relative shift function between two curves. To align a sample of curves to a common time scale, we need a reference curve. Then all curves can be aligned to this reference curve, and hence the average timing can be computed. It is assumed in this subsection that all curves are observed continuously and are noise free. In applications a further preliminary step of smoothing the data is required.

The principle is simple. Let f_e be the chosen reference curve. Warp each curve f_i, $i = 1, \ldots, m$, to f_e and denote the warping function by $h_i(t)$, $t \in [0, 1]$. Then

$$h(t) \equiv \frac{1}{m} \sum_{i=1}^{m} h_i(t)$$

is the average timing with respect to f_e. Since each h_i is strictly increasing, the function h is strictly increasing and it has an inverse h^{-1}. Now it is clear that

$$u_i(t) \equiv h_i\big(h^{-1}(t)\big)$$

is the correct shift function to transform f_i to the average time scale. A structural average of $\{f_i\colon i = 1, \ldots, m\}$ is then computed as

$$f_0(t) = \frac{1}{m} \sum_{i=1}^{m} f_i\big(u_i(t)\big).$$

The reference curve should be close to the typical pattern of the sample curves and should have more or less the same features as most sample curves. A consideration in choosing a reference curve is the trade-off between accuracy and computational effort. Several possibilities are given here.

1. In principle one could choose a curve randomly from the sample as reference curve, following the arguments given above. This is computationally attractive even when m is large, but the statistical quality may suffer if an atypical curve were selected. This would inevitably make it more difficult to estimate the warping function well.
2. Take each f_i, $i = 1, \ldots, m$, as reference curve. Warp every other curve to f_i and compute the total cost (i.e., the sum of the cost for warping f_j to f_i, $j \neq i$). Now choose f_e to be the curve corresponding to the maximum total cost. The main problem with this procedure is that it can require prohibitive computing time if m is large. Note that one would need to solve the variational problem (8) $m(m-1)/2$ times.
3. An iterative method can be used. First take $f_e(t) = (1/m)\sum_{i=1}^{m} f_i(t)$, the cross-sectional average. Then compute a structural average based on f_e. In the following steps, take the structural average computed in the previous step as the reference curve of the next step and iterate. Computation is not a problem since a few iterations are enough. This proposal shows good statistical properties if the relative shifts among curves are small. If the shifts are large, then the cross-sectional average might be too atypical to start with, since structure gets lost.
4. For large m, one could select a random sample of size k from $\{f_i\}$. Assume $k = 2^j$. Partition this selected sample into 2^{j-1} pairs and compute a structural average for each pair by a single warping. Now we have 2^{j-1} structural averages. Partition this group into 2^{j-2} pairs and compute a structural average for each pair. Repeat this procedure till only one structural average is left. Take this one as the reference curve f_e.

Now the question arises as to which of proposals (1)–(4) should be used in practice. Our suggestion is the following. If m is small, (2) should be the choice. If m is large but one has confidence that the relative shifts among sample curves are small and that no outlier should be in the sample, (3) would be a good choice. Finally, if m is large and no prior information about the quality of the sample is available, (4) is recommended. Another possibility would be to combine (2) and (4): select a random sample of size k from $\{f_i\}$ and perform (2) on this subsample to compute a reference curve. As mentioned before, analysis of the methods proposed in this subsection is not considered in this paper.

3. Alignment for some semiparametric models. It has become clear that dynamic time warping effects a nonparametric technique for the alignment of regression functions. The questions is now whether dynamic time warping achieves its goal when some parametric or semiparametric model is assumed to be known. In the following we study some semiparametric models. Most of these models have been studied in recent years in the statistical literature. We want to investigate whether dynamic time warping identifies the right alignment in the absence of noise. Let us postulate a functional model of the following form for the regression model (1):

$$f_i(t) = s(t, \theta_i), \qquad t \in [0,1],\ i = 1, \ldots, m. \tag{9}$$

Here s is some prespecified function with individual parameters $\theta_i \in \mathbf{R}^d$. When data for many functions are available, and when some general structure for s can be postulated, the semiparametric problem of estimating θ_i in the presence of the infinite-dimensional nuisance parameter s can be successfully treated [Kneip and Gasser (1988)], and no specific form for s needs to be specified. In the context of this paper, it is interesting that most of the semiparametric classes of models considered so far have amplitude and shift variation as the basic structure. In the simplest case this variation is modeled linearly, leading to the so-called shape-invariant model (SIM):

$$f_i(t) = a_i s\left(\frac{t - b_i}{c_i}\right) + d_i. \tag{10}$$

Estimating parameters a_i, b_i, c_i, d_i and the shape function s for such a model has been studied in Lawton, Sylvestre and Maggo (1972), Kneip and Gasser (1988), and Kneip and Engel (1995). This class contains logistic, Gompertz and other important nonlinear regression models. It is successful for modeling human growth [Stützle, Gasser, Molinari, Largo, Prader and Huber (1980)] by postulating two components for prepubertal and pubertal growth, respectively.

It is easy to see that the optimal shift function between f_i and f_j is given by the solution (u, α) of (8), where

$$(u(t), \alpha) = \big(c_j(t - b_i)/c_i + b_j, 0\big).$$

Thus, the newly introduced cost function is able to identify linear shifts within SIM correctly. To appreciate this, note that using other cost functions in Table 1 will not lead to correct shift functions within this simple model.

For comparing two functions f_1 and f_2, Härdle and Marron (1990) considered a somewhat more general model with amplitude-phase modulation:

$$f_2(t) = S_{\theta_0}^{-1} f_1\left(T_{\theta_0}^{-1} t\right), \tag{11}$$

where S_θ and T_θ are invertible parametric transformations. For linear transformations S_θ and T_θ this reduces to (10). They proposed to estimate θ_0 by minimizing the loss function

$$L(\theta) = \int \left[f_1(t) - S_\theta f_2(T_\theta t) \right]^2 r(t)\, dt,$$

with some nonnegative weight function r. None of the cost functions in Table 1 is successful for identifying T_{θ_0} in general, but a back-fitting method [such as the one proposed in Kneip and Gasser (1988)] would work.

A more general nonlinear shift model (NLSM) allows a shift function u_i to be specified nonparametrically:

$$f_i(t) = a_i s(u_i(t)), \tag{12}$$

where $u_i \in C^1$ is strictly increasing and a_i is a real and positive parameter. Despite the great generality allowed for shifts, this model is still identifiable. The optimal shift function from f_j to f_i can be obtained by dynamic time warping with cost function (7) as

$$(u(t), \alpha) = \left(u_j^{-1}(u_i(t)), 1\right).$$

Thus, shift functions can only be extracted in relative terms with respect to some function chosen as reference (f_i in this example). Again, other cost functions in Table 1 are not successful for extracting the correct shift function.

An interesting generalization emerges when replacing the individual factor α_i by some parametric function $a_i(t, \beta_i)$:

$$f_i(t) = a_i(t, \beta_i) s(u_i(t)) + d_i. \tag{13}$$

Here α_i is an a priori known function with individual parameter $\beta_i \in \mathbf{R}^d$, and d_i is an unknown constant. Possibly, this quite general semiparametric model is also identifiable when requiring proper conditions for β_i, g_i, and d_i. In any case, recovering the shift function between two such functions would require a more complicated cost function than (7). It is plausible that a back-fitting procedure—such as the one used in Kneip and Gasser (1988)—could solve this problem. However, the general amplitude-phase modulated model

$$f_i(t) = a_i(t) s(u_i(t))$$

is clearly not identifiable. This is true even when requiring obvious conditions such as $Ea_i(t) \equiv 1$ and $Eg_i(t) \equiv t$. Nonetheless, simulations show that cost function (7) yields reasonable results even for this model.

