An approach for reference curve selection in curve registration

Cheng, Yu-Hsiang
National Chengchi University, Department of Statistics
NO.64, Sec.2, Zhinan Rd., Wenshan District
Taipei City 11605, Taiwan (R.O.C)
E-mail: 96354501@nccu.edu.tw

Huang, Tzee-Ming
National Chengchi University, Department of Statistics
NO.64, Sec.2, Zhinan Rd., Wenshan District
Taipei City 11605, Taiwan (R.O.C)
E-mail: tmhuang@nccu.edu.tw

1 Introduction

Functional data are usually functions of time, and it is common that a group of data curves follow the same pattern after they are aligned. The common pattern can be described by a shape function. In order to estimate the shape function, it is important to align the curves. Studies on curve alignment and shape function estimation can be found in literature. References can be found in Telesca and Inoue (2008).

A natural approach for synchronizing a set of curves is to first choose a curve as a reference, align other curves to the reference curve, and then estimate the shape function based on the aligned curves. There are two interesting questions related to this approach. First, does the choice of reference curve have a significant effect on the estimation accuracy of shape functions? If so, how should we choose the reference curve? We have not found much studies on these questions in literature, so we carried out several simulation experiments to find out the answers. From our simulation results, the choice of reference curve does have a significant effect on estimation.

Below we describe a model for data curves follow the same pattern after alignment. Suppose that m misaligned curves are observed at n equal spaced time points $t_1, \ldots, t_k$. Let y_{ik} denote the observed value of the i-th curve at time t_k. Suppose that for $1 \leq i, j \leq m$ and $1 \leq k \leq n$,

$$\begin{aligned} y_{ik} &= f_i(t_k) + \epsilon_{ik} \\ &= f_j(\mu_{ij}(t_k)) + \epsilon_{ik}, \end{aligned} \tag{1}$$

where $f_i(t_k)$ is the shape function for the i-th curve evaluated time t_k, which is also the j-th curve's shape function evaluated at warped time $\mu_{ij}(t_k)$, and ϵ_{ik} is the error term. In (1), μ_{ij} is the warping function for the i-th curve when the j-th curve is used as the reference.

In (1), if the j-th curve is used as the reference and the warping functions μ_{ij}'s are estimated by $\hat{\mu}_{ij}$'s, one can estimate the shape function f_j treating y_{ik} as the observation of f_j at time $\hat{\mu}_{ij}$ plus an error term. If the estimation of f_j at a point x involves only the $\hat{\mu}_{ij}$'s in a neighborhood of x, and there are few $\hat{\mu}_{ij}$'s in the neighborhood, the estimation of f_j at x may be unstable. This phenomenon gives us the idea to choose a reference function so that the estimation of the common shape function is stable.

We consider approximating the common shape function using B-spline basis functions. That is, when the j-th curve is used as the reference, f_j is approximated by

$$a_1 B_1 + \cdots + a_k B_k,$$

where B_1, ..., B_k form a B-spline basis. When $\hat{\mu}_{ij}$'s are used to estimate μ_{ij}'s and the estimation errors are small, we have

$$y_{ik} \approx a_1 B_1(\hat{\mu}_{ij}(t_k)) + \cdots + a_k B_k(\hat{\mu}_{ij}(t_k)) + \epsilon_{ik},$$

so the coefficients a_1, ..., a_k can be regarded as regression coefficients and estimated using least squared method. In such case, the estimation of the common shape function is stable if the estimation of the coefficients is stable, which is related to the notion of A-optimal design or D-optimal design. By considering an A-optimal design, one seeks to minimize the average of variances of coefficient estimators; by considering a D-optimal design, one seeks to minimize the volume of a classical confidence ellipsoid for the coefficients.

In this paper, we offer reference curve selection algorithms based on A-optimal/D-optimal designs. The algorithms are in Section 2. We also carry out several simulation studies to examine the performance of the proposed algorithms. Descriptions for the simulation studies and the results are presented in Section 3.