With the new cost function (7), shift functions can be fully recovered by dynamic time warping in models (SIM) and (NLSM) if data are noise free. This seems to us an important achievement, since dynamic time warping is a relatively easy, automatic method. It can be attributed to the inclusion of derivatives in the cost function. For noisy data, some nonparametric function fitting method like kernel estimators or local polynomial fitting allows the estimation of the function itself and of its derivatives as a preliminary step. The problem of estimating derivatives from noisy data might have prevented their earlier use in a cost function.

4. Estimation and asymptotics. In practice, a regression function f is unobservable and has to be estimated from noisy data $\{y_1, \ldots, y_n\}$, where $y_j = f(t_j) + \varepsilon_j$. We use convolution-type kernel smoothing to estimate derivatives of order $\nu \geq 0$ of f. Specifically, let K_ν be a kernal of order $(\nu, \nu + 2)$ [Gasser, Müller and Mammitzsch (1985)] for $\nu = 0, 1$. That is,

$$\int_{-1}^{1} K_\nu(t) t^j \, dt = \begin{cases} 0, & j \leq \nu \text{ and } 0 \leq j \leq \nu + 1, \\ (-1)^\nu, & j = \nu, \\ \beta_j \neq 0 & j = \nu + 2, \end{cases}$$

and the support of K_ν is $[-1, 1]$. Note that optimal kernels are explicitly known as polynomials of order $(\nu + 2)$. Then we define

$$\hat{f}(t) = \frac{1}{b_0} \sum_{j=1}^{n} y_j \int_{s_{j-1}}^{s_j} K_0\left(\frac{v - t}{b_0}\right) dv,$$

$$\hat{f}'(t) = \frac{1}{b_1^2} \sum_{j=1}^{n} y_j \int_{s_{j-1}}^{s_j} K_1\left(\frac{v - t}{b_1}\right) dv.$$

Here $s_j = (t_{j-1} + t_j)/2$ for $1 \leq j \leq n$, $s_0 = 0$, and $s_n = 1$. Following asymptotic theory we take $b_0 = O(n^{-1/5})$ and $b_1 = O(n^{-1/7})$. Other smoothing methods like local polynomial fitting or smoothing splines can be employed here instead of kernel smoothing. It would not change the convergence rates given in Theorem 4.1 below, since we assume a fixed design.

For any two functions f and g, the optimal shift function between f and g is then estimated by the solution of

$$\inf\left\{C\left(\hat{f}, \hat{f}', \hat{g}, \hat{g}', u, \alpha\right) : u \in C^1, \alpha \in \mathbf{R}\right\}.$$

Obviously some theoretical and practical questions need to be addressed. Does the variational problem (8) have a solution? Is a solution unique? How does dynamic time warping perform with noisy data? We address these questions in this section. As stated before, the analysis focuses on the alignment of two functions. First we prove the existence of a solution to (8).

LEMMA 4.1. *If f, $g \in C^1$, then there exists a solution of the variational problem* (8).

A proof is given in the Appendix. Since $\hat{f}$, $\hat{g}$, $\hat{f}'$, $\hat{g}'$ are continuous, this lemma shows that (8) has a solution if one replaces f, g, f', g' by $\hat{f}$, $\hat{g}$, $\hat{f}'$, $\hat{g}'$.

It is easy to see that (8) has many solutions in some cases. An example is $g(u(t)) = cf(t)$ for some strictly increasing u and some constant c, where $f(t) =$ constant on some open interval $(a, b) \subset [0, 1]$. Obviously $(u, 1)$ is a solution of (8). Now take $(a_1, b_1) \subset (a, b)$ and define v by $v(t) = u(t)$ on $[0, 1] \setminus (a_1, b_1)$ and anything on (a_1, b_1) such that v is strictly increasing and $v \in C^1[0, 1]$. Then $(v, 1)$ is also a solution of (8). To make the solution unique, one can replace ϕ in (8) by a strictly convex function with minimum at 1, but this may drive the solution of (8) away from the optimal shift function, if the identity $u(t) = t$ is not the optimal shift function. Similarly, if $g'(u(t)) = cf'(t)$ and $f'(t) =$ constant on $(a, b) \in [0, 1]$, then (8) has many solutions. Note that even though (8) has many solutions in each of these cases, the alignment is unique. That is,

$$\|g(u) - g(v)\|_2 \|g'(u) - g'(v)\|_2 = 0.$$

Here $\|\cdot\|_2$ stands for the L^2 norm of square integrable functions on $[0, 1]$. We suspect that this equation is true in general if (8) has more than one solution, but we could not prove it.

In data analysis, parametric linear function fitting is the approach most often used. Here, as in other problems in nonparametric function fitting, a linear regression function becomes a degenerate case. However, it is more a theoretical than a practical problem, since linear or almost linear functions can easily be spotted in an exploratory analysis. If both f and g are linear and if both their derivatives are positive or negative, then $f'/\|f'\| - g'/\|g'\| \equiv 0$. Thus for any u defined on $[0, 1]$, the pair $(u, 0)$ is a solution of (8). This results in incorrect alignment in this case. There are two ways to get around this problem. First, choose the penalty function ϕ as a strictly convex function with minimum $\phi(1) = 0$. Then the solution of (8) is $(u, 0)$ with $u(t) = t$. As mentioned before, this may drive the solution of (8) away from the optimal alignment between two functions in general. The second way is the following. Observe that $dg(u(t))/dt = g'(u(t))u'(t)$. It is therefore natural to replace $g'(u)$ in (6) by $g'(u)u'$. When both f and g are linear, the second term of (6) becomes $(1 - u'(t))^2$. The theoretical analysis will not change much if this replacement is made, but computation becomes difficult mainly because of the need to find a way of computing u' in order that it is still possible to solve the problem (8) by using dynamic programming.

Some notation is needed for statistical analysis. Let

$$A(t) \equiv A(f, f', g, g', \bar{u}, \bar{\alpha})(t) = \begin{pmatrix} F_{uu}(f, f', g, g', \bar{u}, \bar{\alpha}) & F_{u\alpha}(f, f', g, g', \bar{u}, \bar{\alpha}) \\ F_{u\alpha}(f, f', g, g', \bar{u}, \bar{\alpha}) & F_{\alpha\alpha}(f, f', g, g', \bar{u}, \bar{\alpha}) \end{pmatrix}.$$

We have used the notation $F_{uu}(f, f', g, g', \bar{u}, \bar{\alpha})$ for $F_{uu}(f(t), f'(t), g(t), g'(t), \bar{u}(t), \bar{\alpha})$ since t will be fixed. Here $(\bar{u}, \bar{\alpha})$ is an optimal solution of (8). Let $(\hat{u}, \hat{\alpha})$ denote a solution of (8) with f, g, f', g' replaced by $\hat{f}$, $\hat{g}$, $\hat{f}'$, $\hat{g}'$.

Recall that $\bar{\alpha} = 0, 1$ correspond to models where true shift functions can be recovered (Section 3). Below and in the remainder of this paper, let $\hat{f}^{(k)}(t) = d^k\hat{f}/dt^k$ and similarly for the derivatives of kernels and other estimators. We have the following theorem.

THEOREM 4.1. *Assume conditions* (i)–(iii).

(i) $f, g \in C^4(\mathbf{R})$ *and the optimal shift function* $\bar{u}$ *between* f *and* g *satisfies* $\delta + r < \bar{u}'(t) < M - r$, $t \in [0, 1]$.

(ii) $A(t)$ *is invertible at some* $t \in (0, 1)$.

(iii) $b_0 = O(n^{-1/5})$ *and* $b_1 = O(n^{-1/7})$.

Then the following conclusions hold:

(a) $$E(\hat{u}, \hat{\alpha})(t) - (\bar{u}, \bar{\alpha})(t) = \begin{cases} O(b_1^2) + o(n^{-1/2}b_1^{-3/2}\log^2(n)), & \bar{\alpha} \neq 1, \\ O(b_0^2) + o(n^{-1/2}b_0^{-1/2}\log^2(n)), & \bar{\alpha} = 1. \end{cases}$$

(b) *If* $\bar{\alpha} \neq 1$ *then*

$$\sqrt{nb_1^5}\left[(\hat{u}, \hat{\alpha})(t) - E(\hat{u}, \hat{\alpha})(t)\right] \Rightarrow N\left(0, A(t)^{-1}V(\bar{\alpha})A(t)^{-1}\right)$$

with N *a multinormal distribution and*

$$V(\bar{\alpha}) = \left[4(1-\bar{\alpha})^2\sigma^2\left(\frac{f'(t)}{\|f'\|} - \frac{g'(\bar{u}(t))}{\|g'\|}\right)^2\int_{-1}^{1} K_1^{(1)}(x)^2\,dx\right]\operatorname{diag}\left(\|g'\|^{-2}, 0\right).$$

If $V(\bar{\alpha}) = 0$ *then* $\sqrt{nb_1^5}[(\hat{u}, \hat{\alpha})(t) - E(\hat{u}, \hat{\alpha})(t)] \to 0$ *in probability.*

(c) *If* $\bar{\alpha} = 1$ *then*

$$\sqrt{nb_0^3}\left[(\hat{u}, \hat{\alpha})(t) - E(\hat{u}, \hat{\alpha})(t)\right] \to N\left(0, A(t)^{-1}V_1A(t)^{-1}\right)$$

with N *a multinormal distribution and*

$$V_1 = \left[4\sigma^2\left(\frac{f(t)}{\|f\|} - \frac{g(\bar{u}(t))}{\|g\|}\right)^2\int_{-1}^{1} K_0^{(1)}(x)^2\,dx\right]\operatorname{diag}\left(\|g\|^{-2}, 0\right).$$

If $V_1 = 0$ *then* $\sqrt{nb_0^3}[(\hat{u}, \hat{\alpha})(t) - E(\hat{u}, \hat{\alpha})(t)] \to 0$ *in probability.*

The proof is given in the Appendix. It shows that dynamic time warping performs reasonably well with noisy data, though the convergence rate is not very fast ($n^{-1/7}$ for $\bar{\alpha} \neq 1$ and $n^{-1/5}$ for $\bar{\alpha} = 1$). We point out that these results on bias and variance hold for any cost function with three continuous Fréchet derivatives. This will be clear from the proof. We make some remarks about this theorem.

REMARKS. (a) Condition (i) is not a restriction since δ, r and M^{-1} can be as small as one wants.

(b) It would be nice to have a convergence rate for $\|\hat{u} - \bar{u}\|$. As a result of the nonuniqueness of solutions (8), it could, however, be complicated to show just that there exists an optimal solution $\hat{u}$ such that $\|\hat{u} - \bar{u}\| \to 0$.

(c) Since $(\overline{u}, \overline{\alpha})$ is an optimal solution, the second variation of the cost function at $(\overline{u}, \overline{\alpha})$, given by

$$D^2C(f, f', g, g', \overline{u}, \overline{\alpha})[u_1, \alpha_1; u_1, \alpha_1]$$
$$= \int_0^1 \left[(u_1, \alpha_1) A(t)(u_1, \alpha_1)^T + \phi''(\overline{u}')(u_1')^2\right] dt,$$

is nonnegative for any admissible (u_1, α_1). See the proof in the Appendix for the definition of the second variation, or second Fréchet derivative. Since $\overline{u}$ satisfies (i), $\phi''(\overline{u}') \equiv 0$. Hence $A(t)$ is semipositive definite for all $t \in [0, 1]$. Since $A(t)$ is continuous in t, condition (ii) implies that $A(t)$ is positive definite in a neighborhood of t. This makes the solution of (8) unique in a neighborhood of t, and therefore makes it possible to prove the pointwise result given in this theorem.

(d) One can check condition (ii) for some interesting models. SIM is an easy example while the proof for NLSM is not so easy.

(e) In the structural analysis proposed by Kneip and Gasser [(1992), Theorem 3, page 1289], the estimated shift function from noisy data converges to the true shift function at a rate of $O(n^{-1/5})$. Here the rate $\sqrt{nb_1^5} = O(n^{-1/7})$ when $\overline{\alpha} \leq 1$ is slower because the second derivative of $\hat{g}$ is involved when solving (8) (see the proof given in the Appendix).

5. Simulations. We undertook a small scale simulation (100 runs) to evaluate the practical performance of dynamic time warping. Both the classical and our new cost function are evaluated. The evaluation is performed for the basic problem of warping one function to a second one. The assumption $u(0) = 0$, which is natural and useful in many applications, is not made in this section.

The base function (or shape function) is shown in Figure 3 and is defined by

$$(14)\quad s(t) = \begin{cases} s_1(t) + 4t\sin(31.4(0.45 - t)), & t \leq 0.45, \\ s_1(t) + 10(s_2(t) - s_1(t))(0.45 - t), & 0.45 \leq t \leq 0.55, \\ s_2(t) + 4(1 - t)\sin(31.4(t - 0.55)), & t \geq 0.55. \end{cases}$$

with

$$s_1(t) = 0.25(t-5)^2 - 8(0.45-t), \qquad s_2(t) = -0.25(t+4)^2 + 8(t-0.55).$$

We consider the model

$$f_i(t) = a_i s(h_i(t)) + d_i, \qquad i = 1, 2,$$

where a_i and d_i are constants while h_i is a strictly increasing shift function. The base function s has 8 extremes, denoted by τ_j, $j = 1, \ldots, 8$. Let $\tau_{ij} = h_i^{-1}(\tau_j)$, $j \in K_i$, be the corresponding extremes of f_i which are present on the graph $\{(t, f_i(t))\colon t \in [0, 1]\}$. K_i is a set of indices. Note that τ_{ij} is present on the graph if and only if $\tau_{ij} \in [0, 1]$.