2 Algorithm

In this section, we describe our algorithm for reference curve selection. Suppose that we have m different curves of the same pattern after alignment, and each curve is observed at n time points t_1, ..., t_n. For $1 \le j \le m$ and $1 \le k \le n$, let y_{jk} be the observed value for the j-the curve at time t_k. If the i-th curve is chosen as the reference curve, then for $1 \le j \le m$ and $1 \le k \le n$,

$$y_{jk} = f_i(\mu_{ji}(t_k)) + \epsilon_{jk},$$

where f_i is the shape function for the i-th curve, μ_{ji} is the warping function for the j-th curve when the i-th curve is used as the reference, and ϵ_{jk} is the error.

In this algorithm, we use cubic B-splines to approximate shape functions f_i's and warping functions μ_{ji}'s, where approximately $2n^{1/3}$ equally spaced knots are used. That is, each f_i is of the form of a B-spline f_θ and each μ_{ji} is of the form of a B-spline μ_η where θ and η are the vectors of B-spline coefficients.

The algorithm is composed of the following 3 steps.

- Step 1. For a fixed i, the B-spline coefficients for the shape function f_i are estimated using least square method based on $(y_{i1}, t_1), \ldots, (y_{in}, t_n)$. That is, f_i is estimated by f_{θ^*}, where

  $$\sum_{k=1}^{n} (y_{ik} - f_\theta(t_k))^2$$

  is minimized at $\theta = \theta^*$. The estimator f_{θ^*} is denoted by $\hat{f}_i$. Similarly, for each $j \neq i$, the warping function μ_{ji} is estimated by μ_{η^*}, where

  $$\sum_{k=1}^{n} (y_{jk} - \hat{f}_i(\mu_\eta(t_k)))^2 \tag{2}$$

  is minimized at $\eta = \eta^*$. The estimator of μ_{η^*} is denoted by $\hat{\mu}_{ji}$. For $j = i$, $\hat{\mu}_{ji}$ is defined as the identity function.

- Step 2. For a fixed i, let $\tilde{t} = (t_1, \ldots, t_n)$, and let $\hat{\mu}_{ji}(\tilde{t}) = (\hat{\mu}_{ji}(t_1), \ldots, \hat{\mu}_{ji}(t_n))$ be the vector of adjusted times for the j-th curve with the i-th curve as the reference.

The m vectors of adjusted times are combined into a vector $\tilde{t}^*$. That is,

$$\tilde{t}^* = (\hat{\mu}_{1i}(\tilde{t}), \ldots, \hat{\mu}_{ni}(\tilde{t}))$$

Suppose that B_1, ..., B_ℓ denote the cubic spline basis functions correspond to approximately $2(mn)^{1/3}$ equally-spaced knots in $[0,1]$. Let $X = ((B_1^*)^T, \ldots, (B_\ell^*)^T)$, where B_i^* is the row vector of B_i evaluated at $\tilde{t}^*$.

The shape function f_i is re-estimated as a linear combination of B_1, ..., B_ℓ, where the vector of the coefficients is given by

$$(X^T X)^{-1} X^T Y, \tag{3}$$

where Y is the column vector $(y_{11}, \ldots, y_{1n}, \ldots, y_{m1}, \ldots, y_{mn})^T$. This re-estimated f_i is denoted by $\hat{f}_i^*$. The warping functions μ_{ji}'s are also re-estimated using (2) with the $\hat{f}_i$ replaced by $\hat{f}_i^*$. Let $\hat{\mu}_{ji}^*$'s denote the re-estimated μ_{ji}'s.

Note that (3) gives the least square solution for the regression problem

$$Y = X\beta + \varepsilon.$$

For the design matrix X, we also compute two indexes

$$Index_A = trace((X^T X)^{-1})$$

and

$$Index_D = det((X^T X)^{-1})$$

for evaluating choosing the i-th curve as the reference based on the A-optimal and D-optimal design criteria.