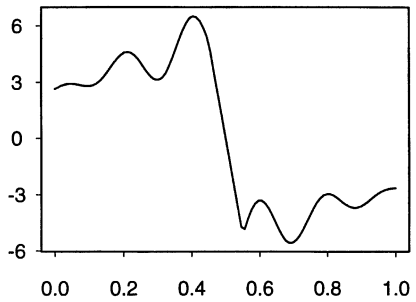

FIG. 3. *Shape function.*

Under optimal shifting of the second function to the first one, the extreme τ_{2j} should be shifted to τ_{1j} for $j \in K \equiv K_1 \cap K_2$. Let $\hat{\tau}_{1j}$ be the image of τ_{2j} under dynamic time warping for $j \in K$. We use

$$\rho \equiv \frac{1}{|K|} \sum_{j \in K} \left(\hat{\tau}_{1j} - \tau_{1j}\right)^2 \tag{15}$$

as an error measurement for aligning extremes. This choice is made for the following reasons. First, the norm $\|\hat{u} - u\|_2$, with $\hat{u}$ and u estimated and true shift functions, respectively, is not very sensitive and thus not appropriate as criterion (a different standardization might, however, help). One could also compute the difference $\|f_1(\cdot) - f_2(u(\cdot))\|_2$, but this is often not informative enough about the appropriate alignment.

5.1. *Shape-invariant model.* In this simulation we consider the shape-invariant model; that is, the shift functions are

$$h_i(t) = \frac{t - b_i}{c_i}.$$

For 100 runs, first 200 sample curves are generated and grouped into 100 pairs. Then we apply dynamic time warping with a prespecified cost function to warp one curve in a pair to the other and compute the warping error by (15). The bandwidth for kernel smoothing was chosen data adaptively via a plug-in rule [Gasser, Kneip and Köhler (1991)]. The parameters in the sample curves are generated as follows:

$$a_i = \max(1 + 5\sigma N(0,1), 0.5), \qquad b_i = 0.1\sigma N(0,1),$$
$$c_i = 1 + \sigma(U(0,1) - 0.5), \qquad d_i = 20\sigma N(0,1).$$

Here $N(0,1)$ stands for a standard normal variable and $U(0,1)$ for a uniform variable on $(0,1)$. The noise is generated as $\varepsilon_{ij} = N(0, \sigma_\varepsilon{}^2)$. The data are then formed as

$$y_{ij} = f_i(t_{ij}) + \varepsilon_{ij}.$$

TABLE 2
Simulation with SIM model

Cost function	σ_ε	Mean	Variance
Classical	0.1	1.81*e*-4	1.53*e*-7
New	0.1	8.77*e*-5	4.59*e*-9
Classical	0.2	9.71*e*-4	1.92*e*-6
New	0.2	2.74*e*-4	1.75*e*-7
Classical	0.4	3.28*e*-3	2.52*e*-5
New	0.4	1.59*e*-3	8.89*e*-6
Classical	0.6	3.95*e*-3	2.80*e*-5
New	0.6	2.80*e*-3	3.13*e*-5
Classical	1.0	2.87*e*-2	8.43*e*-3
New	1.0	2.80*e*-2	6.94*e*-3

The sample size for each curve is 100. We happen to take $\sigma = \sigma_\varepsilon$ in this and subsequent subsections, but there is no special reason for doing so.

Table 2 gives the results of 100 runs. Let ρ_i, $i = 1, \ldots, 100$, be the error (15) of the ith run. The column "Mean" is the sample means of $\{\rho_i\}$, while the column "variance" is the sample variance of $\{\rho_i\}$, leading easily to standard errors of simulation. The rows starting with "Classical" give the results for the classical quadratic cost function, and the rows starting with "New" give the results for cost function (7).

A typical run with $\sigma = \sigma_\varepsilon = 0.2$ is shown in Figure 4. The error of aligning extremes using the classical cost function is $3.33e - 3$, and the error using the new cost function is $1.66e - 4$. The true shift function is linear for the SIM model.

The results are visually appealing, and even more so for the new cost function.

5.2. *Nonlinear shift model.* We consider a nonlinear shift model for the simulation. The shift function is modeled by

$$h_i(t) = t + \frac{\alpha_i}{2\pi} \sin\bigl(2\pi(\beta_i t + \gamma_i)\bigr)$$

and the model is again

$$f_i(t) = a_i s\bigl(h_i(t)\bigr) + d_i.$$

The parameters a_i, d_i are generated as in the last simulation, while α_i, β_i, γ_i are generated as follows:

$$\alpha_i = \max\{-0.7, \min\{0.35N(0,1), 0.7\}\},$$

$$\beta_i = \min\{1 + 0.2N(0,1), 1.4\}, \qquad \gamma_i = 0.5(U(0,1) - 0.5).$$

Note that typically $|\beta_i| \le 1.4$ and therefore

$$h_i^*(t) = 1 + \alpha_i \beta_i \cos\bigl(2\pi(\beta_i t + \gamma_i)\bigr) \ge 1 - |\alpha_i \beta_i| > 0.$$

Thus we have strictly increasing shift functions.

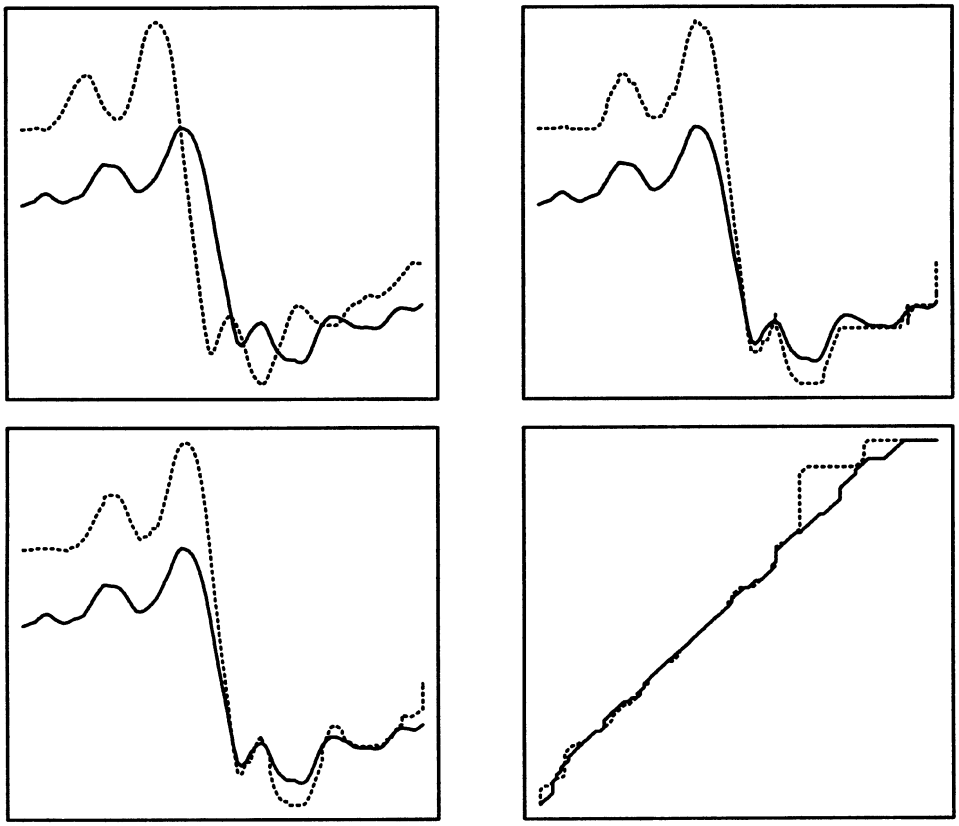

FIG. 4. *Alignment for SIM model. Top left*: *two smoothed sample curves*; *top right*: *warping result using the classical cost function*; *bottom left*: *warping result using the new cost function*; *bottom right*: *two warping functions* (*dotted line for classical cost function*, *solid line for new cost function*).

The simulation is done as follows. First 100 shift functions are generated and the data for 100 curves are formed as

$$y_{ij} = f_i(t_{ij}) + N(0, \sigma_\varepsilon).$$

Each curve is sampled at 100 points. After kernel smoothing, each curve is aligned to the shape function $s(t)$ and the error ρ_i is computed. Finally, we compute the sample mean and variance of $\{\rho_i\}$. The results from 100 runs are given in the Table 3.

A typical run with $\sigma = \sigma_\varepsilon = 0.2$ is shown in Figure 5. The error of aligning extremes using the classical cost function is $1.2e-4$, and the error using the new cost function is $2.55e-5$. Evidently, dynamic time warping with the classical cost function has difficulty in aligning the signals properly. This is not due to noise but to the more complicated shift function. The new cost function (7) performs much better.

TABLE 3
Simulation with NLSM model

Cost function	σ_ε	Mean	Variance
Classical	0.1	1.44*e*-3	7.56*e*-6
New	0.1	8.08*e*-4	2.24*e*-6
Classical	0.2	2.16*e*-3	7.04*e*-6
New	0.2	1.10*e*-3	3.61*e*-6
Classical	0.4	2.32*e*-3	1.71*e*-5
New	0.4	8.75*e*-4	2.82*e*-6
Classical	0.6	2.01*e*-3	7.49*e*-6
New	0.6	1.01*e*-3	3.09*e*-6
Classical	1.0	2.75*e*-3	1.36*e*-5
New	1.0	1.79*e*-3	8.94*e*-6

5.3. *Conclusions from simulations*. Dynamic time warping is an appropriate method for aligning two functions. The new cost function outperforms the classical cost function, because of the use of derivatives. Furthermore, dynamic time warping with the new cost function produces smoother shift functions and prevents the breakdown shown in Figure 5 (top left graphics).

As expected, the performance of warping depends mainly on the amount of true shifts—the difference between true shift functions and the identity function (see the simulation in Section 5.1). The noise level does not affect the warping results very much (see the simulation in Section 5.2) since data are smoothed before warping.

Compared to the structural analysis as described in steps (1)–(4) in the introduction, the advantage of dynamic time warping is that it automatically aligns structural points of two functions. It needs thus less a priori knowledge and less manpower, but structural analysis could still be preferable in difficult situations.

APPENDIX A

Proofs.

A.1. *Proof of Lemma* 4.1. First note that we can restrict $\alpha \in [0,1]$, because for any $\alpha \in [0,1]^c$, $C(f, f', g, g', u, \alpha) \geq \min\{C(f, f', g, g', u, 0), C(f, f', g, g', u, 1)\}$. Since $C(f, f', g, g', u, \alpha)$ is bounded from below, there exists a feasible sequence (u_n, α_n) such that

$$\lim_{n\to\infty} C(f, f', g, g', u_n, \alpha_n) = \sum_{u,\alpha} C(f, f', g, g', u, \alpha).$$

Let $I(\delta, M) = \{u \in \mathbf{C}^1[0,1]: |u(0)| \leq M, \alpha \leq u' \leq M\}$. Then $I(\delta, M)$ is compact under the norm $|u| = \sup\{|u(t)| + |u'(t)|: t \in [0,1]\}$. Therefore $I(\delta, M) \times [0,1]$ is compact and $\{u_n, \alpha_n\}$ has a subsequence which converges to a limit uniformly. This limit is a solution to (8) because C is continuous in (u, α). □

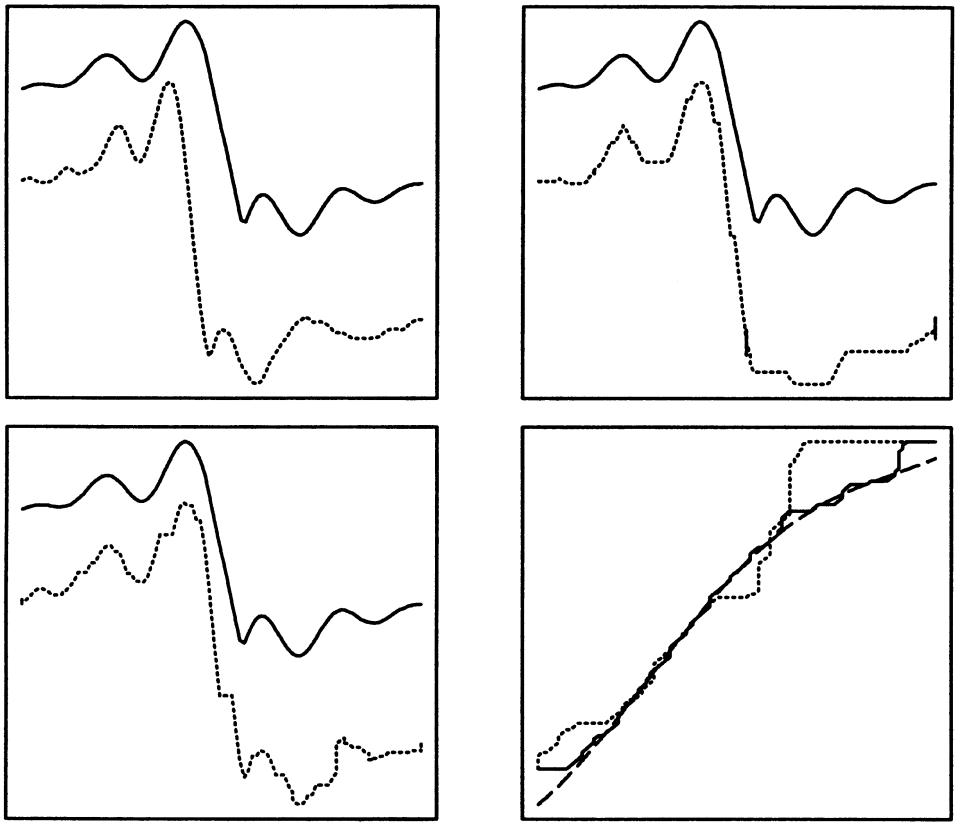

FIG. 5. *Alignment for NLSM model. Top left*: *two smoothed sample curves*; *top right*: *warping result using the classical cost function*; *bottom left*: *warping result using the new cost function*; *bottom right*: *warping functions* (*dotted line for classical cost function*, *solid line for new cost function*, *dashed line for true shift function*).

A.2. *Proof of Theorem* 4.1. The proof proceeds as follows. First, the difference $(\hat{u}, \hat{\alpha}) - (\hat{u}, \overline{\alpha})$ can be represented as a linear functional of the basic statistics plus higher order error terms. These basic statistics are $\hat{f}(t) - f(t)$, $\hat{f}^{(1)}(t) - f'(t)$, $\hat{f}'(t) - f'(t)$, $\hat{f}'^{(1)}(t) - f''(t)$, $\|\hat{f}\| + \|f\|$, $\|\hat{f}'\| - \|f'\|$, and similar statistics with $f(t)$ replaced by $g(\overline{u}(t))$. Note that $\hat{f}^{(1)}(t)$ is different from $\hat{f}'(t)$. To treat bias and variance asymptotically, we need only to take care of these basic statistics. The bias of the linear functional is simply the linear combination of the bias of those basic statistics, available in the literature. To deal with the variance we note that the dominating terms are those involving second derivatives: $\hat{f}'^{(1)}(t) - f''(t)$ and $\hat{g}'^{(1)}(\overline{u}(t)) - g''(\overline{u}(t))$. All other terms can be neglected as far as asymptotic variance is concerned. Therefore the main issue is to develop such a representation, which will be the first thing to do.