- Step 3. Carry out Step 1 and Step 2 for each $i \in \{1, \ldots, m\}$, and we obtain m $Index_A$ values and m $Index_D$ values, which correspond to the m reference curves.

3 Simulation studies

In this section, we examine the performance of the proposed reference curve selection algorithm via simulation experiments. In the simulation experiments, data curves are generated using the shape function

$$m(t) = \sin(2\pi t) + \sin(t^2),$$

which is also used in Sangalli et al. (2010) for curve generation. The warping functions are of form t^α and the error terms are distributed as $N(0, \sigma^2)$, where $\alpha \in \{1/1.9, 1/1.7, 1/1.5, 1/1.3, 1, 1.3, 1.5, 1.7, 1.9\}$ and $\sigma \in \{0.05, 0.1\}$. The time points are n equally-spaced points in $[0,1]$, where $n \in \{10, 15, 20, 25, 30\}$. Each experiment corresponds to a (n, σ) combination and is carried out 500 times. In each of the 500 replications, 9 data curves are generated using the 9 warping functions, the shape function and the errors term distributions mentioned above.

For the simulation study, we use the quantity

$$SE(i) = \sum_{j=1}^{m} \sum_{k=1}^{n} (f_i(\mu_{ji}(t_k)) - \hat{f}_i^*(\hat{\mu}_{ji}^*(t_k)))^2$$

to assess estimation accuracy when the i-th curve is used as the reference. A smaller SE value indicates that the estimation of shape function is more accurate. For each set of 9 simulated curves, let

$$SE_1 = \min_{1\le i\le 9} SE(i),$$

$$SE_9 = \max_{1\le i\le 9} SE(i),$$

SE_A be the SE value when the reference curve is chosen based on $Index_A$, and SE_D be the SE value when the reference curve is chosen based on $Index_D$, and

$$AV(SE) = \frac{1}{9}\sum_{i=1}^{9} SE(i).$$

Two ratios SE_A/SE_1, SE_D/SE_1 and the quantities $AV(SE)/SE_1$, SE_9/SE_1 are computed. The averages of the four quantities over the 500 replications are given in Table 1. The results in Table 1 indicate that choosing a reference curve using the proposed algorithm gives smaller estimation error comparing to choosing a reference curve at random.

	$\sigma = 0.05$				$\sigma = 0.1$			
	$\frac{SE_A}{SE_1}$	$\frac{SE_D}{SE_1}$	$\frac{AV(SE)}{SE_1}$	$\frac{SE_9}{SE_1}$	$\frac{SE_A}{SE_1}$	$\frac{SE_D}{SE_1}$	$\frac{AV(SE)}{SE_1}$	$\frac{SE_9}{SE_1}$
$n = 10$	1.063	1.044	1.497	2.721	1.060	1.052	1.198	1.609
$n = 15$	1.167	1.167	2.214	5.241	1.085	1.081	1.383	2.246
$n = 20$	1.226	1.192	2.228	5.329	1.097	1.082	1.378	2.172
$n = 25$	1.407	1.236	2.687	6.791	1.140	1.100	1.508	2.621
$n = 30$	1.295	1.071	2.609	5.550	1.093	1.055	1.472	2.293

Table 1: Averages for $\frac{SE_A}{SE_1}$, $\frac{SE_D}{SE_1}$, $\frac{AV(SE)}{SE_1}$ ***and*** $\frac{SE_9}{SE_1}$

Acknowledgement

Part of this research is supported by National Science Council in Taiwan under grant NSC 98-2118-M-004 -003 -.

REFERENCES (RÉFERENCES)

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

[2] Donatello Telesca and Lurdes Y. T. Inoue. Bayesian hierarchical curve registration. Journal of the American Statistical Association, 103(481):328–339, 2008.