Some more notation is needed. Since, $f, g \in C^4$, $C(f, f', g, g', u, \alpha)$ has three continuous Fréchet derivatives with respect to (u, α). Denote these derivatives by DC, D^2C and D^3C, respectively. That is,

$$\begin{aligned}
&DC(f, f', g, g', u, \alpha)[u_1, \alpha_1] \\
&\quad = \lim_{\delta \to 0} \frac{C(f, f', g, g', u + \delta u_1, \alpha + \delta\alpha_1) - C(f, f', g, g', u, \alpha)}{\delta}, \\
&D^2C(f, f', g, g', u, \alpha)[u_1, \alpha_1; u_2, \alpha_2] \\
&\quad = \lim_{\delta \to 0} \big(DC(f, f', g, g', u + \delta u_2, \alpha + \delta\alpha_2)[u_1, \alpha_1] \\
&\qquad\qquad - DC(f, f', g, g', u, \alpha)[u_1, \alpha_1]\big)\delta^{-1},
\end{aligned}$$

and D^3C is defined in the same way.

Since $(\hat{u}, \hat{\alpha})$ is the minimizer of $C(\hat{f}, \hat{f}' \hat{g}, \hat{g}', u, \alpha)$, for any (u_1, α_1) with $u_1 \in C^1[0, 1]$,

$$\begin{aligned}
0 &= DC\Big(\hat{f}, \hat{f}', \hat{g}, \hat{g}', \hat{u}, \hat{\alpha}\Big)[u_1, \alpha_1] \\
&= DC(f, f', g, g', \hat{u}, \hat{\alpha})[u_1, \alpha_1] \\
&\quad + R\Big(f, f', g, g', \hat{f}, \hat{f}', \hat{g}, \hat{g}', \hat{u}, \hat{\alpha}\Big)[u_1, \alpha_1] \\
&= D^2C(f, f', g, g', \overline{u}, \overline{\alpha})\Big[u_1, \alpha_1; \hat{u} - \overline{u}, \hat{\alpha} - \overline{\alpha}\Big] \\
&\quad + O\big(|(\hat{u} - \overline{u}, \hat{\alpha} - \overline{\alpha})|^2\big) \\
&\quad + R\Big(f, f', g, g', \hat{f}, \hat{f}', \hat{g}, \hat{g}', \hat{u}, \hat{\alpha}\Big)[u_1, \alpha_1].
\end{aligned} \tag{16}$$

We have used the fact that $DC(f, f', g, g', \overline{u}, \overline{\alpha})[u_1, \alpha_1] = 0$. The O term equals

$$D^3C(f, f', g, g', \psi_u, \psi_\alpha)\Big[u_1, \alpha_1; \hat{u} - \overline{u}, \hat{\alpha} - \overline{\alpha}; \hat{u} - \overline{u}, \hat{\alpha} - \overline{\alpha}\Big]$$

for some (ψ_u, ψ_α) between $(\overline{u}, \overline{\alpha})$ and $(\hat{u}, \hat{\alpha})$. The term R is simply

$$DC\Big(\hat{f}, \hat{f}', \hat{g}, \hat{g}', \hat{u}, \hat{\alpha}\Big)[u_1, \alpha_1] - DC(f, f', g, g', \hat{u}, \hat{\alpha})[u_1, \alpha_1].$$

Since ϕ does not depend on $f, f', g, g', \hat{f}, \hat{f}', \hat{g}, \hat{g}'$ and their derivatives, R does not depend on ϕ (canceled out). Consequently

$$\begin{aligned}
&R\Big(f, f', g, g', \hat{f}, \hat{f}', \hat{g}, \hat{g}', \hat{u}, \hat{\alpha}\Big)[u_1, \alpha_1] \\
&\quad = \int_0^1 \Big[\Big(F_u\Big(\hat{f}, \hat{f}', \hat{g}, \hat{g}', \hat{u}, \hat{\alpha}\Big) - F_u(f, f', g, g', \hat{u}, \hat{\alpha})\Big)u_1 \\
&\qquad + \Big(F_\alpha\Big(\hat{f}, \hat{f}', \hat{g}, \hat{g}', \hat{u}, \hat{\alpha}\Big) - F_\alpha(f, f', g, g', \hat{u}, \hat{\alpha})\Big)\alpha_1\Big]\, dt.
\end{aligned} \tag{17}$$

Easy computations show that

$$D^2C(f,f',g,g',\overline{u},\overline{\alpha})\big[u_1,\alpha_1;\hat{u}-\overline{u},\hat{\alpha}-\overline{\alpha}\big]$$
$$=\int_0^1\Big[(u_1,\alpha_1)A(t)(\hat{u}-\overline{u},\hat{\alpha}-\overline{\alpha})^T+\phi''(\overline{u}')u_1'(\hat{u}'-\overline{u}')\Big]\,dt.$$

Since (u_1,α_1) can be chosen arbitrarily, it follows from (16), (17) and $\phi''(\overline{u}')\equiv 0$ that

$$A(t)(\hat{u}-\overline{u},\hat{\alpha}-\overline{\alpha})^T=\hat{B}(t)+O\big(|(\hat{u}-\overline{u},\hat{\alpha}-\overline{\alpha})|^2\big), \tag{18}$$

with

$$\hat{B}(t)=\begin{pmatrix}F_u\big(\hat{f},\hat{f}',\hat{g},\hat{g}',\hat{u},\hat{\alpha}\big)-F_u(f,f',g,g',\hat{u},\hat{\alpha})\\ F_\alpha\big(\hat{f},\hat{f}',\hat{g},\hat{g}',\hat{u},\hat{\alpha}\big)-F_\alpha(f,f',g,g',\hat{u},\hat{\alpha})\end{pmatrix}.$$

One could also derive (18) from Euler equations. Since the inverse A^{-1} exists and since F_u, F_α are continuous, one immediately gets $(\hat{u}-\overline{u},\hat{\alpha}-\overline{\alpha})\to 0$ in probability from the fact that $(\hat{f},\hat{f}',\hat{g},\hat{g}',\hat{f}^{(1)},\hat{g}^{(1)},\hat{f}'^{(1)},\hat{g}'^{(1)})-(f,f',g,g',f',g',f'',g'')\to 0$ in probability.

We now deal with $\hat{B}$. First note the following two algebraic equalities

$$\frac{\hat{f}}{\|\hat{f}\|}=\frac{f}{\|f\|}+\frac{\hat{f}-f}{\|f\|}-\frac{f}{\|f\|}\frac{\|\hat{f}\|-\|f\|}{\|f\|}-\frac{\|\hat{f}\|-\|f\|}{\|f\|}\left(\frac{\hat{f}}{\|\hat{f}\|}-\frac{f}{\|f\|}\right),$$
$$\frac{\hat{f}^{(1)}}{\|\hat{f}\|}=\frac{f'}{\|f\|}+\frac{\hat{f}^{(1)}-f'}{\|f\|}-\frac{f'}{\|f\|}\frac{\|\hat{f}\|-\|f\|}{\|f\|}-\frac{\|\hat{f}\|-\|f\|}{\|f\|}\left(\frac{\hat{f}^{(1)}}{\|\hat{f}\|}-\frac{f'}{\|f\|}\right).$$

The same relations hold if we replace, f, f', $\hat{f}$, $\hat{f}^{(1)}$ by f', f'', $\hat{f}'$, $\hat{f}'^{(1)}$. For any (u,α), we simply write f for $f(t)$, g for $g(u(t))$, and so on. Then it follows from the two algebraic equalities that

$$F_u\big(\hat{f},\hat{f}',\hat{g},\hat{g}',u,\alpha\big)-F_u(f,f',g,g',u,\alpha)$$
$$=2\alpha^2\left[\left(\frac{\hat{g}}{\|\hat{g}\|}-\frac{\hat{f}}{\|\hat{f}\|}\right)\frac{\hat{g}^{(1)}}{\|\hat{g}\|}-\left(\frac{g}{\|g\|}-\frac{f}{\|f\|}\right)\frac{g'}{\|g\|}\right]$$
$$-2(1-\alpha)^2\left[\left(\frac{\hat{g}'}{\|\hat{g}'\|}-\frac{\hat{f}'}{\|\hat{f}'\|}\right)\frac{\hat{g}'^{(1)}}{\|\hat{g}'\|}-\left(\frac{g'}{\|g'\|}-\frac{f'}{\|f'\|}\right)\frac{g''}{\|g'\|}\right]$$
$$=2\alpha^2G_1\big(g,\hat{g},f,\hat{f},u,\alpha\big)+o\big(|G_1|\big)+2(1-\alpha)^2G_2\big(g',\hat{g}',f',\hat{f}',u,\alpha\big)$$
$$+o\big(|G_2|\big).$$

The G functionals are defined by

$$\begin{aligned} &G_1(g,\hat{g},f,\hat{f},u,\alpha) \\ &\quad = \frac{g'}{\|g\|}\frac{\hat{g}-g}{\|g\|} - 2\frac{g'}{\|g\|}\frac{g}{\|g\|}\frac{\|\hat{g}\|-\|g\|}{\|g\|} - \frac{g'}{\|g\|}\frac{\hat{f}-f}{\|f\|} \\ &\qquad + 2\frac{g}{\|g\|}\frac{\|\hat{f}\|-\|f\|}{\|f\|} + \left(\frac{g}{\|g\|} - \frac{f}{\|f\|}\right)\frac{\hat{g}^{(1)}-g'}{\|g\|}, \\ &G_2(g',\hat{g}',f',\hat{f}',u,\alpha) \\ &\quad = \frac{g''}{\|g'\|}\frac{\hat{g}'-g'}{\|g'\|} - 2\frac{g''}{\|g'\|}\frac{g'}{\|g'\|}\frac{\|\hat{g}'\|-\|g'\|}{\|q'\|} - \frac{g''}{\|g'\|}\frac{\hat{f}'-f'}{\|f'\|} \\ &\qquad + 2\frac{g'}{\|g'\|}\frac{\|\hat{f}'\|-\|f'\|}{\|f'\|} + \left(\frac{g'}{\|g'\|} - \frac{f'}{\|f'\|}\right)\frac{\hat{g}'^{(1)}-g''}{\|g'\|}. \end{aligned}$$

F_α is treated in the same way and one gets

$$\begin{aligned} &F_\alpha(\hat{f},\hat{f}',\hat{g},\hat{g}',u,\alpha) - F_\alpha(f,f',g,g',u,\alpha) \\ &\quad = 4\alpha H(f,\hat{f},g,\hat{g},u,\alpha) - 4(1-\alpha)H(f',\hat{f}',\tilde{g},\hat{g}',u,\alpha) + o(|H|). \end{aligned}$$

The H functional is defined by

$$\begin{aligned} &H(f,\hat{f},g,\hat{g},u,\alpha) \\ &\quad = \left(\frac{f}{\|f\|} - \frac{g}{\|g\|}\right)\left(\frac{\hat{f}-f}{\|f\|} - \frac{f}{\|f\|}\frac{\|\hat{f}\|-\|f\|}{\|f\|} - \frac{\hat{g}-g}{\|g\|} + \frac{g}{\|g\|}\frac{\|\hat{g}\|-\|g\|}{\|g\|}\right), \end{aligned}$$

and the o term is

$$o(|H|) = o\big(|H(f,\hat{f},g,\hat{g},u,\alpha)|\big) + o\big(|H(f',\hat{f}',\tilde{g},\hat{g}',u,\alpha)|\big).$$

Combining all the above considerations, we have shown that

$$\hat{B}(t) = \hat{B}_0(t) + \text{higher order error terms}$$

with

$$\hat{B}_0(t) = \begin{pmatrix} 2\hat{\alpha}^2 G_1(g,\hat{g},f,\hat{f},\hat{u},\hat{\alpha}) + 2(1-\hat{\alpha})^2 G_2(g',\hat{g}',f',\hat{f}',\hat{u},\hat{\alpha}) \\ 4\hat{\alpha}H(f,\hat{f},g,\hat{g},\hat{u},\hat{\alpha}) - 4(1-\hat{\alpha})H(f',\hat{f}',g',\hat{g}',\hat{u},\hat{\alpha}) \end{pmatrix}.$$

Let us make one more observation here. If we replace $(\hat{u},\hat{\alpha})$ by $(\overline{u},\overline{\alpha})$ in $\hat{B}_0(t)$, the error terms are those like $(\hat{g}(\xi)-g(\xi))(\hat{u}-\overline{u})$, with some ξ between $\hat{u}$ and $\overline{u}$, by a first order Taylor expansion. Therefore,

$$\hat{B}(t) = \hat{B}_1(t) + \text{higher order error terms},$$

where $\hat{B}_1(t)$ is obtained from $\hat{B}_0(t)$ by replacing $(\hat{u}, \hat{\alpha})$ with $(\overline{u}, \overline{\alpha})$. Now (18) becomes

$$(\hat{u} - \overline{u}, \hat{\alpha} - \overline{\alpha})^T = A(t)^{-1}\hat{B}_1(t) + \text{higher order error terms.} \tag{19}$$

We are ready to complete the proof of Theorem 4.1.

PROOF OF (a). This part follows from (19) and the lemma of Kneip and Gasser [(1992), pages 1291 and 1292], which shows that

$$E(\hat{f} - f) = O(b_0^2), \qquad E(\hat{f}^{(1)} - f') = O(b_0^2),$$
$$E(\hat{f}' - f') = O(b_1^2), \qquad E(\hat{f}'^{(1)} - f'') = O(b_1^2).$$

Also,

$$\left|E\|\hat{f}\| - \|f\|\right| \le E\|\hat{f} - E\hat{f}\| + \|E\hat{f} - f\| = o\left(\frac{\log^2(n)}{\sqrt{nb_0}}\right) + O(b_0^2),$$

$$\left|E\|\hat{f}'\| - \|f'\|\right| \le E\|\hat{f}' - E\hat{f}'\| + \|E\hat{f}' - f'\| = o\left(\frac{\log^2(n)}{\sqrt{nb_1^3}}\right) + O(b_1^2).$$

The same is true for g. With $b_0 = O(n^{-1/5})$ and $b_1 = O(n^{-1/7})$, we have

$$O(b_0^2) = o(b_1^2), \qquad (nb_0)^{-1} = o\left((nb_1^3)^{-1}\right).$$

Part (a) follows.

PROOF OF (b). If $\overline{\alpha} \neq 1$ then $\hat{B}_1(t)$ depends on $\hat{g}'^{(1)}$. It follows from (b) of the Lemma of Kneip and Gasser (1992) that the variances of $\hat{f}'$, $\hat{g}'$, $\hat{f}^{(\gamma_0)}$ and $\hat{g}^{(\gamma_0)}$ are dominated by the variance of $\hat{g}'^{(1)}$ if $\gamma_0 = 0, 1$. We need also that the variances of $\|\hat{f}'\|$, $\|\hat{g}'\|$, $\|\hat{f}\|$ and $\|\hat{g}\|$ are dominated by that of $\hat{g}'^{(1)}$. This can be seen, for example, for $\|\hat{f}'\|$, as follows:

$$\begin{aligned}
E\left(\|\hat{f}'\| - E\|\hat{f}'\|\right)^2 &= E\left(\|f'\hat{\|} - \|E\hat{f}'\| + \|E\hat{f}'\| - E\|\hat{f}'\|\right)^2 \\
&\le 2E\left(\|\hat{f}'\| - \|E\hat{f}'\|\right)^2 + 2\left(\|E\hat{f}'\| - E\|\hat{f}'\|\right)^2 \\
&\le 4E\|\hat{f}' - E\hat{f}'\|^2 = o\left(\frac{\log^4(n)}{nb_1^3}\right) \\
&= o\left(n^{-1}b_1^{-5}\right) = o\left(E\left(\hat{g}'^{(1)} - E\hat{g}'^{(1)}\right)^2\right).
\end{aligned}$$

Now it is clear that the covariance of any two different terms in $\hat{B}_1(t)$ is an $o(\cdot)$ term of $\mathrm{Var}(\hat{g}'^{(1)})$ because $|\mathrm{Cov}(X, Y)| \le (\mathrm{Var}(X)\mathrm{Var}(Y))^{-1/2}$ for any X and Y.

Part (b) now follows from (19) and the fact that [cf. the lemma of Kneip and Gasser (1992)]

$$\operatorname{Var}\bigl(\hat{g}'^{(1)}\bigr) = \frac{\sigma^2}{nb_1^5}\int_{-1}^{1} K_1^{(1)}(x)^2\,dx + o\left(\frac{1}{nb_1^5}\right)$$

and that $\hat{g}'^{(1)}(t) - E\hat{g}'^{(1)}(t)$ converges to a normal variable if properly scaled; compare Gasser and Müller (1984).

PROOF OF (c). If $\overline{\alpha} = 1$, then $\hat{B}_1(t)$ does not depend on $\hat{f}'$, $\hat{g}'$ and $\hat{g}'^{(1)}$. The dominating term is then the term involving $\hat{g}^{(1)}$, as far as asymptotic variance-covariance is concerned. Again the covariance of any two terms in $\hat{B}_1(t)$ is an $o(\cdot)$ term of $\operatorname{Var}(\hat{g}^{(1)})$. Note that

$$\operatorname{Var}\bigl(\hat{g}^{(1)}\bigr) = \frac{\sigma^2}{nb_0^3}\int_{-1}^{1} K_0^{(1)}(x)^2\,dx + o\left(\frac{1}{nb_0^3}\right).$$

The rest of the proof for (c) is the same as in the proof of (b). □

Acknowledgments. We thank two referees and an Associate Editor for helpful comments.

REFERENCES

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

GASSER, TH., KNEIP, A. and KÖHLER, W. (1991). A flexible and fast method for automatic smoothing. *J. Amer. Statist. Assoc.* **86** 643–652.

GASSER, TH., KNEIP, A., ZIEGLER, P., MOLINARI, L., PRADER, A. and LARGO, R. (1994). Development and outcome of indices of obesity in normal children. *Ann. of Human Biology* **21** 275–286.

GASSER, TH. and MÜLLER, H. (1984). Estimating regression functions and their derivatives by the Kernel method. *Scand. J. Statist.* **11** 171–185.

GASSER, TH., MÜLLER, H. and MAMMITZSCH, V. (1985). Kernels for nonparametric curve estimation. *J. Roy. Statist. Soc. Ser. B* **47** 238–252.

HÄRDLE, W. and MARRON, J. S. (1990). Semiparametric comparison of regression curves. *Ann. Statist.* **18** 63–89.

HÖHNE, H., COKER, C., LEVINSON, S. and RABINER, L. (1983). On temporal alignment of sentences of natural and synthetic speech. *IEEE Trans. Acoust. Speech Signal Process.* **31** 807–813.

KNEIP, A. and ENGEL, J. (1995). Model estimation in nonlinear regression under shape invariance. *Ann. Statist.* **23** 551–570.

KNEIP, A. and GASSER, TH. (1988). Convergence and consistency results for self-modeling nonlinear regression. *Ann. Statist.* **16** 82–112.

KNEIP, A. and GASSER, TH. (1992). Statistical tools to analyze data representing a sample of curves. *Ann. Statist.* **20** 1266–1305.

LAWTON, W. H., SYLVESTRE, E. A. and MAGGO, M. S. (1972). Self-modeling regression. *Technometrics* **14** 513–532.

PARSONS, T. (1986). *Voice and Speech Processing*. McGraw-Hill, New York.

QI, Y. (1992). Time normalization in voice analysis. *J. Acoust. Soc. Am.* **92** 2569–2576.

RABINER, L. and SCHMIDT, C. (1980). Application of dynamic time warping to connected digit recognition. *IEEE Trans. Acoust. Speech Signal Process.* **28** 377–388.

RAMSAY, J. and DALZELL, C. (1991). Some tools for functional data analysis (with discussion). *J. Roy. Statist. Soc. Ser. B* **53** 539–572.

RAO, C. (1958). Some statistical methods for the comparison of growth curves. *Biometrics* **14** 1–17.

RICE, J. and SILVERMAN, B. (1991). Estimating the mean and covariance structure nonparametrically when data are curves. *J. Roy. Statist. Soc. Ser. B* **53** 233–243.

ROBERTS, K., LAWRENCE, P., EISEN, A. and HOIRCH, M. (1987). Enhancement and dynamic time warping of somatosensory evoked potential components applied to patients with multiple sclerosis. *IEEE Trans. Biomed. Eng.* **BME-34** 397–405.

SAKOE, H. and CHIBA, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Speech Signal Process.* **26** 43–49.

SILVERMAN, B. (1995). Incorporating parametric effects into functional principal components analysis. *J. Roy. Statist. Soc. Ser. B* **57** 673–689.

STÜTZLE, W., GASSER, TH., MOLINARI, L., LARGO, R., PRADER, A. and HUBER, P. (1980). Shape-invariant modeling of human growth. *Ann. of Human Biology* **7** 507–528.

DEPARTMENT OF PSYCHIATRY
NEURODYNAMICS LAB
STATE UNIVERSITY OF NEW YORK
450 CLARKSON AVE., BOX 1203
BROOKLYN, NEW YORK 11203-2098
E-MAIL: kmw@mendel.neurodyn.hscbklyn.edu

DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF ZÜRICH
CH-8006 ZÜRICH
SWITZERLAND
E-MAIL: tgasser@ifspm.unizh.ch